The effects of music education on students' well-being. Empirical evidence from a field experiment

Preliminary - work in progress

This draft: October 17, 2020

Vera Schramm
University of Halle-Wittenberg

Abstract *Objective.* This thesis examines the effect of a music project which was carried out parallel at different middle schools in Bavaria on life satisfaction and on certain areas of satisfaction. *Methods.* I review prior work pertaining to music's impact mainly on cognitive outcomes. My analysis applies Bayesian inference and multilevel modeling on a longitudinal data set to asses music involvement and possible effects on life satisfaction, satisfaction with friends, satisfaction with the class and satisfaction with the school. A difference-in-difference framework allows to draw causal conclusion even despite the fact that the treatment group and the control group differ in some aspects *Results.* The students from the treatment group tend to have a more _____ Although _____, this effect is only minimal... *Conclusion.* Music participation in form of the studied music project caused little changes...

Contents

List of Abbreviations

BMU	Bundesverband für Musikunterricht
IQ	Intelligence Quotient
KIP	klasse.im.puls
FAU	Friedrich-Alexander-Universität
LS	Life satisfaction
SEA	... Average
SD	Standard Deviation
PWI-SC	Personal Wellbeing Index – School Children

List of Tables

List of Figures

1 Introduction

In German schools, subjects like arts and music are often considered less important than the typical hard subjects like math and science. Due to a lack of teachers, many classes are canceled, of which 80% are in the subject of music. In Saxony we see ongoing efforts to eliminate the subject of music from the curriculum entirely. Furthermore, the quality of music lessons suffers from the fact that 80% of its teaching staff are foreign to the subject (Möller, 2017).

Music education experts are concerned about this development. According to them, music should not be regarded as a private matter. Regardless of their socioeconomic background, school children must have the opportunity to receive high level music education because it is as important for a proper education as literacy and mathematics (Gebert, 2018). Prof. Höppner, the General Secretary of the German Music Council (Generalsekretär des Deutschen Musikrats), said in an interview that music education helps to build stable self-esteem by learning to access ones own emotions (Stoverock, n.d.). He points out that the phase where music can shape a young person explicitly well is completed by the age of 13. That stresses the importance of high quality music education for students in pre- and secondary school.

The Federal Association of Music Education (Bundesverband Musikunterricht (BMU)) has set up the "Agenda 2030" to initiate an improvement in music education. Similar to Höppner (Stoverock, n.d.), they consider music education valuable and essential for a social and cultural society. It encourages children to take responsibility and to increase their sense of self-determination (Bundesverband Musikunterricht (BMU), 2016, p. 2).

In views of this broad societal debate, empirical research has concentrated on the effect of music education on cognitive abilities (school grades and IQ in particular). Not so much effort was spent on observing the connection of music and well-being. This seems surprising because life satisfaction and happiness have become central research areas in the social sciences. My thesis addresses this research gap and investigates the effects of music education on children's overall life satisfaction and on satisfaction in specific areas, namely satisfaction with the class, satisfaction with friends, and satisfaction with the school. It analyzes music education in the classroom where fifth and sixth grade students have one additional hour of music education per

week. The project that is studies is called “klasse.im.puls” and it promotes the establishment of musical training in secondary schools in Bavaria. The program was implemented with the intention to give every child the opportunity to learn how to play an instrument. Additional positive outcomes were expected: an increase in self-confidence and social competence, as well as a reduction in violent behavior.[1] The project was supervised by the FAU Erlangen-Nuremberg Music Teaching Department in collaboration with the Bavarian Ministry of Education. My analysis focuses on the change in the overall life satisfaction as a consequence of participating in the music project. Further outcome variables are of interest: Satisfaction with friends, satisfaction with the class, and satisfaction with the school. I will approach the problem using a multilevel model that accounts for differences among students and differences among schools by introducing a hierarchical structure where students represent level 1 and schools level 2. Data are available for different points in time. Therefore a third level, time periods are included. To estimate the treatment effects, Bayesian analysis is the method of choice. Using Bayesian inference allows to make predictions on the outcomes of future projects of this kind. As the KIP project is still continuing to be applied in many schools over Bavaria, the predicted effects also carry important information.

I am particularly interested in life satisfaction as an outcome of the music project because there is the idea of higher values of LS coming with many benefits also in later life. Among other positive correlates, adolescents reporting very high levels of LS are less likely to be affected by depression, anxiety, negative affect and social stress compared to adolescents with very low life satisfaction. Also, they achieve higher SEAs and demonstrate higher mean scores of school satisfaction (Gilman & Huebner, 2006, p. 316; Proctor, Linley, & Maltby, 2009, p. 928). These results go in line with a study by Suldo & Huebner (2004, p. 94) that shows that LS could be a moderating variable in predictions of the development of psycho-pathological behaviors. Low life satisfaction may be an indication for externalizing behavior problems in the future. When life satisfaction is on a higher level, those behavior problems are less likely to occur (Suldo & Huebner, 2004, p. 100) The authors conclude that life satisfaction might operate as a buffer against the development of subsequent externalizing behavior problems “in the face of stressful

[1] For more information:

life events" (Suldo & Huebner, 2004. p. 101). Kim, Conger, Elder, & Lorenz (2003) add to the discussion that externalizing behavior problems in turn lead to more stressful life events. That reciprocal interrelation of stressful life events and externalizing problems (reported as delinquent behaviors) leads to unhealthy dynamics of increasing stress and behavior problems. If higher LS leads to better coping mechanisms with stressful life events, these dependencies could be reduced. LS is also positively correlated with children having higher measures of self-esteem, internal locus of control, and extraversion (Huebner, 1991a, p. 107). These features help in building a solid foundation for later life. On the other hand, dissatisfaction with life is associated with adolescents having poor mental health (like anxiety and neuroticism) and physical health and being exposed to a higher risk of considering or attempting suicide (Huebner, 1991a, p. 107; Valois, Zullig, Huebner, & Drane, 2004, p. 94). Furthermore, Zullig, Valois, Huebner, Oeltmann, & Drane (2001, pp. 284–185) show that adolescents reporting low levels of overall life satisfaction are more likely to use drugs and alcohol earlier in life and in higher amounts than adolescents with medium or high life satisfaction.

Considering the statement of Höppner, one would expect the project to positively effect students' life satisfaction. Evidence for this relation would lend credibility to the project and support its continuation. It could also stress the importance of music education and signal the Ministry of Culture to keep music education in the curriculum and work on its implementation in federal states alongside Bavaria. My thesis is a first try to reveal some information about the influence of the KIP project on the students. However, regardless of the findings, there will be no discussion on the reasons of any observed effects from a music education background. The analysis will be strictly observable and any interpretation of the results must be left to music and education experts.

The structure of my thesis is as follows: At first there will be an overview of the current literature on music effecting students' lives in Chapter 2. Next, in Chapter 3 the project and the data set will be explained along with descriptive statistics and a detailed diagnostic on pre-treatment differences in the treatment group compared to the control group. The data section also critically discusses challenges arising with measuring self reported life satisfaction in children. The

identification strategy is presented in Chapter **??**, following the estimation in the 5th chapter. Finally, Chapter 6 shows the results and Chapter 7 concludes and discusses.

2 The role of music in childrens' lives

With regard to the effect of music on students' lives, most of the researchers are interested in academic outcomes and intelligence among students who are actively involved in music. Generally, there is the predominant perception of a positive link between music and cognitive abilities. Osborne, McPherson, Faulkner, Davidson, & Barrett (2015, p. 14) observed improved math skills and higher subjective well-being scores in children that were part of a music project. He also found them to have a better self-control over impulsive behavior. Yang (2015) (p. 385),Wetter, Koerner, & Schwaninger (2008) (p. 372) , Hille (2014) (p. 62), and Guhn, Emerson, & Gouzouasis (2019) (p. 316) present evidence that children playing music have better grades at school. But this conclusion is not very meaningful as all of these studies follow a correlational design and do not allow for causal inference. Nonetheless, the very optimistic and as we will see not quite realistic belief has prevailed that playing music causes children to achieve better results at school. This conclusion needs to be revised. In an extensive review of the available evidence concerning associations between music and cognitive abilities, the picture is not so clear any longer (Schellenberg & Weiss, 2013). Small associations between music training and mathematical ability in correlational and quasi-experimental studies might result from individual differences in general intellectual ability (p. 527). The available evidence simply indicates that high-functioning children (i.e., higher IQ, better performance in school) are more likely than other children to take music lessons and to perform well in mathematics and other tests of cognitive ability (p. 534). This fits with the Hille (2014) study where the outcome difference in cognitive skills between musically active and inactive children reduces greatly when holding constant observable characteristics. An other plausible interpretation of study outcomes that fail to detect a causal relationship comes from Wetter et al. (2008). He points out the possibility that the relationship is explained by affluent parents being more likely to afford music lessons for their child and thus, the socio-economic background may be the cause of higher performance at school. Despite the weakness of the above studies to draw causal inference, there is slight

evidence that there may be a causal direction *from* music training *to* cognitive abilities in a study by Schellenberg (2004). He compares two treatment groups who receive piano lesson and voice lesson respectively to a control group in which the children have drama lessons. Random assignment to the different conditions allowed for inference that music lessons caused small increases in cognitive abilities (namely larger increases in full-scale IQ). However, this does not preclude the possibility that high-functioning children are more likely driven to play music. The misconception of music being a predictor for academic achievement is also discussed by Southgate & Roscigno (2009) (p. 17) who states that music is rather a mediator, to some degree, of family background and student status. Results from a meta-analysis, suggest that music training does not reliably enhance children and young adolescents' cognitive or academic skills, and that previous positive findings were probably due to confounding variables, such as placebo effects and lack of random allocation of participants (Sala & Gobet, 2016, p. 64). The effect size was reduced in studies that applied a proper study design (random allocation of participants to the treatment group and comparison to an active control group). With respect to the mathematical outcomes, the only study comparing a music training to an active control group and with random allocation of the participants to the group (Mehr, Schachner, Katz, & Spelke, 2013) found a negative effect size. These considerations are in clear contrast to the popular perception that music training enhances any non-music related cognitive skill.

However positive outcomes due to music could be found outside academics. The Norwegian psychologist and musicologist Even Ruud who in 1978 was the head of the first music therapy training performed extensive studies on music and identity. He states that cultural activities, explicitly music, can "contribute to a feeling of quality of life and the subjective sense of health." (p. 96). In an experimental study by Costa-Giomi (2004) a positive causal effect was detected of piano lessons on self-esteem [144] but she did not find an effect on math computation scores. One especially popular project that was conducted to improve children's lives is Venezuela's National Music Education Program "El Sistema". It is a large scale social music education program established by José Abreu in the 1970s. 300,000 children are equipped with instruments every year and receive regular after-school lessons and are playing in orchestras. The initial goal was to prevent children from using drugs and being involved in violence and crime which

was successfully achieved. Being in orchestras also enhanced social behavior of the students through greater concern for others and their own well-being (Uy, 2012, p. 13). However positive effects go far beyond keeping adolescents away from drugs and violence: El sistema teaches the participating students to "reflect and act upon the world in order to transform it" [7]. Playing in an orchestra means joy, motivation, teamwork, the aspiration to success [6]. The students pick up management and organizational skills and responsibility due to many roles and rules that they need to follow to sty in the program[p. 10]. Also, being in an orchestra gives the students the change to conceptualize themselves as part of something much larger and greater (p. 11) and they learn to express greater concern for others' and their well being (p. 13). However, it remains open if those positive effects are a result of being musically active or if they are induced by the natrue of an orchestra being a place of social interaction that positively affects childrens' lives. Anyway, El Sistema became internationally popular and was replicated in several countries. Osborne et al. (2015) reviewed the outcome of El Sistema inspired projects in Australia and found significantly higher subjective well-being scores in the participating group (p. 14). Students from the music program also had better self-control over impulsive behavior (p. 15). However, other studies come to contrary conclusions and fail to show a significant effect of music participation on well-being, social skills, emotional intelligence or self-esteem (Portowitz, Lichtenstein, Egorova, & Brand, 2009, p. 121; Schellenberg, 2011, p. 190 association). Again, findings are diffuse and therefore interpretation of any results needs to be done with caution.

The takeaway from this section is that causal conclusion must be done carefully and with adequate methods. This thesis is an attempt to do this and to complements existing literature by using different music indicators and by adjusting estimation strategies to provide further insight on the relationship between music and education. A multilevel model will help to draw correct inference by avoiding an overstatement of statistical significance. This often happens when failing to recognize hierarchical structures because then standard errors of regression coefficients will be underestimated. It also allows for inference beyond the groups that are observed because in a multilevel model, the groups in the sample are treated as a random sample from a population of groups. http://www.bristol.ac.uk/cmm/learning/multilevel-models/what-why.html

3 Data

In this section, I will first give more background information on the project itself, including the data collection process, before describing the data and variables individually. After general summary statistics, there is a section about handling potential issues arising with differences in the treatment and control group.

3.1 The project

The data come from the years 2012-2014 when KIP was conducted in six different secondary schools (Mittel- und Realschulen) in Nuremberg. Generally, there are different types of implementations of the KIP project (e.g. choir, brass band, string orchestra) among which the rock band model was implemented in all of the schools that are subject to this study. Probably due to parents' and students' music preferences, this type has been the most popular. In that concept, a class was typically divided into four bands, practicing concurrently in two rooms with one teacher each, talking turns by one band playing their instrument and another doing the vocals.

A school had to fulfill certain criteria in order to get support in the form of advanced training for music teachers and acquiring equipment in order to implement ensemble music class teaching. For example, the school had to make sure that there was a full-time music teacher who was qualified to carry out the project. That person was trained to lead a band class and attended yearly meetings with other music teachers to share experiences. This way, it can be assured that the treatment in each school looked pretty much the same and there were no noteworthy differences in the way the teachers arranged the music lessons in the band classes. Participating schools were supported in acquiring equipment and instruments to implement ensemble music class teaching. In each school was a band class and a control class. The band classes received three music lessons per week, whereas the control classes only received two lessons of regular music education. Music education in the treatment classes differed significantly from that in the control group. In the band classes, the teachers had a more practical approach and there was less theoretical music education than in the control group. The three music lessons per week were split into a) Instrumental instructions in small groups, b) ensemble playing, and c) general music

education (music theory or history). The latter was similar to how regular music education was implemented in the control group. Therefor not only the amount of music education is higher in the treatment group but also the way that music education is delivered differs a lot from normal music lessons. It is more practical and the equipment is of higher quality. The students played concerts in regular intervals and were thus able to share what they had learned.

In five different points of time, the students received the same anonymized questionnaire which they were asked to complete within about 30 minutes.

The assessments were performend in: wave 1 – Beginning of 5th grade (pre-treatment) \ wave 2 – Mid of 5th grade \ wave 3 – end of 5th grade \ wave 4 – mid of 6th grade \ wave 6 – end of 6th grade \

3.2 Variables

The sample that is studied has a hierarchical structure with three levels: Each student is observed multiple times, ideally at five points throughout the project if he was present on every assessment day. The different time periods are level one. The students themselves are level two and as the students are nested within schools, the third level is the school level. More details on the specific variables follow now:

3.2.1 Satisfaction

The term *life satisfaction* refers to a cognitive evaluation of a person's reaction to his or her life in contrast to *affect*, an ongoing emotional reaction. Combined, LS and affect yield subjective well-being (Diener, 2009, p. 71).

To measure life satisfaction and any area of satisfaction, I use an 11-point scale, with responds on a scale from 0 to 10 with "0" being completely unsatisfied and "10" being completely satisfied. Life satisfaction is the variable of interest in my thesis, alongside with Satisfaction with friends, Satisfaction with class, and Satisfaction with school. An extract of the questionnaire with the exact wording of the question can be found in the index (Figure **??**fig:questionnaire). Measuring life satisfaction on a 0-10-scale is quite common in that field of literature [Mellor, Stokes, Firth,

Hayashi, & Cummins (2008) Cheung & Lucas (2014)} but also 7-point scales are found a lot (Diener, Emmons, Larsen, & Griffin, 1985). For the model, the satisfaction measures are assumed to be cardinal comparable both across and within individuals across time. There is broad consensus in the literature to interpret satisfaction responds as cardinally scaled (Ferrer-i-Carbonell & Frijters, 2004). Mean values and standard deviations of life satisfaction and the different areas of satisfaction from wave 1 are presented in Table 1.

Table 1: Mean satisfaction

Variable	Mean	SD
Life satisfaction	7.30	2.56
Satisfaction friends	8.85	1.82
Satisfaction class	7.58	2.03
Satisfaction school	7.72	2.19

As the standard deviations are quite big, the life satisfaction reports are presented in Figure 1 for a better overview. As already cited above in Chambers & Johnston (2002), children tend to report values at the extreme ends of a scale. This is also what happened in the KIP sample: The histogram hows a striking portion of the students indicating to have a life satisfaction of 10. Also, a high amount of the students respond with a 5 on the scale. Almost 50% of the students assign themselves a life satisfaction of 5 or of 10. This might indicate the issue that children at that age are having a hard time understanding what the question asked them to do. It also suggests difficulties to make the transfer from their individual assessment of their life satisfaction to pair it with a number and so they just chose 5 randomly, hoping that it was neither “right or wrong”.

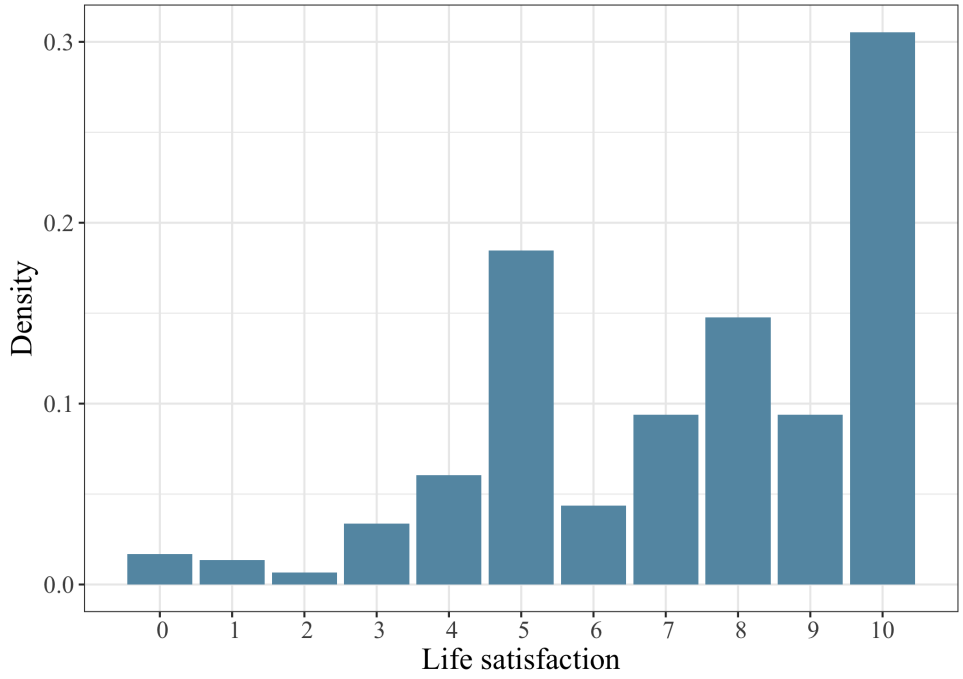

Figure 1: Distribution of life satisfaction in wave 1

3.2.2 Treatment

The treatment is represented by a binary variable, indicating whether a student participated in the treatment ("1") or not ("0"). The treatment is conducted on the class level. All students in the same class are assigned the same indicator of whether or not they attend the treatment. Students did not change the class. Treatment also appears in the data set as an interaction term with each time period in that a treatment effect can be measured. Multiplying the binary indicator of treatment with the dummy variable for each period gives four indicators with the variable indicating 1 if the observation was in the treatment group and was made in the respective wave and 0 otherwise.

3.2.3 Individual characterisitc background

In the first wave, students were asked to state their gender (Tick either *"Are you a girl/are you a boy?"*). This question did not appear in subsequent waves which leads to the issue that children who were not in school on the day the first questionnaire was handed to them, appear as NAs in the sample. The same happened for migration background. To avoid drop out of too many observations in the estimation process, gender and migration background were manipulated as

category variables and NAs were replaced by the category *unknown*. This leads to the following distribution of time-invariant individual characteristics: Both the gender and the migration background is not known for around 20% of the total of 390 students. The available data show that out of the remaining students, 45% are female and 36% have a migration background.

The fact that it is an unbalanced panel where not all the children attended each wave should be considered more thoroughly. In Table 2 , the number of students in each school and each wave is presented. It shows that the number of students is varying from one wave to another, most severely in class five. In that class, the number of participants of the questionnaire drops by a third in wave 5 compared to the first period. The overall effect of this decline on the whole sample is shown in Table 3. In the treatment group, the number of students decreases from 154 in wave 1 and is the lowest in wave 4 with 137 participants. In the control group, there are 144 to start with and only 134 in wave 5. This could be a reason to exclude school five from the observation completely as the attrition rate in the remaining schools is not that alarming. In total there were 1500 students observed throughout the five waves, however I excluded the students that did not report their satisfaction. If I had left them in the sample, this would have lead to problems in the posterior predictive checks because when the 'brmsfit' fits the model, any N/As are excluded automatically. Without these observations, there are 1429 left.

Table 2: Number of observations by school

School	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Sum
1	47	40	46	48	52	233
2	53	49	52	48	49	251
3	49	46	48	52	52	247
4	47	45	47	50	49	238
5	56	57	57	43	37	250
6	46	42	44	38	40	210
Sum	298	279	294	279	279	1429

As a third individual background variable, I included if a student had some musical experience before the project had started. I expected students who already played an instrument before or received private music lessons would be more drawn to the treatment class. Also, as I want to

find out if the project made students more satisfied, any previous effects of music on satisfaction should be minimized.

Table 3: Number of observations by treatment

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Treatment group	154	142	146	137	145
Control group	144	137	148	142	134

3.3 Pre-treatment differences

The treatment groups were not randomized, but formed by choice. This makes it a quasi-experimental study where the decision who attended the treatment group was made before the entrance to grade 5. Parents could choose to put their child into a music class. Therefore, the experiment is not fully randomized. Children who were already practicing music in the first place might have been more likely to enter the music class. Also, the decision can be a a consequence of parental and socio-economic background of the student. To have a better idea of possible differences between the treatment and the control group, I computed standardized mean differences:

Table 4: Standardized mean differences

	treated = 1 vs 0
Life satisfaction	0.15
Female	0.16
Satisfaction friends	0.16
Satisfaction class	0.16
Problems in class	0.04
Exclusion in class	0.12
Grade german	0.10
Grade math	0.13
Grade music	0.19
Hobby music	0.54
Duration music	0.45
Instrument	0.10
Musically active	0.16
Migration background	0.15

The standardized mean differences are not a reason for concern in this setting. For most of the observed variables, there are no notable differences between the treatment group and the control group. However, as already suspected, there are differences in prior musical knowledge. Among the treated, there are relatively more children who have music as a hobby before the project has even started. Of the 87 students who answered with "yes" to the question if music was their hobby, only 24 were assigned to the control group and the remaining 63 students attended the band classes. Also, the duration of making music is very different in the two subgroups. In the Treatment group, 56% are practicing at least 30 min, while in the control group that share is only 35%. However, the model addresses this issue by applying a difference-in-difference approach to estimate the causal effect of additional music lessons on students' well-being. Possible differences in the control group and the treatment group are not a problem for inference.

3.4 Measuring LS in children

To investigate ones quality of life, research uses both objective and subjective indicators. Typical objective indicators are the income levels, crime rates, and access to medical services – measures that are external and quantifiable. Subjective indicators on the other hand comprise subjective evaluations of ones' individual life circumstances (Gilman & Huebner, 2000, p. 178). Only a modest relation between both measures was found which indicates that each approach carries unique information that are relevant for a comprehensive understanding of overall life quality (Veenhofen, 1996). Over time, measurement methods for subjective indicators have evolved, leading to substantial growth in the life satisfaction research. However, for a long time, most of the measurements were designed to assess adults' life satisfaction. Only recently, investigating correlates of life satisfaction in adolescents has begun. One reason why life satisfaction research in children was put off for a long time is probably that measuring life satisfaction in children is more challenging than for adults. Instruments for assessing children's subjective life satisfaction reports have been less intensively developed which is probably due to the fact that a Likert-type ratings scale is more difficult to use for younger children than for adolescents (Chambers & Johnston, 2002, p. 28). For example, younger children or children with poorer readings skills are less able to respond appropriately to negative items on questionnaires and this effect biases

the interpretation of children's responses (Marsh, 1986, p. 45). It is also common for children rating their subjective life satisfaction to show elevated extreme scores. As children become older, this tendency subsides and they are more capable of providing graded ratings in between the two extremes. These results have potentially substantial implications for the interpretation of self-report ratings form children *this tendency might have an erroneous and invalid impact on the interpretation of children's self-reports*(Chambers & Johnston, 2002, pp. 33–34). When dealing with self-reported life satisfaction in children, it is crucial that the respective child fully understands the question in order to give a valid response (Gluskie, 2012; Tomyn, Fuller-Tyszkiewicz, Cummins, & Norrish, 2016). One must make sure that a child is old enough to know how to use a satisfaction scale. This requires children to distance themselves from the current situation, cognitively evaluate their life satisfaction (considering all relevant areas of life) and rate the degree to which certain items on the scale apply to them. This requires abstract thinking, which children develop in early adolescent years (10-14 years) (Gluskie, 2012; Piaget, 1955, 1969). Gilman & Huebner (2000) have reviewed five different (both unidimensional and multidimensional) measurements explicitly developed to asses adolescents' life satisfaction.[2] The authors evaluated those measures in terms of validity and reliability and found all of the scales to be appropriate for research with adolescents [181-188]. The demographic characteristics of the available samples show that all of the adolescents observed were older than 12 years. Therefor it remains unclear if children younger than that age are able to report valid satisfaction. An other instrument was developed by (Cummins & Lau, 2005), the Personal Well-being Index (PWI-SC). Again, studies demonstrated reliability for this instrument as well (Casas & Rees, 2015; Casas et al., 2011; Tomyn & Cummins, 2011; Tomyn, Stokes, Cummins, & Dias, 2019). But also in those studies, all of the adolescents were at least 12 years of age, mostly even older. There is only very little evidence on the psychometric properties of the PWI-SC for children below the age of 12. One of them is González-Carrasco, Casas, Malo, Viñas, & Dinisman (2016, p. 70) who applied the instrument for children as young as only 9 years and also found adequate fit of the data. On the other hand, Tomyn et al. (2016) conducted a study with children aged

[2]The Students' Life Satisfaction Scale (Huebner, 1991b), the Satisfaction With Life Scale (Diener et al., 1985), the Perceived Life Satisfaction Scale (Adelman, Taylor, & Nelson, 1989), the Comprehensive Quality of Life Scale – School Version (Cummins, 1997; Gullone & Cummins, 1999), and the Multidimensional Student's Life Satisfaction Scale (Huebner, 1994)

10-12 and concluded that subjective well-being data of children must be interpreted with caution. They also show that response bias towards the extreme positive end of a scale is higher with decreasing age. The authors do not recommend using the PWI-SC for children younger than 12 years. As for the specific sample, the PWI-SC did not serve as a valid instrument for measuring the SWB.

In conclusion, measuring life satisfaction in children is more challenging than for adults and is still in progress. It is advisable to check the validity and reliability of their data when testing children. Considering, the majority of the children from the KIP project are 10 year sold in wave 1 (72% of those participating in wave 1), they might be just too young to give valid responses when asked about their life satisfaction. This must be kept in mind when evaluating the results.

4 Identification strategy

As I already mentioned in the introduction, I will make use of a multilevel model, precisely a varying-intercept and varying-slope multilevel model. A multilevel model is a good choice for drawing causal inference but also for prediction and descriptive modeling (Andrew Gelman, 2013, p. 6). In the following, I will summarize important aspects of the methodology of those models and explain how multilevel models are best to navigate between underfitting (too few parameters included in the model) and overfitting (too many parameters included in the model). In the second part, I will give an introduction to Bayesian analysis that is the ideal method for multilevel modeling because of its underlying assumption that all parameters are random quantities.

4.1 Multilevel modeling

Since the sample is structured hierarchically, with students nested within schools, observed in several time periods, using a simple nonhierarchical model would be inappropriate. In practice, simple nonhierarchical models with few parameters generally cannot fit large data sets accurately, whereas with many parameters, they tend to overfit such data gelman 101. Though fitting the data well, often seems to be the goal of data analysis, these models are usually very poor in making predictions for new data (Andrew Gelman (2013), 101). (fit to sample always – not

true of multilevel models – improves as we add parameters.) In the context of the intervention under study, each school may well have a different average life satisfaction tendency or respond differently to the treatment. The data should benefit from a model that expects such variation.

A multilevel model is a linear model in which parameters are given a probability model. Mcelreath (2020) (p. 14) suggests to think of parameters as placeholders for a missing model. That probability model itself has parameters called hyperparameters which are also estimated from the data. In other words, there are parameters for parameters for parameter and so on. Technically, there is no limit to the number of levels, however infeasible computation and ability to understand the model are in practice a restriction. Though multilevel models are more complicated, it is worth using them because they produce better estimates mcelreath, 15. In the context of the observed study, a multilevel model works extremely well because of several reasons (elreath 15):

- Adjusting estimates for repeat sampling – having more than one observation arising from the same student, a traditional, single-level model may be misleading
- Adjusting for imbalance in sampling – some students are sampled more than others, again, a single-level model could be misleading in this case
- Studying variation – I am interested in the variation of the treatment effect among schools. Multilevel models are of big help in this case because the model variation explicitly.

Multilevel models involve predictors at different levels of variation. In the KIP setting that means that it is possible to measure different effect sizes throughout time in each of the schools. It is likely to observe different response behaviors among the schools. Depending on the school, students might react differently to the project. This is an important investigation to find out if the intervention was more effective in some schools than in others. One school might have done something explicitly well or had to face challenges other schools did not. Learning from these findings could possibly help to improve the project in the future.

In a classical regression, assigning varying effects to each group is done by using interaction terms or running separate regressions for each subgroup of the sample. However, this has the disadvantage that estimates can be very imprecise, in particular when there are only few

observations per group. Estimating parameters for each group separately is referred to as *no-pooling* because each group is observed independently without considering the rest of the sample. On the other hand, ignoring the hierarchical structure and *completely pooling* the information is also not recommended, because existing effects might disappear as illustrated in Figure @(fig:lsat-vs-time): if solely taking the mean over all schools (solid purple line), there is almost no change in life satisfaction through time. But when each school taken separately (dashed blue lines), there are clear dynamics that substantially differ from each other. While in some schools average life satisfaction is decreasing with different rates, in other school opposite trends are visible.

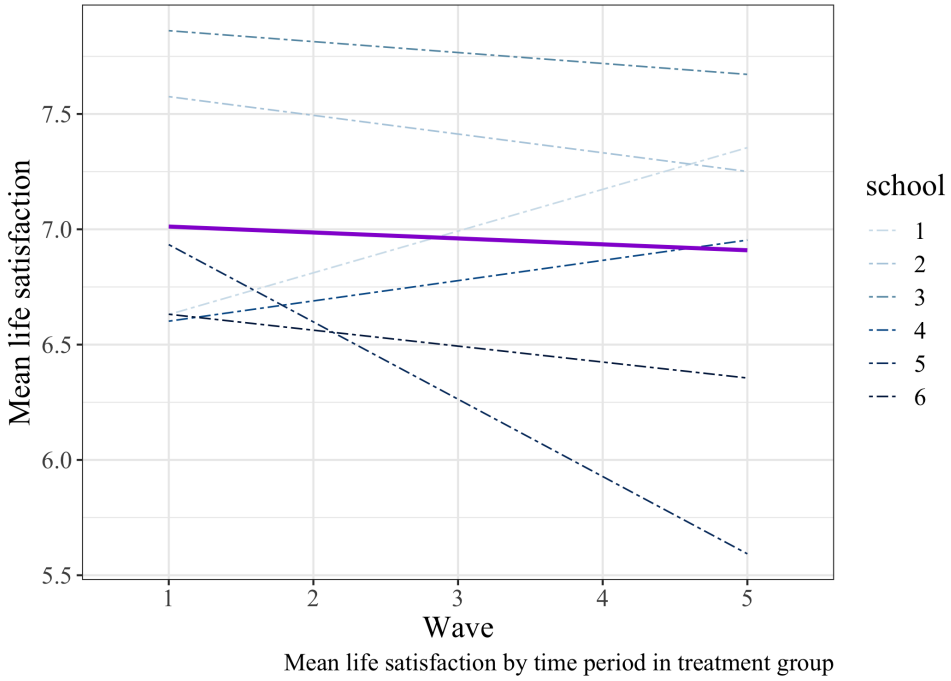

Figure 2: LS in treatment group

Both approaches, no-pooling and complete pooling, have downsides for different reasons but they can be useful preliminary estimates, eventually leading up to what is known as partial pooling that comes out of a multilevel analysis. Therefore I will illustrate both the extremes, the complete pooling and the no-pooling approach in the following, before I get to the details of partial pooling. The complete pooling and no-pooling model in comparison to the partial pooling model will be presented in a very simplified way. Life satisfaction will be explained with a varying-intercept model without any predictors. The purpose is to emphasize the characteristic

features of the three approaches without unnecessary complexity in the model. In this example, α and σ are given weakly informative priors and the models are estimated using `rstanatrm` functions `stan_glm` and `stan_glmer`.

The simplest approach of the three is the pooled model, where life satisfaction is modeled as independent and identically distributed draws from a common distribution. The intercept α, the average life satisfaction in this case, is the same for each school.

$$\begin{aligned} LS_i &\sim \text{Normal}(\alpha, \sigma) && \text{for } i \in 1, ..., n \\ \alpha &\sim \text{Normal}(0, 10) \\ \sigma &\sim \text{Exponential}(1) \end{aligned}$$

In the complete pooling approach, the population of schools is assumed to be invariant. This ignores the fact that schools are different and assigns the same intercept to each of the schools. If done like this, there is the risk of ignoring important variation in how schools correspond to the treatment (Mcelreath, 2020, p. 416). The total sample mean underfits the data, meaning that the model is insensitive to the details in the data and is learning too little from it. Ignoring the group-level variation is very likely misleading.

Whereas complete pooling ignores variation between schools, the no-pooling analysis overstates it and gives a different mean to each school:

$$\begin{aligned} LS_i &\sim \text{Normal}(\alpha_{j[i]}, \sigma) && \text{for } i \in 1, ..., n \\ \alpha_j &\sim \text{Normal}(0, 10) && \text{for } j \in 1, ..., m, \end{aligned}$$

where $j[i] \in 1, ..., m$ is the school of student i. Although the α_j is drawn from the same prior distribution, it has fixed parameters and thus no information is shared between observations in different schools. This makes the model very sensitive to the details in the data, you could say it is learning too much from it. The no-pooling approach includes the assumption that the schools are completely different and one school cannot tell anything about the other schools. This is equivalent to the variation among schools being infinite. But in reality, though schools

are different in some ways, they are also very similar so that each school helps to estimate the treatment effect in other schools. One should choose a model that incorporates the idea of learning from other groups while still accounting for systematic differences among the groups. This is exactly what partial pooling is attempting.

Partial pooling represents a compromise between the two extremes of excluding a categorical predictor from a model (complete pooling), or estimating separate models within each level of the categorical predictor (no pooling). For this simplified model with no predictors, the multilevel estimate for a given school j can be approximated as a weighted average of the mean of the observations in the school (the unpooled estimate, $\bar{y}_j$) and the mean over all schools (the completely pooled estimate, $\bar{y}_{\text{all}}$):

$$\hat{\alpha}_j^{\text{multilevel}} \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{\text{all}}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}, \tag{1}$$

where n_j is the number of students in school j, σ_y^2 is the within-school variance in LS, and σ_j^2 is the variance among the average LS levels in the different schools. For keeping the example simple, I assume the within-school variance to be constant among the schools (gelman hill 254). The weighted average describes the ratio of information available about the individual school and the average of all the schools. Depending on that ratio, the weighted average converges to one of the two extremes, either no-pooling or complete pooling.

- A school with only few students carries less information, and the weighting pulls the multilevel estimate closer to the overall state average. If $n_j = 0$, the multilevel estimate is simply the average, $\bar{y}_{\text{all}}$.
- Averages from larger schools carry more information, and the corresponding multilevel estimates are close to the school average. If $n \to \infty$, the multilevel estimate is simply the country average, $\bar{y}_j$
- In intermediate cases, the multilevel estimate lies between the no-pooling and the complete pooling estimate.

This means that in the partially pooled model, still each school has its own mean values but with the difference that these school-means share a prior which has its own parameters.

$$LS_i \sim \text{Normal}(\alpha_{j[i]}, \sigma) \qquad \text{for } i \in 1, \ldots, n$$
$$\alpha_j \sim \text{Normal}(\mu, \tau) \qquad \text{for } j \in 1, \ldots, m$$

gelman hill p 253/254 We could also write the model with the school-level average in the mean equation for y, and the α_j values distributed around the country level average, γ.

$$LS_i \sim \text{Normal}(\alpha_{j[i]}, \sigma) \qquad \text{for } i \in 1, \ldots, n$$
$$\alpha_j \sim \text{Normal}(\gamma, \tau) \qquad \text{for } j \in 1, \ldots, m$$
$$\tau \sim \text{Exponentional}(1)$$

After fitting the above models, I extracted the estimates of each of them and plotted them in Figure 3. (https://jrnold.github.io/bayesian_notes/multilevel-models.html)

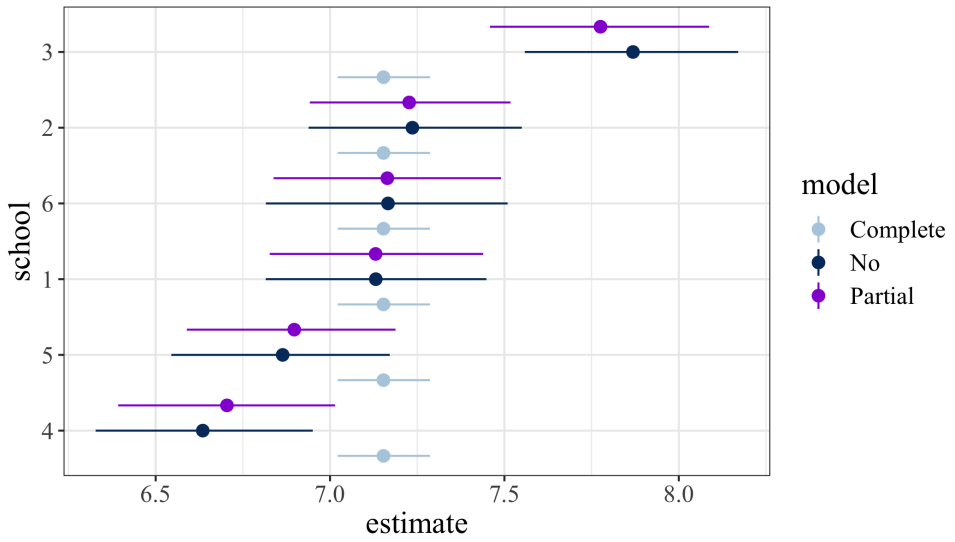

Figure 3: Mean

It shows clearly that when partially pooling information, both similarities and differences among groups are comprised. Partial pooling uses all the data available to perform inference for groups instead of just using local information which is especially useful when there is only a small

number of observations in each group. The group estimates are supported to the extend of the whole data set and are not being limited to the number of observations per group. These estimates are less underfit than the grand mean and less overfit than the no-pooling estimates. As a consequence they tend to be better estimates of the true per-school means (mcelreath p423). Visually, this is shown by the purple dots (partial pooling), always being somewhere between the no-pooling estimates and the partial pooling estimates and being pulled towards the grand mean. The difference between the classical no-pooling estimates and the results from partial pooling is not very drastic. This is probably because the group-level variation is relatively small and that makes the multilevel model reduce to a classical regression with no group indicators. Also, the number of groups (6) is quite small which might be a reason why group-level variation cannot be estimated accurately. Still, it is worth the effort of expanding a classical regression in this way when adding the third level for the time periods later in the estimation. Generally, partial pooling will be especially helpful when there are only few observations in one school because then the no-pooling estimate will be overfit significantly. The difference between no-pooling and partial pooling becomes smaller, the more observations there are per cluster. The partial pooling produces noticeably better estimates and therefore will be used for my estimation. There are different tools to create a model that partially pooles information. In practice the variance parameters together with the α_j's are either estimated with an approximated program such as `lmer()` or using fully Bayesian inference. Since multilevel models, have a natural Bayesian representation (as they are models with multiple levels of uncertainty), a Bayesian data analysis fevourable mcelreath, 14, gelman und hill, 143

4.2 Bayesian inference

To set up what will follow in Chapter 5 which is about the explicit estimation, stylized components of a Bayesian statistical model are explained and formally written down. The notation is based on Gelman etal 2020(chapter 1).

The interest of Bayesian data analysis lies in drawing statistical conclusions about unobservable vector quantities of population parameters, denoted by θ, and making predictions about unobserved data $\tilde{y}$. These conclusions are made in terms of probability statements that are conditional

on the observed data y. So the goal is to find out what is $p(\theta|y)$ and $p(\tilde{y}|y)$. For every parameter, there needs to be a distribution of prior probability, its prior $p(\theta)$. This is the initial plausibility assignment for each possible value of the parameter from which the Bayesian model is updating with the data y.

- $p(\theta|y)$:

  To make a probability statement about θ given y, a joint probability distribution for θ and y is needed in order to apply Bayes' rule. The joint probability function is the product of two densities: the prior distribution $p(\theta)$ and the sampling distribution $p(y|\theta)$. Then the updating of all of the prior distributions lead to the posterior distribution

  $$p(\theta|y) = \frac{p(\theta,y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)},$$

  where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, and the sum is over all possible values of θ (or $p(y) = \int p(\theta)p(y|\theta)d\theta$ in the case of continuous θ). Omitting the factor $p(y)$, which does not depend on θ and, with fixed y, can thus be considered a constant, the result is the unnormalized posterior density:

  $$p(\theta|y) \propto p(\theta)p(y|\theta).$$

  This expresses that the posterior is proportional to the product of the prior and the likelihood.

- $p(\tilde{y}|y)$: To make predictive inferences, a similar logic is applied: The distribution of the unknown but observable y is

  $$p(y) = \int p(y,\theta)d\theta = \int p(\theta)p(y|\theta)d\theta,$$

  which is referred to as the prior predictive distribution. The prior predictive distribution is

  - not conditional on previous observations of the process
  - a distribution for a quantity that is observable.

From the same process, $\tilde{y}$ can be predicted after observing the data y. The distribution of $\tilde{y}$ is called the posterior predictive distribution. The posterior predictive distribution is

- conditional on y
- a prediction for an observable $\tilde{y}$:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y},\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta,y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta. \end{aligned}$$

The equations express the posterior predictive distribution as an average of the conditional predictions over the posterior distribution of θ. Given θ, y and $\tilde{y}$ are assumed to be conditional independent. Having decided on a specific probability model, the data y affect the posterior inference *only* through $p(y|\theta)$, the likelihood function.

To wrap up the above: Bayesian inference refers to statistical procedures that model unknown parameters (and also missing and latent data) as random variables. It starts with a prior distribution, beliefs about the parameter before having examined the new data, and updates it with the likelihood of the data, yielding a posterior distribution which is used for inferences and predictions (gelman and hill p143) (–> der absatz ist kopiert, aber ich finde keine guten Worte. Kannst du da ein bisschen umformulieren?)

5 Estimation

For model estimation I decided to use the `brms` package by Bürkner (Bürkner (2017)). It implements Bayesian multilevel models in `R` using the probabilistic programming language `Stan`. It allows for individual prior specifications and facilitates incorporating prior distributions that reflect the users' beliefs. Different numerical techniques for computing the posterior distribution. In the `brms` package, Markov chain Monte Carlo (MCMC) is the technique used to fit the data. MCMC does not approximate the posterior distribution directly but draws samples from the

posterior. This leads to a collection of parameter values of which the frequencies correspond to the posterior plausibilities (elreath, 45).

5.1 The model

The goal of my estimation is to identify the causal effect of the treatment on the well-being of the participating students. But the development of the satisfaction measures could have altered due to other variables, besides the introduction of the music project. Therefor, I make use of the Difference-in-Difference (DD) method and compare the treatment group to the control group. This helps removing the confounding effect after the intervention period and arrive at the real causal impact. The underlying common trend assumption implies that both groups follow the same trend in the pre-treatment and the treatment group would have the same trend as the control group in the the-post intervention period in the absence of the treatment. Then the difference in the change of life satisfaction is the actual treatment effect.

This way, the problem of significant pre-treatment differences is avoided because DD-regressions allow for systematic differences due to group-specific time-invariant characteristics. The model I am using to fit the data includes varying intercepts on the school level and on the student level and varying slopes for each school. This assumes that schools have similar features but it also recognizes that unmeasured aspects on the students or schools can lead to variation in treatment effects. It allows for differences in how individuals or groups respond to the same circumstance. The specific regression equation is:

$$y_{ijt} = \beta_0 + \beta_{1j[i]} T_i + \sum_{l=2}^{5} \delta_{lj[i]} (T_i \cdot 1[l = t]) + \sum_{l=2}^{5} \lambda_l \cdot 1[l = t] + \mu_{j[i]} + \alpha_i + \varepsilon_{ijt}, \qquad (2)$$

where y_{ijt} is the outcome of pupil i in school j at time t. T_i is a binary indicator of whether or not pupil i attends a class with additional music education. δ_l is the period-specific treatment effect at time l. λ_l are period fixed effects. We model school-specific effects and individual-specific effects by including μ_j and α_i, respectively. ε_{ijt} is an error term.

The multilevel model in equation 2 allows the period-specific coefficients δ_l (i.e. the coefficient on the interaction terms between the treatment group indicator, T_i, and the dummy variables for

the periods) to vary across school. In doing so, we are able to examine potential heterogeneity in the treatment effects of the intervention across schools, which may provide insights into contextual factors of schools that promote or hinder pupils to benefit from the intervention. Furthermore, the model includes random intercepts at the level of schools, μ_j, as well as at the level of pupils, α_i, represents an individual-specific effect. The random intercepts capture that average well-being may vary across schools and/or across pupils. The multilevel model takes into account that our data has a hierarchical structure that is defined in terms of three clusters. We have repeated measurements of outcomes (level 1) for the same pupils (level 2). The pupils are nested within schools (level 3). The hierarchical structure implies, for instance, that pupils within a school are more similar than pupils from different schools.[3]

5.2 Prior distribution

As stated above in Section @ref{sec:identification}, in a Bayesian framework parameters are not point estimates but they have distributions. The specification for these parameters are called prior distribution because they must be specified *before* the model is fit to data and it assigns them a probability to every possible value of each parameter to be estimated. If the specification is done properly, for all parameters in the model, a Bayesian model yields a joint prior distribution on parameters and data, and hence a prior marginal distribution for the data (Gabry p5). This process can be described as the prior distributions and data interacting to finally produce the posterior distribution. The posterior can be seen as a compromise between the prior distribution and the likelihood function. Therefore, the choice of the prior should be done carefully since it can determine the outcome severely. After identifying the type of data being described, creating a descriptive model with meaningful parameters, and defining a likelihood function, the next step is to establish a prior distribution over the parameter values (kruschke 110). The prior distribution indicates the believes about the distribution of each parameter without knowing the specific data. Depending on the choice of prior and the respective data set, the "compromise" favors one of

[3]School-level effects and pupil-level effects may be modeled by including separate sets of indicator variables for each school, each class, and each pupil. However, these effects are not identified separately in a balanced panel because the school-level effects are a linear combination of the individual-level effects. Hence, μ_j and α_i cannot be estimated together because of perfect collinearity. Furthermore, estimating a potentially large number of parameters is inefficient and generalizations to the population of pupils are not straight forward (**???**).

them over the other. Flat priors or super-vague priors ($\mathsf{Exp}(0.5)$) are usually not recommended. They lead to a posterior distribution that is mainly influenced by the data which means that though the data set is very well described by the model, predictions are very poor. The model learned about the distribution from the data only because the prior did not carry any information. However if the prior has strong beliefs about the distribution of the parameters, meaning the prior distribution is sharply peaked, and there are only relatively few data, the posterior distribution is more influenced by the prior. Generally, weakly informative priors ($\mathsf{N}(0, 10)$) are recommended because they help by providing a very gentle nudge towards reasonable values of the parameter (mcelreath p 299). If there is a reasonable large amount of data, they will dominate while the prior becomes less important. In case of not very meaningful data though, the "weakly informative prior" makes up for it by strongly influencing the posterior inference. This relation is illustrated in Figure @ref{fig:flat-peaked}. The upper row shows the posterior that appears when having a prior distribution $LS \sim \mathsf{Exponential}(0.5)$. This prior gives barely any useful information about the distribution of life satisfaction among the population. The shape of the posterior is identical to the one of the likelihood. The opposite is the case when the prior is highly informative. A normal distribution with parameters of $\mu = 5$ and $\sigma = 1$ suggests that life satisfaction is very narrow around the mean of 5. The prior is so sharp that the posterior distribution is noticeably influenced by the prior. Bayesian models with proper priors are generative models (Gabry p5)

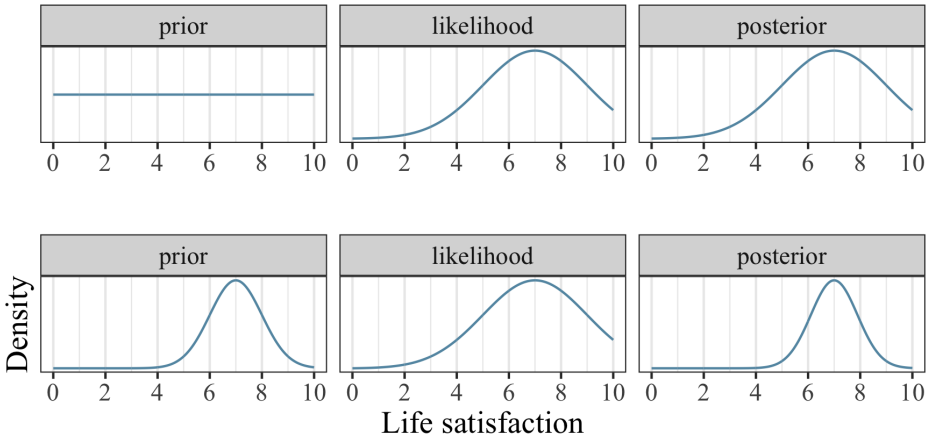

Figure 4: Flat vs peaked prior

A good prior also tackles the problem of overfitting in favor of making better predictions. As discussed in Section **??**, overfitting happens when the model is too sensitive to the sample. Choosing a flat prior means that every parameter values is equally plausible and therefore the posterior encodes as much of the likelihood function as possible. Overfitting can be avoided by choosing a skeptical prior. The term skeptical refers to the prior being skeptical about values outside the range of parameter values that are not reasonable. The most common skeptical prior is the regularizing prior which effectively reduces overfitting while still allowing the model to learn about the regular features of a sample rlreath, 219. To evaluate the chosen prior, a prior predictive check can be helpful. It examines what data sets would be consistent with the prior. Prior predictive checks will not be calibrated with actual data, but extreme values help diagnose priors that are either too strong, too weak, poorly shaped, or poorly located. This method is useful for understanding the implication of a prior. It simulates predictions from a model, using only the prior distribution instead of the posterior distribution. Once priors are chosen, they imply a joint prior distribution of individual life satisfaction (elreath p85). This procedure is based on choosing priors conditional on pre-data knowledge of the data, on general facts so to say (elreath, p100). Only when this step is done, the model is applied to the data. To decide on proper priors, I plotted the influence of different priors on the distribution of the outcome variable, life satisfaction. This helped me deciding on priors that influenced the outcome in a way that the largest part of the distribution was on a reasonable scale while extreme values are still possible. To begin with, the prior for the estimate of the population intercept is $\mathsf{N}(7, 2)$ (Figure @ref(fig:chosen-priors, upper left plot). Life satisfaction as it is measured in this setting, spreads on a range from 0 to 10 with most of the responses in the upper third. For the standard deviation I chose $\mathsf{Exp}(0.5)$ which is wide enough for all values on the satisfaction scale to remain possible. For the population-level and group-level effects, I came to the conclusion that a weakly-informative prior of $\mathsf{N}(0, 1)$ is best. To show this, I plotted predicted values from different prior predictive distributions:

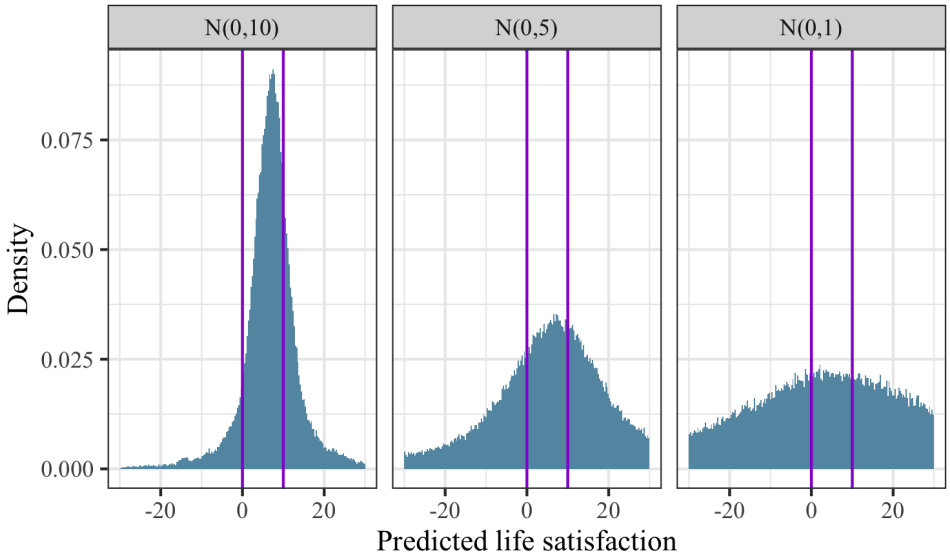

Figure 5: Prior predictive distribution

In each of the plots 100,000 values are drawn solely from the priors, ignoring the likelihood, which allows among other to generate samples from the prior predictive distribution. On the left, the prior for the population- and group-level estimates is noramally distributed with $\mu = 0$ and $\sigma = 1$. The parameters for the prior that produced the middle polot are $\mu = 0$ and $\sigma = 5$, and for the right plot $\mu = 0$ and $\sigma = 10$. The purple lines represent the lowest and the highest value from the life satisfaction scale. The prior that I used for the final estimation is the one on the left because in that plot, most of the density is within the limits of the scale.

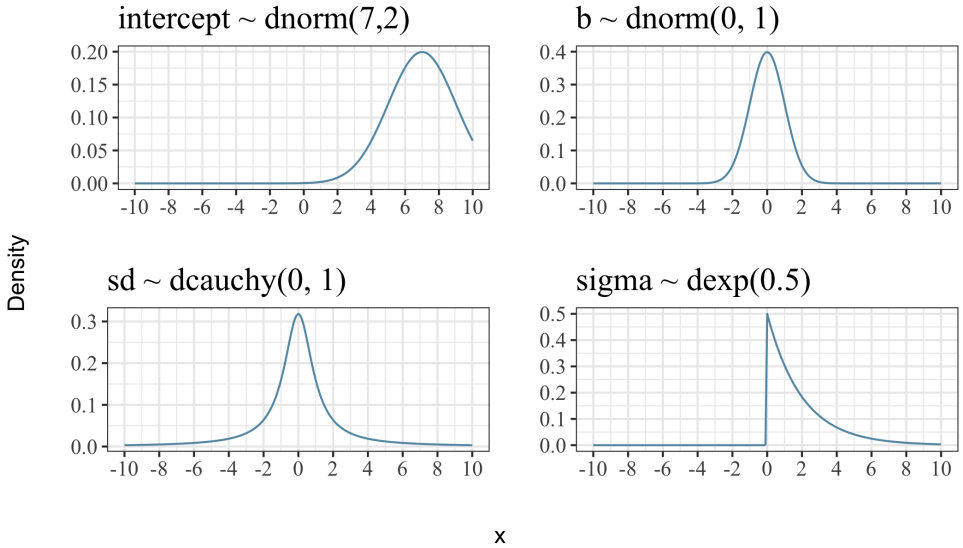

Figure 6: Plots of chosen priors

The assumption of a normally distributed effects is hard to justify, considering the distribution of life satisfaction as it is in wave 1. (Figure 1). The peaks for values of life satisfaction at 5 and at 10 do not really represent a normal distribution. The main reason for the use of normal distributions is mainly mathematical tractability. However, if the family of Bayesian models is inapporpriate, Bayesian answers can be quite misleading (Rubin (1981) 394)

5.3 Model validation

After fitting the model in the previous section, I will now check its fit to data.A basic approach to asses the models adequacy is what is described as "phenomenological Bayesian monitoring" in Rubin (1981) (394) also: Gelman Bayesian Data Analsis ch. 5.5. The method is as follows. The posterior distribution, the distribution of the model parameters conditional on the observed data, can be used to generate a posterior predictive distribution. The implied predictions come from sampling from the posterior distribution to simulate predictions. The idea behind this approach is then to compare the posterior predictive distribution to a) the actual data and b) scientific judgment about plausible values of such data. If the model fits well, there should not be any systematic differences between what the model predicts to what the real data show and they will not contradict with plausible values but will be typical of them. Recalling Chapter 4, the posterior predictive distribution is

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta,$$

where y is the current data, $\tilde{y}$ is to be predicted, and θ are the model parameters. I chose a graphical representation of the PPCs. Instead of calculating posterior probability, simulated data are plotted, to be visually compared to observed data. Figure @re(fig:post-dens-overlay) shows many replicated data sets drawn from the posterior predictive distribution (thin light lines). Thinking predictively, these are data that would be observed in the future if the experiment was replicated with the same model and the same value of θ that produced the observed data y. They are compared to the empirical distribution of the observed outcome (thick dark line). It is evident that the model is able to simulate new data that are similar to the observed life satisfaction with the exception that the exceptional high frequencies of the value 5 and 10 on the satisfaction scale

cannot be reflected correctly. On the right side of Figure 7, the light bars show the distribution of the mean valaue in the replication which captures the computed mean of the observed y well.

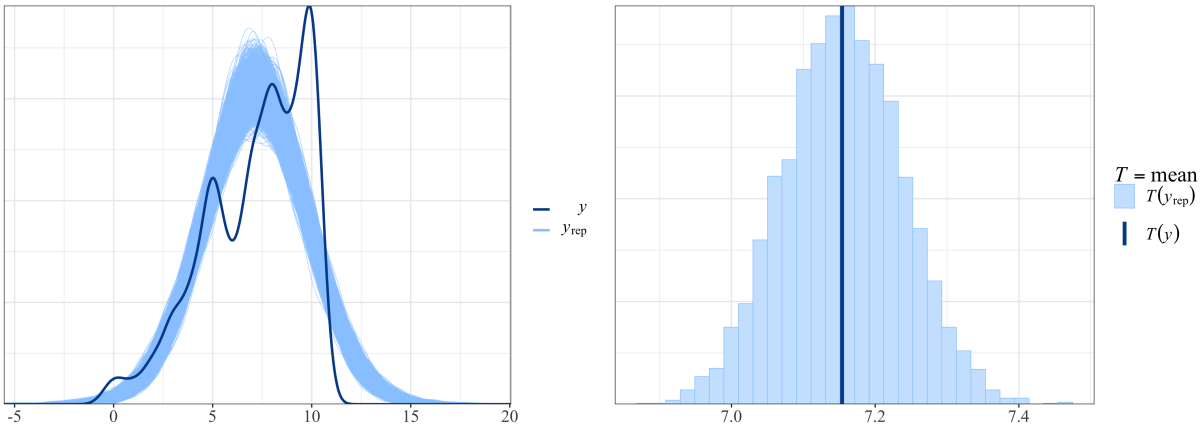

Figure 7: Posterior predictive check

In Figure **??**, the PPCs are repeated for the three regressions of satisfaction with friends, class, and school. With the exception of the density of satisfaction with friends, all of the models predictions coincide weill with the actual data.

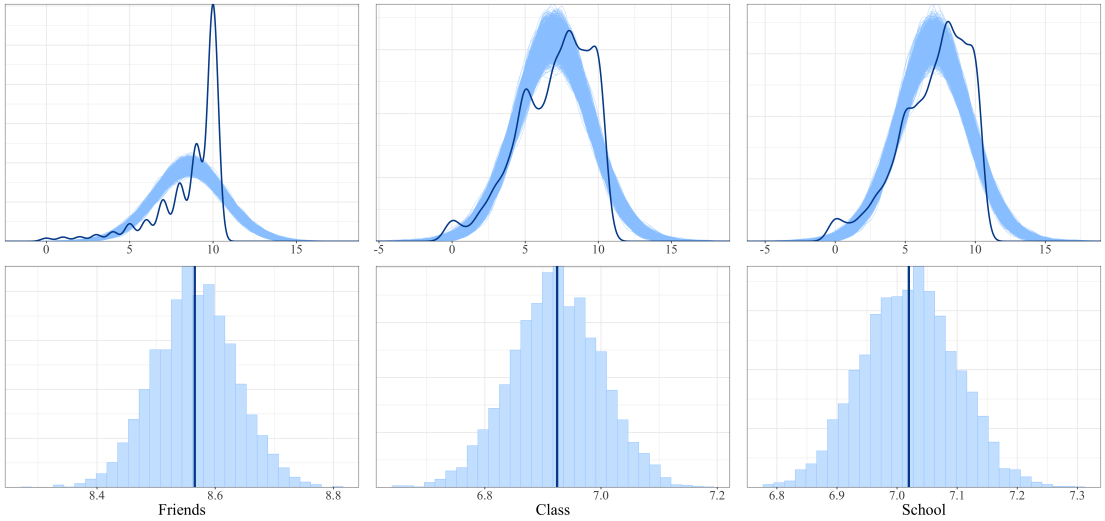

Figure 8: Posterior predictive check

6 Results

In this chapter I will focus on the causal effect of the intervention on the students' life satisfaction. I will also bring in the results for the three observed areas of satisfaction. The general life

satisfaction will be discussed in more detail than the satisfaction areas because this is where my main interest lies.

As for the model checking, good way for summarizing and interpreting the posterior distribution is to sample parameter values drawn from it. These samples can then be used to produce intervals and point estimates. The resulting samples will have the same proportions as the exact posterior density. Therefore the individual values of the parameters will appear in the samples in proportion to the posterior plausibility of each value. elreath p 52

Figure 9 summarizes point estimates for the median of the posterior distribution. When analyzing this figure, it is important to keep in mind that any point estimate in Bayesian analysis discards information. A Bayesian parameter consists of an entire posterior distribution and not just a single value. Besides the estimates, the figure includes specified intervals. Different names exist for the interval of posterior probability but I will refer to elreath 54, who prefers to call them compatibility intervals. These intervals indicate the range of parameter values compatible with the model and data. The compatibility intervals in Figure&nbso;9 state that out of 4000 values that are drawn from the posterior distribution, 90%(50%) lie within the boundaries of the thick(thin) lines. In other words, posterior intervals report two parameter values that contain between them a specified amount of posterior probability, a probability mass. elreath 54

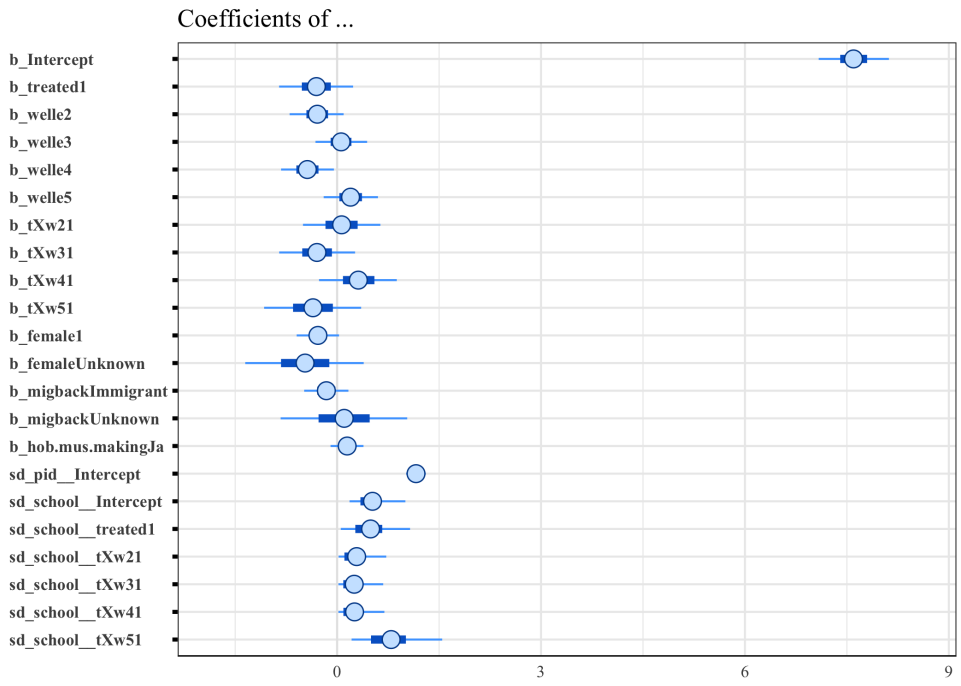

Figure 9: Parameter estimates

6.1 Life satisfaction

I will now come to the identification of the treatment effects. To give a general picture of how life satisfaction changed in the treatment and control group respectively, I averaged the predicted life satisfaction for each wave over all schools (10).

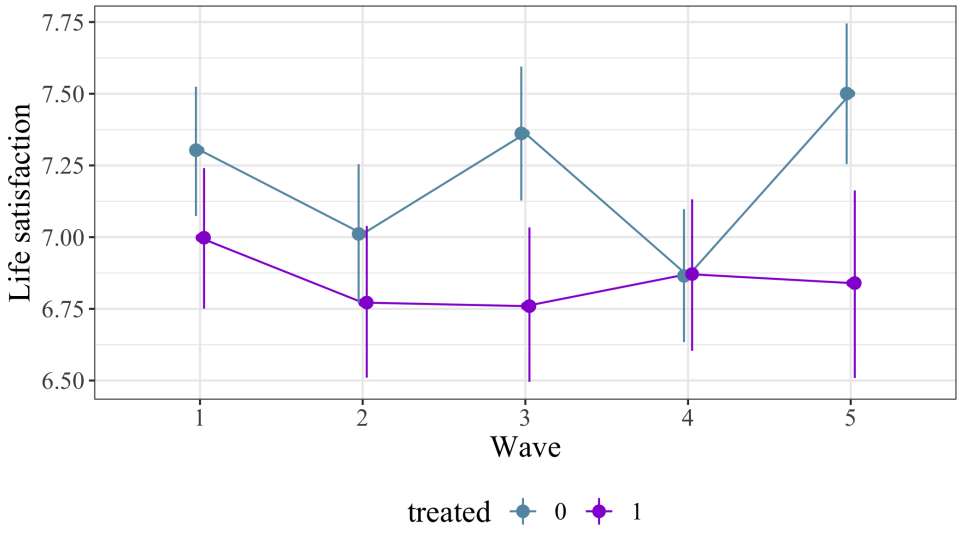

Figure 10: Predicted life satisfaction for average school

Overall, students from the treatment group (blue line) have a lower life satisfaction than the ones from the control group and do not show any large changes in their satisfaction level. After the first period, they become even slightly less satisfied and remain on the same level until the end of the project. The control group shows a different dynamic. Their satisfaction level drops by around 0.25 points on the satisfaction scale after wave 1, almost parallel to the treatment group. But the both groups then depart from each other as the control group goes up to the initial level in wave 3 to drop again in period 4. From period 4 to period 5, life satisfaction jumps back and even exceeds the level of period 1. Evaluating the average treatment effects in Figure 11, the differences between the control and the treatment group, indicate that the treatment effect appear not due to changes in the LS of the treated but are due to varying LS among the untreated. In other words, the differences between the control group and the treatment group are resulting from the the control group being quite volatile in their life satisfaction while the treatment group is rather consistent and generally on a lower level. The supposedly positive treatment effect that Figure @(fig:lsat-teff) indicates in waves 2 and 4 simply appear because the difference between the treatment group and the control group are shrinking, compared to the previous period. Not because the treatment class is actually becoming happier. With all of the 80% uncertainty intervals crossing the zero-line, treatment effects are rather weak.

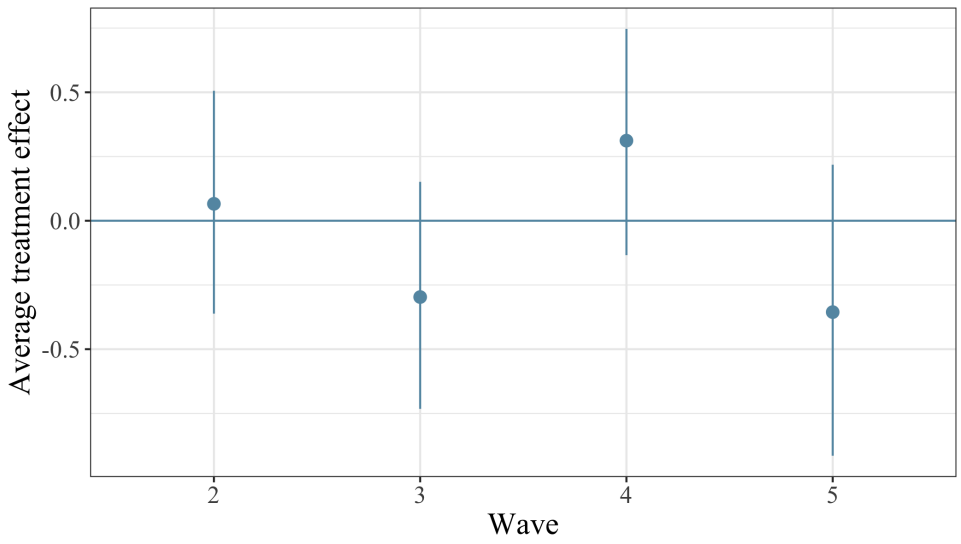

Figure 11: Predicted life satisfaction for average school

To reflect the heterogeneity across schools, Figure 12 represents the school-specific predicted life satisfaction. The corrseponding treatment effects are shown in Figure 13. The pattern from Figure 11 also appears in Figure 13: There is a tendency of the treatment effects alternating from positive to negative from one wave to another. The magnitude of the effects however is extremely small and the 80% certainty intervals always cross the zero line. Wave 5 shows a diffuse situation where school five seems to be completely off. The effect appears because from wave 4 to wave 5, the control group in school 5 becomes more satisfied while at the same time, the average life satisfaction in the treatment class suddenly drops even though it showd a constant level in the previous periods. The problem of school five being not very trustworthy was already mentioned when the descriptive statistics were introduced in Section 3.2.

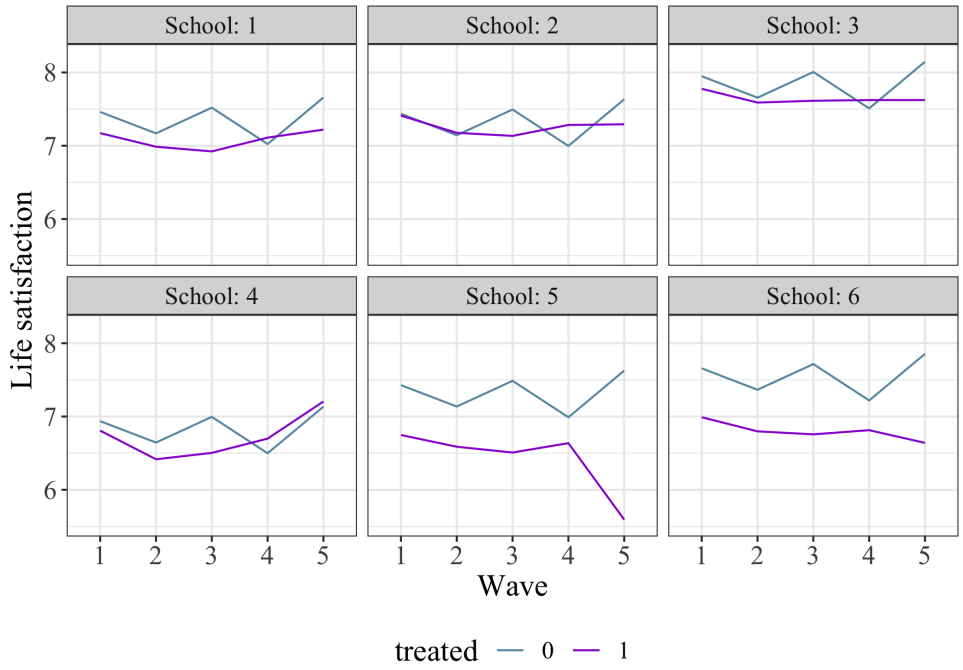

Figure 12: Predicted life satisfaction across schools

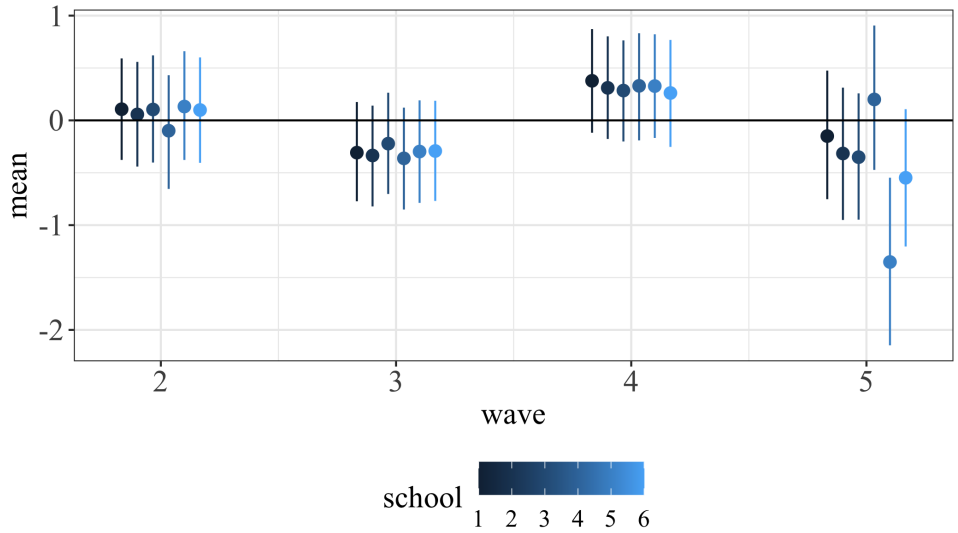

Figure 13: Treatment effects across schools

6.2 Satisfaction with friends

The satisfaction with friends is negatively effected by the treatment. In both of the groups, the students tend to become less satisfied with their friends over the course of the five waves. (Figure 19). The control group becomes a little more satisfied, by around 0.2 points on average, in wave 2 compared to the previous wave but then drops by more than 0.2 points in wave 3. It then stays on the same level of about 8.6. The treatment group drops consistently, being less satisfied with friends from period to period. The treatment group starts with a mean satisfaction at almost 9, which is the highest values observed in all of the areas, and reaches a level of even less than 8.2 in wave 5. The point estimates of the treatment effects are below zero in each school (Figure 14, right plot), as well as the 80% uncertainty intervals (Figure 14, left plot). Again, it is school 5 where the treatment group shows the most rapid drop in satisfaction with friends from 9 in wave 1 to 7.5 in wave 5. The other schools also also a decrease in their satisfaction levels but school 1 and school 3 show a tendency for increasing satisfaction towards the end of the project (Figure 20).

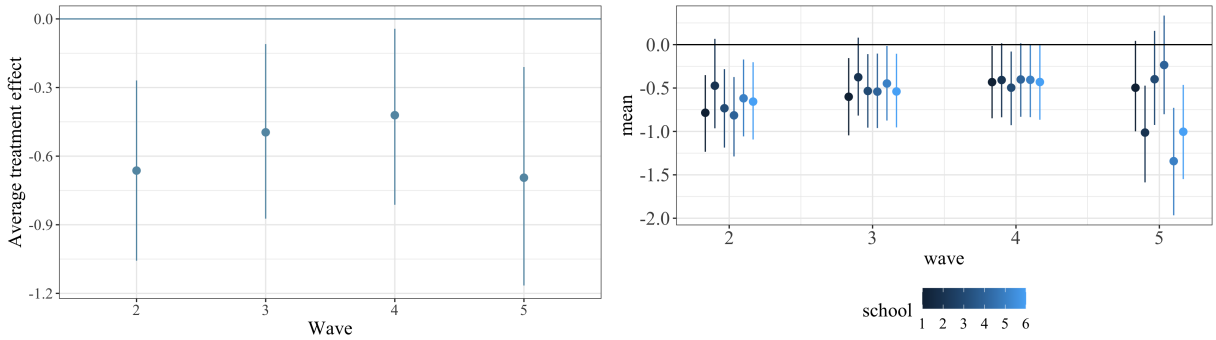

Figure 14: Treatment effects by area of satisfaction for average school

6.3 Satisfaction with the class

Averaging over all schools shows similar developments of the satisfaction with the class in the treatment and the control group (Figure 21). Both of them drop sharply in the first two periods and then remain on a level of around seven with the treatment group generally showing slightly lower satisfaction levels than the control group. The similarity between the groups leads to treatment effects that are close to zero in all of the waves (Figure 15, left plot). When looking at each school separately, it stands out that in school 1 and 2, it is the students in the treatment class that are on average happier than the ones in the control group. The treatment effects across schools are represented in (Figure 15, right plot). The strikingly positive effects in wave 2 in schools 2 and 3 appear because in both cases the control group becomes more satisfied with friends while the treatment group becomes less satisfied (Figure 22).

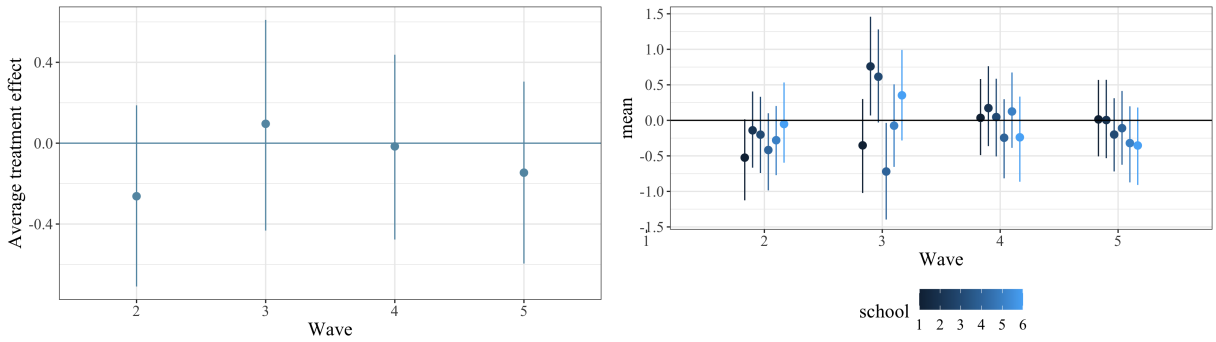

Figure 15: Treatment effects by area of satisfaction for average school

6.4 Satisfaction with school

The level of satisfaction with school is quite volatile throughout the time of the project. There is a pattern observable in the control group which shows that students are more satisfied with school at the beginning and at the end of a school year and less satisfied in between, with an overall general trend to be less satisfied. The students in the treatment group generally show a level of satisfaction of more than eight and show values between 7 and 7.25 in all the other periods (Figure 23). The drop from wave 1 to wave 2 can be seen in each of the in each of the observed classes. But while in school 5 it continues to drop in the remaining waves with the lowest value in wave 5 being only 6. In all the other schools satisfaction with school goes up after that initial drop and eventually levels between 7 and 8. Figure 16 shows the treatment effects for the average school satisfaction over the schools (left plot) and across the schools. The effects lie close to zero but have a tendency to be negative.

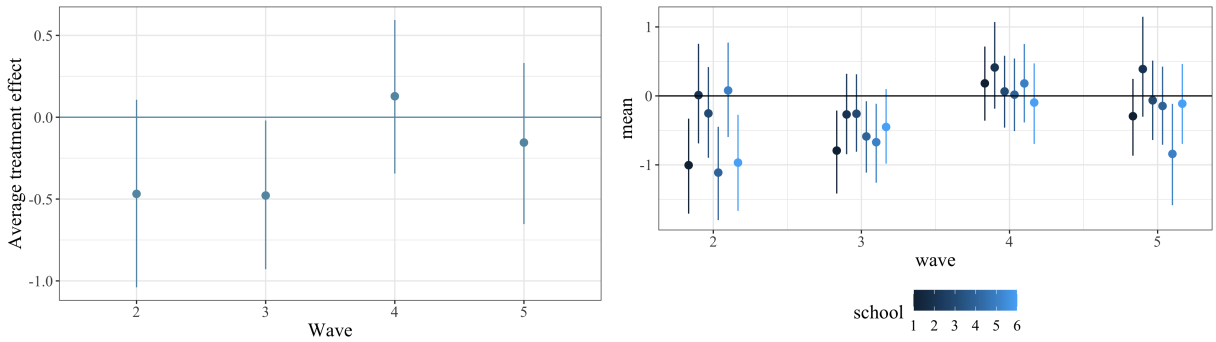

Figure 16: Treatment effects by area of satisfaction for average school

7 Conclusion

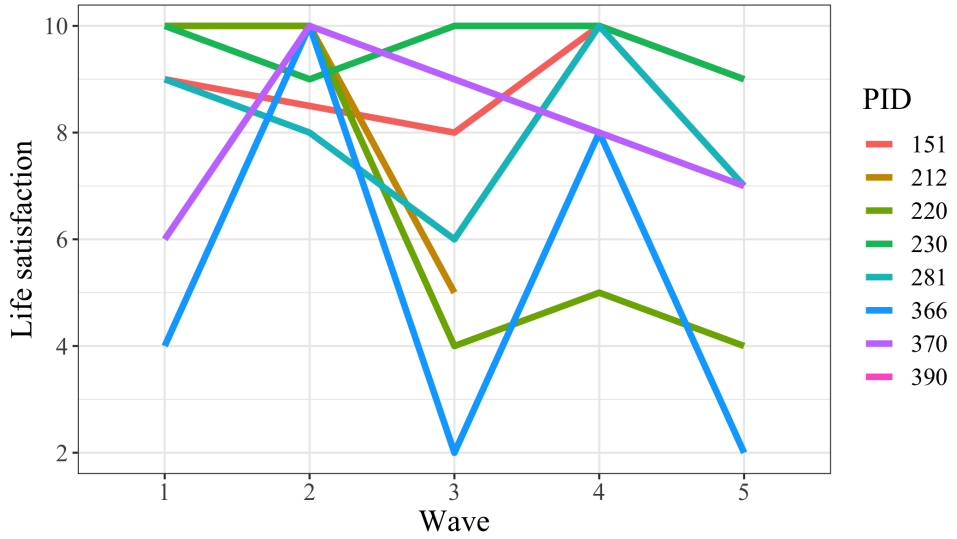

Figure 17: Excerpt from questionnaire

Is the common trend assumption too strong? There are no pre-treatment observations that compare the groups. Can we compare the classes???

References

Adelman, H. S., Taylor, L., & Nelson, P. (1989). Minors dissatisfaction with their life circumstances. *Child Psychiatry & Human Development*, *20*(2), 135–147. https://doi.org/10.1007/bf00711660

Andrew Gelman, J. H. (2013). *Data analysis using regression and multilevel/hierarchical models*. Retrieved from https://www.ebook.de/de/product/6522123/andrew_gelman_jennifer_hill_data_analysis_using_regression_and_multilevel_hierarchical_models.html

Bundesverband Musikunterricht (BMU). (2016). *Grundsdatzpapier*. Retrieved from https://www.bmu-musik.de/ueber-uns/positionen/agenda-2030-bmu-positionen-9-2016.html

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Casas, F., & Rees, G. (2015). Measures of children's subjective well-being: Analysis of the potential for cross-national comparisons. *Child Indicators Research*, *8*(1), 49–69. https://doi.org/10.1007/s12187-014-9293-z

Casas, F., Sarriera, J. C., Alfaro, J., González, M., Malo, S., Bertran, I., … Valdenegro, B. (2011). Testing the personal wellbeing index on 1216 year-old adolescents in 3 different countries with 2 new items. *Social Indicators Research*, *105*(3), 461–482. https://doi.org/10.1007/s11205-011-9781-1

Chambers, C. T., & Johnston, C. (2002). Developmental differences in childrens use of rating scales. *Journal of Pediatric Psychology*, *27*(1), 27–36. https://doi.org/10.1093/jpepsy/27.1.27

Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: Results from three large samples. *Quality of Life Research*, *23*(10), 2809–2818. https://doi.org/10.1007/s11136-014-0726-4

Costa-Giomi, E. (2004). Effects of three years of piano instruction on children's academic achievement, school performance and self-esteem. *Psychology of Music*, *32*(2), 139–152. https://doi.org/10.1177/0305735604041491

Cummins, R. A. (1997). *Manual for the comprehensive quality of life scale – student (grades 7-12): ComQol-s5* (5th ed.). Malbourne: Deakin University, School of Psychology.

Cummins, R. A., & Lau, A. L. D. (2005). *Personal wellbeing index - school children (PWI-SC)* (3rd ed.). Australian Centre on Quality of Life, School of Psychology, Deakin University, Australia.

Diener, E. (2009). *Culture and well-being* (T. Moum, M. A. G. Sprangers, J. Vogel, R. V. C. Michalos, E. Diener, & W. G. and, Eds.). https://doi.org/10.1007/978-90-481-2352-0

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Ferrer-i-Carbonell, A., & Frijters, P. (2004). How important is methodology for the estimates of the determinants of happiness? *The Economic Journal*, *114*(497), 641–659. https://doi.org/10.1111/j.1468-0297.2004.00235.x

Gebert, S. (2018). *Musische erziehung ist keine privatangelegenheit*. Retrieved from https://www.deutschlandfunk.de/musikunterricht-an-schulen-musische-erziehung-ist-keine.680.de.html?dram:article_id=419333

Gilman, R., & Huebner, E. S. (2000). Review of life satisfaction measures for adolescents. *Behaviour Change*, *17*(3), 178–195. https://doi.org/10.1375/bech.17.3.178

Gilman, R., & Huebner, E. S. (2006). Characteristics of adolescents who report very high life satisfaction. *Journal of Youth and Adolescence*, *35*(3), 293–301. https://doi.org/10.1007/s10964-006-9036-7

Gluskie, A. L. (2012). *Subjective wellbeing in children* (PhD thesis). Deakin University.

González-Carrasco, M., Casas, F., Malo, S., Viñas, F., & Dinisman, T. (2016). Changes with age in subjective well-being through the adolescent years: Differences by gender. *Journal of Happiness Studies*, *18*(1), 63–88. https://doi.org/10.1007/s10902-016-9717-1

Guhn, M., Emerson, S. D., & Gouzouasis, P. (2019). A population-level analyses of associations between school music participation and academic achievement. *Journal of Educational Psychology*, *112*(2), 308–328. https://doi.org/http://dx.doi.org/10.1037/edu0000376

Gullone, E., & Cummins, R. (1999). The comprehensive quality of life scale: A psychometric evaluation with an adolescent sample. *Behaviour Change*, *16*, 127–139.

Hille, J., Adrian; Schupp. (2014). How learning a musical instrument affects the development of skills. *Economics of Education Review*, *44*, 56–82. https://doi.org/http://dx.doi.org/10.1016/j.econedurev.2014.10.007

Huebner, E. S. (1991a). Correlates of life satisfaction in children. *School Psychology Quarterly*, *6*(2), 103–111. https://doi.org/https://doi.org/10.1037/h0088805

Huebner, E. S. (1991b). Initial development of the students life satisfaction scale. *School Psychology International*, *12*(3), 231–240. https://doi.org/10.1177/0143034391123010

Huebner, E. S. (1994). Preliminary development and validation of a multidimensional life satisfaction scale for children. *Psychological Assessment*, *6*(2), 149–158. https://doi.org/10.1037/1040-3590.6.2.149

Kim, K. J., Conger, R. D., Elder, G. H., & Lorenz, F. O. (2003). Reciprocal influences between stressful life events and adolescent internalizing and externalizing problems. *Child Development*, *74*(1), 127–143. https://doi.org/10.1111/1467-8624.00525

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*(1), 37–49. https://doi.org/10.1037/0012-1649.22.1.37

Mcelreath, R. (2020). *Statistical rethinking*. Retrieved from https://www.ebook.de/de/product/38708673/richard_mcelreath_statistical_rethinking.html

Mehr, S. A., Schachner, A., Katz, R. C., & Spelke, E. S. (2013). Two randomized trials provide no consistent evidence for nonmusical cognitive benefits of brief preschool music enrichment. *PLoS ONE*, *8*(12), e82007. https://doi.org/10.1371/journal.pone.0082007

Mellor, D., Stokes, M., Firth, L., Hayashi, Y., & Cummins, R. (2008). Need for belonging, relationship satisfaction, loneliness, and life satisfaction. *Personality and Individual Differences*, *45*(3), 213–218. https://doi.org/10.1016/j.paid.2008.03.020

Möller, T. (2017). *Ausverkauf musikalischer Bildung?* Retrieved from https://www.deutschlandfunk.de/musikunterricht-in-der-schule-ausverkauf-musikalischer.1992.de.html?dram:article_id=382783

Osborne, M. S., McPherson, G. E., Faulkner, R., Davidson, J. W., & Barrett, M. S. (2015). Exploring the academic and psychosocial impact of el sistema-inspired music programs within two low socio-economic schools. *Music Education Research*, *18*(2), 156–175. https://doi.org/10.1080/14613808.2015.1056130

Piaget, J. (1955). *The construction of reality in the child*. London: Routledge; Keagan Paul.

Piaget, J. (1969). *The child's concept of time*. London: Routledge; Keagan Paul.

Portowitz, A., Lichtenstein, O., Egorova, L., & Brand, E. (2009). Underlying mechanisms linking music education and cognitive modifiability. *Research Studies in Music Education*, *31*(2), 107–128. https://doi.org/10.1177/1321103x09344378

Proctor, C., Linley, P. A., & Maltby, J. (2009). Very happy youths: Benefits of very high life satisfaction among adolescents. *Social Indicators Research*, *98*(3), 519–532. https://doi.org/10.1007/s11205-009-9562-2

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, *6*(4), 377–401. https://doi.org/10.3102/10769986006004377

Sala, G., & Gobet, F. (2016). When the musics over. Does music skill transfer to childrens and young adolescents cognitive and academic skills? A meta-analysis. *Educational Research Review*, *20*, 55–67. https://doi.org/10.1016/j.edurev.2016.11.005

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science*, *15*(8), 511–514. https://doi.org/https://doi.org/10.1111/j.0956-7976.2004.00711.x

Schellenberg, E. G. (2011). Music lessons, emotional intelligence, and IQ. *Music Perception: An Interdisciplinary Journal*, *29*(2), 185–194. https://doi.org/10.1525/mp.2011.29.2.185

Schellenberg, E. G., & Weiss, M. W. (2013). Music and cognitive abilities. In *The psychology of music* (pp. 499–550). https://doi.org/10.1016/b978-0-12-381460-9.00012-2

Southgate, D. E., & Roscigno, V. J. (2009). The impact of music on childhood and adolescent achievement. *Social Science Quarterly*, *90*(1), 4–21. https://doi.org/10.1111/j.1540-6237.2009.00598.x

Stoverock, K. (n.d.). *Ein jahrzehntelanges versagen der bildungspolitik*. Retrieved from https://themen.miz.org/fokus-musikunterricht/interview-hoeppner

Suldo, S. M., & Huebner, E. S. (2004). Does life satisfaction moderate the effects of stressful life events on psychopathological behavior during adolescence? *School Psychology Quarterly*, *19*(2), 93–105. https://doi.org/10.1521/scpq.19.2.93.33313

Tomyn, A. J., & Cummins, R. A. (2011). The subjective wellbeing of high-school students: Validating the personal wellbeing indexSchool children. *Social Indicators Research*, *101*(3), 405–418. https://doi.org/10.1007/s11205-010-9668-6

Tomyn, A. J., Fuller-Tyszkiewicz, M. D., Cummins, R. A., & Norrish, J. M. (2016). The validity of subjective wellbeing measurement for children: Evidence using the personal wellbeing indexSchool children. *Journal of Happiness Studies*, *18*(6), 1859–1875. https://doi.org/10.1007/s10902-016-9804-3

Tomyn, A. J., Stokes, M. A., Cummins, R. A., & Dias, P. C. (2019). A rasch analysis of the personal well-being index in school children. *Evaluation & the Health Professions*, *43*(2), 110–119. https://doi.org/10.1177/0163278718819219

Uy, M. (2012). Venezuela's national music education program el sistema: Its interactions with society and its participants' engagement in praxis. *Music & Arts in Action*, *4*(1), 5–21. Retrieved from http://musicandartsinaction.net/index.php/maia/article/view/elsistema

Valois, R. F., Zullig, K. J., Huebner, E. S., & Drane, J. W. (2004). Life satisfaction and suicide among high school adolescents. *Social Indicators Research*, *66*(1/2), 81–105. https://doi.org/10.1023/b:soci.0000007499.19430.2f

Wetter, O. E., Koerner, F., & Schwaninger, A. (2008). Does musical training improve school performance? *Instructional Science*, *37*(4), 365–374. https://doi.org/10.1007/s11251-008-9052-y

Yang, P. (2015). The impact of msic on educational attainment. *Journal of Cultural Economics*, *39*(4), 369–396. https://doi.org/10.1007/s10824-015-9240-y

Zullig, K. J., Valois, R. F., Huebner, E. S., Oeltmann, J. E., & Drane, J. W. (2001). Relationship between perceived life satisfaction and adolescents' substance abuse. *Journal of Adolescent Health*, *29*(4), 279–288. https://doi.org/10.1016/s1054-139x(01)00269-5

7	**Wie zufrieden bist du …**											
	Bitte kreuze für jeden Bereich auf der Skala einen Wert an: *Wenn du ganz und gar unzufrieden bist, den Wert „0“,* *wenn du ganz und gar zufrieden bist, den Wert „10“.* *Wenn du teils zufrieden/teils unzufrieden bist, einen Wert dazwischen.*											
		☹ ganz und gar unzufrieden 0	1	2	3	4	5	6	7	8	9	☺ ganz und gar zufrieden 10
a)	... zur Zeit, alles in allem, mit deinem Leben?	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
b)	... mit dem, was du hast? Denke dabei an Geld und Dinge, die du besitzt.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
c)	... mit deiner Gesundheit?	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
d)	... mit deiner Familie?	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
e)	... mit deinem Bekannten- und Freundeskreis?	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Figure 18: Excerpt from questionnaire

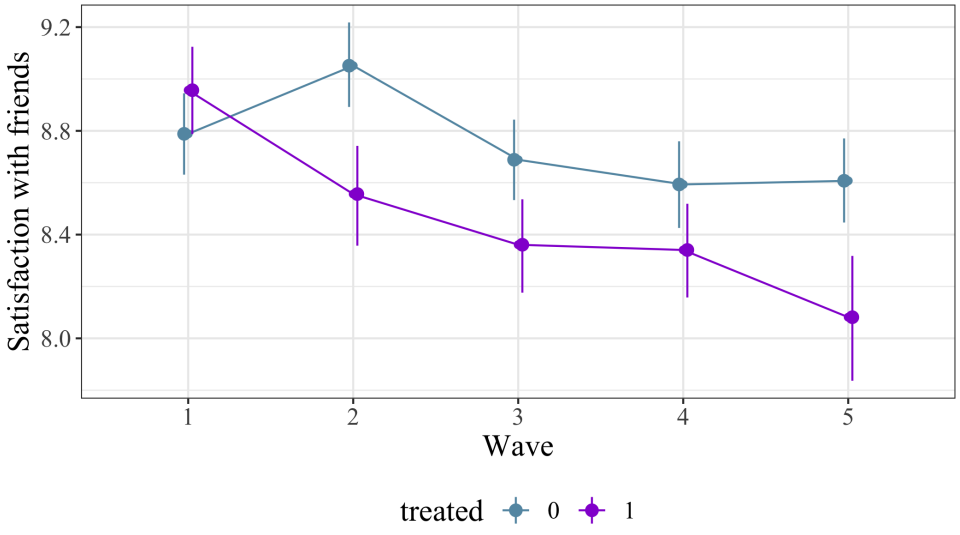

Figure 19: Predicted satisaction by area for average school

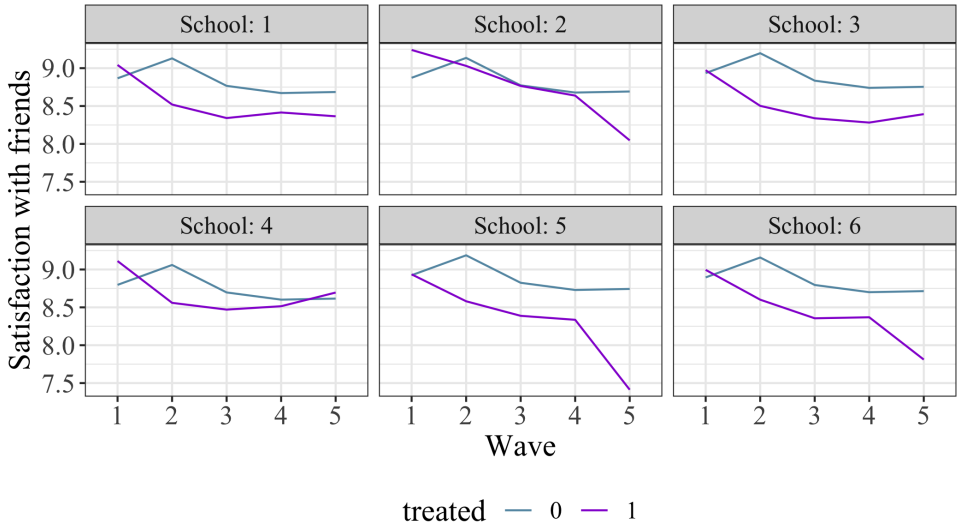

Figure 20: Predicted satisaction by area for average school

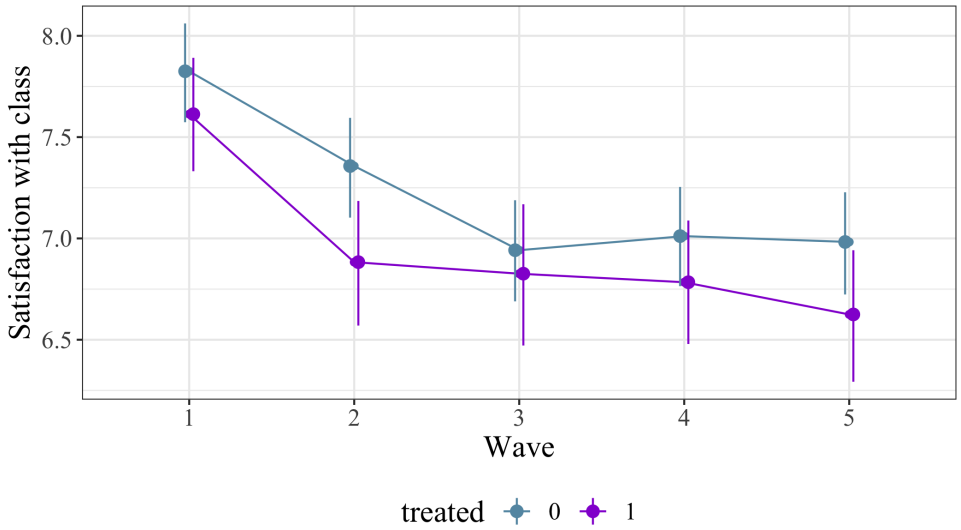

Figure 21: Predicted satisaction by area for average school

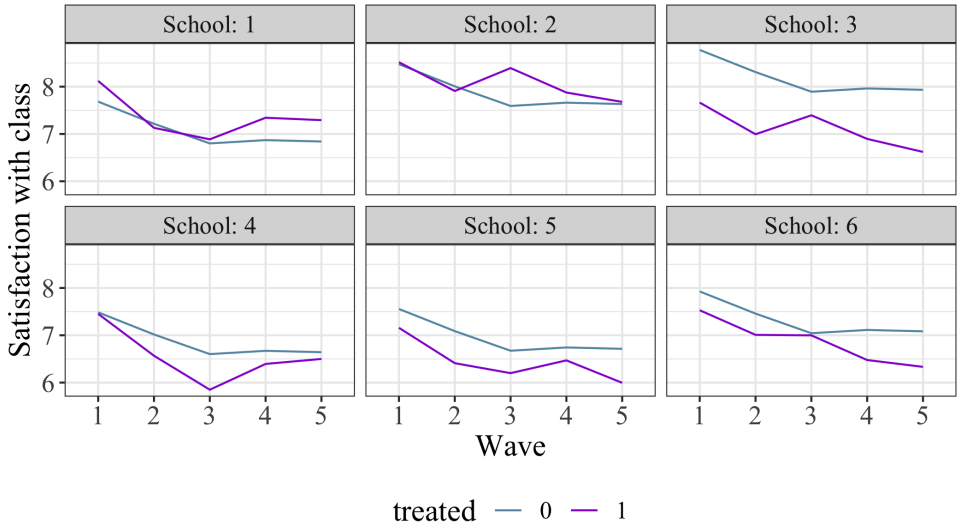

Figure 22: Predicted satisaction by area for average school

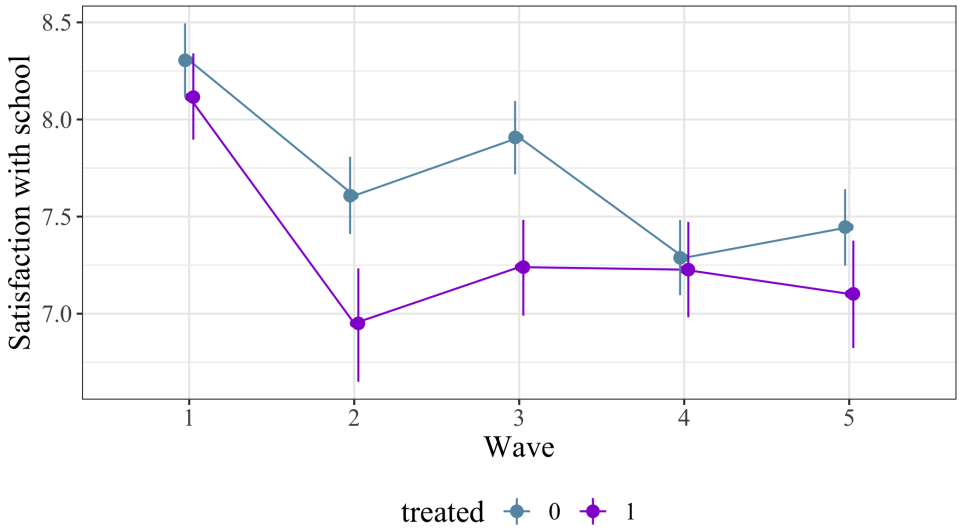

Figure 23: Predicted satisaction by area for average school

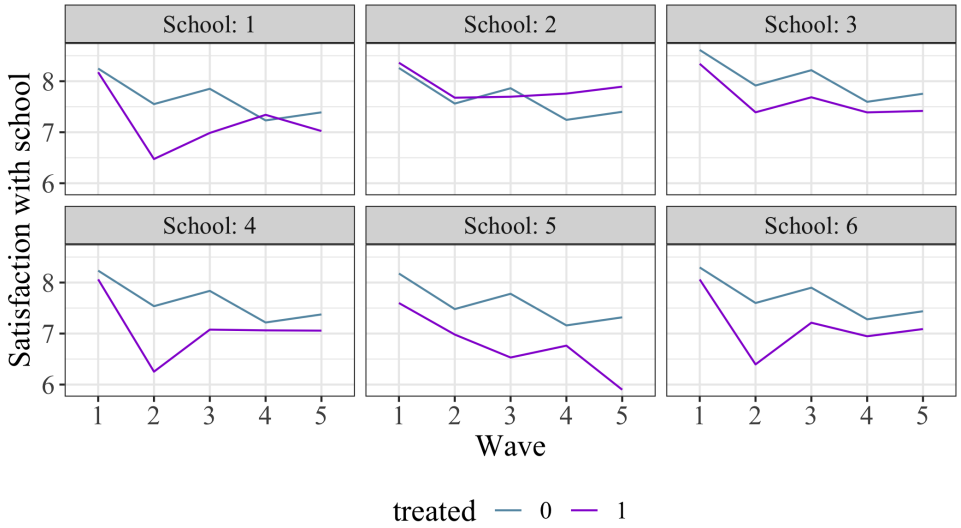

Figure 24: Predicted satisaction by area for average school

Example of nice appendix in Hille (2014)

A Declaration of authorship