Ambiguity in European seasonal comparative research: how decisions on modelling shape results on inequality in learning?

Laura A. Helbling · Martin J. Tomasik · Urs Moser

Received: 20 May 2020 / Revised: 11 December 2020 / Accepted: 9 March 2021 / Published online: 31 March 2021
© The Author(s) 2021

Abstract Summer break study designs are used in educational research to disentangle school from non-school contributions to social performance gaps. The summer breaks provide a natural experimental setting that allows for the measurement of learning progress when school is not in session, which can help to capture the unfolding of social disparities in learning that are the result of non-school influences. Seasonal comparative research has a longer tradition in the U.S. than in Europe, where it is only at its beginning. As such, summer setback studies in Europe lack a common methodological framework, impairing the possibility to draw lines across studies because they differ in their inherent focus on social inequality in learning progress. This paper calls for greater consideration of the parameterization of “unconditional” or “conditional” learning progress in European seasonal comparative research. Different approaches to the modelling of learning progress answer different research questions. Based on real data and constructed examples, this paper outlines in an intuitive fashion the different dynamics in inequality that may be simultaneously present in the survey data and distinctly revealed depending on whether one or the other modeling strategy of learning progress is chosen. An awareness of the

Dr. L. A. Helbling (✉) · Prof. Dr. M. J. Tomasik · Prof. Dr. U. Moser
Institute for Educational Evaluation, University of Zurich, Wilfriedstrasse 15, 8032 Zürich,
Switzerland
E-Mail: laura.helbling@ibe.uzh.ch

Prof. Dr. M. J. Tomasik
E-Mail: martin.tomasik@ibe.uzh.ch

Prof. Dr. U. Moser
E-Mail: urs.moser@ibe.uzh.ch

Prof. Dr. M. J. Tomasik
Department für Psychologie und Psychotherapie, Universität Witten/Herdecke,
Alfred-Herrhausen-Straße 44, 58448 Witten, Germany

parameterization of learning progress is crucial for an accurate interpretation of the findings and their international comparison.

Keywords Seasonal comparative research · Summer gap · Social disparities

Mehrdeutigkeit Sozialer Disparitäten in Europäischen Sommerloch-Studien: Wie methodische Entscheidungen Resultate zu Ungleichheiten in Lernverläufen beeinflussen

Zusammenfassung Sommerlochstudien werden in der erziehungswissenschaftlichen Forschung zur Separation der Beiträge ausserschulischer und schulischer Faktoren im Rahmen sozialschichtabhängiger Lernverläufe verwendet. Die Sommerferien bieten ein Experiment ähnliches Setting, welches es erlaubt, sozialschichtabhängige Lernzuwächse, respektive, Lernverluste über einen längeren schulfreien Zeitraum zu messen. Sozialschichtabhängige Divergenzen in den Lernleistungen über diesen Zeitraum, werden auf ausserschulische Faktoren zurückgeführt. Sommerlochstudien haben eine längere Tradition in den USA im Vergleich zu Europa, wo Untersuchungen zu Ferieneffekten noch weniger verbreitet sind. Die Europäische Forschung zu Ferieneffekten kennt dementsprechend auch keinen gemeinsamen methodologischen Rahmen, was die Vergleichbarkeit der Ergebnisse zu vorgefundenen sozialen Disparitäten erschwert, wenn nicht gar verunmöglicht. Dieser Artikel sensibilisiert für Unterschiede in der Ergebnisinterpretation nach methodischem Zugang. Je nachdem ob sozialschichtabhängiger Lernzuwachs „konditional“ oder „unkonditional“ parametrisiert wird, wird eine unterschiedliche Forschungsfrage beantwortet. Basierend auf einer kürzlich durchgeführten Sommerlochstudie sowie zwei konstruierten Beispielen, versucht dieser Artikel auf intuitive Art und Weise zu zeigen, inwiefern unterschiedliche methodische Zugänge unterschiedliche Dynamiken sozialer Disparitäten fokussieren, welche simultan in den Daten präsent sein können. Kenntnisse der methodischen Divergenzen sind unumgänglich für eine akkurate Interpretation der Ergebnisse und deren internationalem Vergleich.

Schlüsselwörter Sommerloch · Ferieneffekte · Soziale Disparitäten

1 Background

The part played by family and schooling in perpetuating educational stratification is at the heart of sociologically founded educational research. Through standardized educational testing, the schools' accountability for their students' performance development has come increasingly into public focus (William 2010), providing also the empirical basis to trace social inequality in performance development. Social performance gaps have been shown to widen as students progress through school (e.g., Cameron et al. 2015; Caro et al. 2009; Helbling et al. 2019). Yet, separating school from non-school influences on student learning is a complex issue. Learning progress that occurs throughout schooling reflects the effects of attending school as well as the many other developmental influences of the contexts in which students

interact (Bloom et al. 2008, p. 296). Students of different social backgrounds start school with differing levels of initial knowledge and spend their leisure time in different contexts, formed by their familial environments, neighborhoods, and peer-groups, which are all influential to their learning progress and the development of assets conducive to learning in school (e.g., Ainsworth 2002; Boudon 1974; Bourdieu 1971; Bradley and Corwyn 2002; Lee and Burkam 2002; Jencks and Mayer 1990; Sirin 2005). Though school rewards and recognitions may create institutionalized mechanisms that operate in socially selective ways (e.g., Bourdieu 1985; Bourdieu and Passeron 1977; Caro et al. 2009) and teachers may be disproportionately effective based on their student composition (e.g., van Ewijk and Slegers 2010), the empirically observed social divide in students' performance development (e.g., Cameron et al. 2015; Caro et al. 2009; Helbling et al. 2019), may also (or even primarily) be explained by extracurricular influences. In this vein, and in contrast to views of schools as drivers of social inequality, schools have also been considered "great equalizers," preventing the even greater widening of social performance gaps that perhaps would occur in the absence of schooling (Downey et al. 2004). Empirical results pinpointing schools as either equalizers or drivers of inequality derive from distinct strands of research. The strand of *seasonal comparative research* compares social stratification in learning that occurs during the time when school is not in session (summer breaks) to stratification in learning that occurs when school is in session. From this strand, scholars have portrayed schools as balancing inequalities. From the strand of research on *school compositional effects*, which investigates the effects of stratification across schools and teaching practices, schools have been portrayed as a source of inequality (Dumont and Ready 2019). To a certain extent, the conflicting views of these scholarly strands on the role of schools in educational inequality can be explained by the different counterfactual frameworks in which these research strands operate. Seasonal comparative research asks, "what would inequality look like if children were exposed to more or less schooling?". Research on effects of school composition asks, "what would inequality look like if children were exposed to identical schools?" (Dumont and Ready 2019, p. 3; Jennings et al. 2015). Hence, though touching on the same subject, the point of reference of scholars in these research strands inevitably differs. This paper sets out to address social disparities in the framework of seasonal comparative research, particularly shedding light on ambiguities in findings in the European context.

Drawing on a *resource faucet perspective* (Entwisle et al. 2000), the argument supporting the seasonal comparative scholars' view on schools as equalizers is that access to culturally and cognitively demanding resources is less socially stratified when school is in session compared to when it is not. In other words, when school is in session, the resource faucet is turned on for all children, but it is unequally turned on when school is not in session. Particularly during the long summer break, homes and neighborhoods provide unequal learning opportunities and children engage in socioeconomically stratified summer activities, which are assumed to underlie their differential learning progress or learning loss over the summer break (e.g., Burkam et al. 2004; Fairchild and Noam 2007; Gershenson 2013; Meyer et al. 2017). Apart from summer activities, more implicitly, the ways parents' engage with their children, parental role-modeling, and the structuring of daily life during the prolonged

school break is assumed to lead to increased socioeconomic performance gaps at the start of the next term (Burkam et al. 2004), resulting in ever larger social gaps throughout the school career. In a longstanding methodological tradition beginning in the early 20th century (e.g., Patterson and Rensselaer 1925), summer breaks have been used as a perfect setup in seasonal comparative research disentangling non-school from school influences on students' learning progress (Cooper et al. 1996). As summer breaks provide institutionalized phases in which students are suspended from all curricular exposure for several weeks, educational assessments before and after the summer break allow for insights into performance decays and gains over a period in which the effects of schooling are eliminated by design.

Research on summer break effects on learning progress has a stronger tradition in the U.S. compared to the more recent research in Europe, where research on summer learning (losses) is still rather scarce. As Paechter et al. (2015) summarize, in nearly all U.S. studies, student's learning (losses) over the summer break depended on parents' socioeconomic attributes. Among these, for example, earlier works by Entwisle and Alexander (1992) and Alexander et al. (2001) showed that low-SES children in elementary education in Baltimore schools (Beginning School study, BSS) mainly lost ground in mathematics and reading performance when school was not in session, and they could trace large parts of the achievement gaps in early high school back to unequal summer learning (Alexander et al. 2007). In their widely influential work based on the nationally representative Early Childhood Longitudinal Study (ECLS, Kindergarten Cohort of 1998–1999), Downey et al. (2004) showed that disparities in reading and math performance development grew faster during the summer compared to the rest of the school year. More recent U.S. studies that re-analyze ECLS and BSS data as well as analyzing new ECLS cohort data (e.g., Dumont and Ready 2019; Hippel and Hamrock 2019; Quinn et al. 2016), report more nuanced results where schools do play a compensatory role but SES-related inequality in learning rates is smaller than reported by previous studies and not consistent across grades in elementary education.

European studies are mixed¹ regarding the existence of the effects of socioeconomic variables on summer learning. For the UK, Shinwell and Defeyter (2017) investigated summer learning losses in reading and spelling among low-SES children in grades 1–5 from three schools in the North-East of England and West of Scotland. They found summer learning losses occurred in spelling. Based on the two-year panel study SCHLAU, which focused on seasonal trends in learning of students in German schools between grades 5 and 7, Siewert (2013) reports significant socioeconomic stratification in mathematical learning over the summer break. Low-SES students experienced a performance decay whereas high-SES students did not decrease in performance. Though, depending on how SES was operationalized, some widening socioeconomic gaps also occurred during the school year. Focusing on German schools in Berlin, Becker et al. (2008) analyzed pre- and post-summer break assessments of the study ELEMENT for students transiting from grade 4 to grade 5. Their findings also suggest widening gaps in reading skills between chil-

¹ Here we only summarize the state of research published in either German or English. There probably exist more summer break studies published in other national languages in the European context.

dren of different migrant status and socioeconomic backgrounds over the summer break. In contrast, in the project framework of the Jacobs Summer Camps, which focused on the promotion of German language skills for students with migrant background during the intervening summer break between grades 3 and 4 (Stanat et al. 2005), results did not reveal any socioeconomic variation in summer setback (Coelen and Siewert 2008). In a more recent study in Germany, Meyer et al. (2017) analyzed summer learning effects in reading comprehension and writing for students in grade 2 across two schools of differential income composition. Again, they found social disparities in summer learning in reading comprehension to occur between high- and low-income peers: the former gained, the latter stagnated. Paechter and colleagues (2015) investigated summer learning losses in mathematics, reading, and spelling for students from Austria in lower secondary education. They only found summer learning losses in mathematics to vary across the educational background of the mother; generally, they conclude that the impact of socio-economic resources is small. Following a sample of kindergarteners throughout first grade in the Flemish part of Belgium in the SiBO-project, Verachtert et al. (2009) did not find statistically significant socioeconomic differences in mathematical performance development during the intervening summer break. Similarly, in a Swedish study on fifth graders in Stockholm schools, Lindahl (2001) did not find any conclusive evidence for socioeconomic discrepancies in mathematical learning over the summer break to grade 6.

Though all of these studies focus on summer break effects in an effort to pin down the non-school contribution in growing social disparities in academic performance development, the findings of these studies may not be straightforwardly compared. There are different aspects in which summer setback studies differ, which could underlie the *ambiguities in findings*. The lengths of the summer vacations under consideration differ (Meyer et al. 2017). Further, there may be subject-specific differences. While home environments generally provide more opportunities to practice language, learning in mathematics is more strongly tied to schooling (Burkam et al. 2004; Cooper et al. 1996). Thus, stronger social stratification in reading compared to mathematical learning over the summer break has been expected (Becker et al. 2008). Depending on when the pre- and post-assessments are done, different amounts of instructional time has been included in the timeframe under assessment (Cooper et al. 1996; Paechter et al. 2015). Also, different ways of operationalizing socioeconomic attributes of students result in incomparability of findings (e.g., Sirin 2005). Further, researchers have used absolute and relative metrics to report performance change over the summer, which inherently reflect different perspectives on change; in addition, the handling of measurement error can alter results (Quinn 2015; Quinn et al. 2016). Hippel and Hamrock (2019) found results to be sensitive towards the choice of test score scaling methods and to changes in test forms, i.e. whether the same test form is used for pre- and post-testing. Moreover, one very crucial point of divergence in the investigation of social disparities in summer setback effects, is the *choice of the modeling strategy* (see Quinn 2015). Researchers analyzing pre-test and post-test designs are confronted with the choice between a difference score approach or a regressor variable approach (Allison 1990; van Breukelen 2013; Gastro-Schilo and Grimm 2018). Though some studies remark that the findings may

potentially diverge based on the statistical method chosen, which is also discussed as an instance of Lord's paradox in the quasi-experimental literature (Allison 1990; Lord 1967), this issue has gained little attention in applied summer setback research, particularly in the European context, where seasonal comparative research is still in its infancy (see Meyer et al. 2017; Paechter et al. 2015; Shinwell and Defeyter 2017). Among European summer setback studies, there seems to be no agreement on methodology. Depending on the method chosen, the magnitude and the direction of effects can change. In line with earlier works (e.g., Holland and Rubin 1982), Quinn (2015) argues that rather than pointing to a supposedly paradoxical situation of contradictory results, the statistical models applied answer different research questions. Very recently, Dumont and Ready (2019) outlined how differing modeling strategies of student learning influence our understanding of the role schools play in educational inequality. They showed that research strands favoring or opposing the view that schools exacerbate educational stratification differ in their methodological habits concerning the way learning progress is modelled (Dumont and Ready 2019). Complementing Dumont and Ready's (2019) thorough discussion on the differential model foci of strands of scholars on schools' part in educational stratification, which mainly pertains to the U.S. seasonal comparative context, the main aim and contribution of this paper is to draw attention to the methodological inconsistencies in European summer setback research and, in particular, to technically illustrate the distinct dynamics in inequality that may be revealed by the model of choice. This paper intends to make the technicalities underlying the relevance of the parameterization of learning progress for the interpretation of results and for international comparative purposes accessible to applied social science research in general. To do so, we, first, outline the models of choice and their distinct ways of parameterizing learning progress in a simple pre-test post-test setting in formula notation. Then, we introduce the problem of conflicting results by model choice based on a very recent summer setback study conducted in Switzerland. Following this and going more deeply into the technicalities of the generating data structure, we illustrate the problem more in depth with some easy-to-follow simulated examples and graphs, where we explore and illustrate the data structure, which produces the ambiguity in findings depending on model choice.

2 Methodological issues: the choice of model

When we analyze data from pre-test post-test designs in order to recover differences² in the way different groups change between measurement occasions, the literature points towards two commonly and interchangeably used methodological strategies: the *difference score approach* and the *regressor variable approach*, also known as the residualized change or lagged variable model (Allison 1990; van Breukelen 2013; Gastro-Schilo and Grimm 2018; Quinn 2015).

² The problem outlined is the same when analysing the effect of a continuous variable on change (Gastro-Schilo and Grimm 2018).

2.1 The difference score model

The difference score model [DM] regresses a difference score $Y_{2i} - Y_{1i}$ on a group indicator G_i (e.g., SES groups). The difference score is the difference between student i 's post- and pre-test scores $Y_{2i} - Y_{1i}$. The difference score model can be depicted as follows:

$$Y_{2i} - Y_{1i} = \text{Intercept}_{DMi} + b_{DM}G_i + e_{DMi}$$

where b_{DM} recovers the effect of the group indicator at issue on change.

2.2 The regressor variable model (residualized change or lagged variable model)

The regressor variable model [RM] regresses student i 's post-test score Y_{2i} on a group indicator G_i (e.g., SES groups), while controlling for student i 's pre-test score Y_{1i} . The regressor variable model may be depicted as follows:

$$Y_{2i} = \text{Intercept}_{RMi} + b_1Y_{1i} + b_{RM}G_i + e_{RMi}$$

where b_{RM} recovers the effect of the group indicator at issue on (residualized) change. Please note that if we set $b_1 = 1$, then the regressor variable model is equivalent to the difference score model. In reality, this coefficient is usually somewhere between 0 and 1 (see Allison 1990).

Depending on the author(s), model preferences differ. The difference score has gotten a bad reputation for its' susceptibility to increased unreliability when the component measures are affected by measurement error (see Allison 1990; Gastro-Schilo and Grimm 2018). Some have also suggested that measurement error in the pre-test can attenuate discrepancies in results across models (see van Breukelen 2013). Throughout the literature, different suggestions have been made. Becker and colleagues (2008) favor the regressor variable approach for their pre-test post-test data because variances between measurement occasions do not change, so effects estimated by the difference score are not more reliable. Gastro-Schilo and Grimm (2018) note that if the groups differ in baseline levels in the pre-test, the regressor variable model is always biased and should be discarded. They suggest conducting difference score analyses in a latent variable framework, where measurement error can be controlled for (see Gastro-Schilo and Grimm 2018). Allison (1990) shows that the findings based on the regressor variable approach are biased when the difference score model is the true data generating process, where the regressor variable model, for example, estimates a treatment effect though there is in fact none. The attempts to establish guidelines as to "when to use which model" have rested mainly on (untestable) model assumptions and possible reliability adjustments (Allison 1990; van Breukelen 2013; Gastro-Schilo and Grimm 2018). Van Breukelen (2013) suggests that for a non-randomized study design, an agreement between the two methods might offer reassurance about the existence of an effect.

It is a long-standing tradition in the (quasi-)experimental field of research for these models to be applied interchangeably with the intention to recover the same causal effect of some treatment, comparing groups of treated and controls. The finding that the models can yield different conclusions (in direction, magnitude, and significance of b_{DM} vs. b_{RM}) has come to be known as “Lord’s paradox” (Allison 1990). In an example, Lord (1967) showed that the regressor variable approach favored a treatment effect, whereas the difference score model did not (Lord 1967). It can be shown mathematically that the models’ estimates are equal ($b_{DM} = b_{RM}$) (i) when there are no pre-existing group mean differences in the pre-test scores Y_1 or (ii) when the regression weight b_1 is set to 1. This corresponds to a perfect correlation between pre-test Y_1 and post-test Y_2 scores, which at the same time is an implicit assumption of the difference score approach (see Allison 1990; Lord 1967) (see also the formulas in the supplementary material).

In the field of observational surveys such as for summer setback studies, where data is not intended to meet experimental standards and the focus is not on causal effects, the situation of contradictory results across models, described as “paradoxical,” may also be understood as stemming from the ‘different foci’ of the two models (see Dumont and Ready 2019; Holland and Rubin 1982; Quinn 2015). We should note that the two models and their parameterization of conditional or unconditional learning progress³ can also be written as repeated measures models and hence be analyzed in a multilevel mixed regression framework (van Breukelen 2013), extending to analyses that accommodate more than two measurement occasions (e.g., Dumont and Ready 2019). Categorizing the European summer setback studies by their model choices, Verachtert et al. (2009) and Meyer et al. (2017), in line with the U.S. tradition in seasonal research, used piecewise growth curve modelling. This comes down to an unconditional modelling of learning progress⁴, which is similar to a difference score approach (see Dumont and Ready 2019). Also, Siewert (2013), comparing differences in group means between measurement occasions, modelled learning progress in an unconditional fashion. On the other hand, Becker et al. (2008), Lindahl (2001), and Paechter and colleagues (2015) opted for a lagged variable model approach and parameterized disparities in learning progress from a conditional angle.

The study outlined in the following section shows a real data example of a summer setback study, where results diverge in direction across the difference score and the regressor variable model. The purpose of presenting this study is a demonstration that such divergence in results does not only occur under some rare and special circumstances but rather can be encountered in every research setting. Together with the results reported by Quinn (2015) and Dumont and Ready (2019), it is meant to provide cumulative evidence that such diverging results also might be rather a rule

³ There may be some confusion as both models can be rewritten mathematically to include some conditioning on the pretest (see Allison 1990). We mainly use this wording in an intuitive fashion to make the point of interpretative differences in model results clear. Following the formulas as shown in the supplementary materials, conditioning of learning progress (posttest—pretest) on the pretest is further clearly distinct and distinguishes between the two models (see Werts and Linn 1979), whereas conditioning of posttest scores on pretest scores is not (see Allison 1990).

⁴ See footnote 3.

than an exception. The comparison between the two methodological approaches is helpful, because most published results report either the one or the other, so that the difference between the two cannot be assessed by the reader. We explicitly neglect measurement error in performance estimates, as we want to emphasize the fact that in descriptive studies models can arrive at incongruent results also in the absence of measurement error. Our focus in this manuscript is on observed scores rather than on true scores. One should note, however, that incongruent results across and within models can also arise from measurement error adjustments (see Quinn 2015). Further, we do not engage with differences in findings that may occur depending on whether absolute or relative learning progress is modelled. Because test score variances may change over time, the picture of inequality dynamics, also depends on whether and how gaps in learning progress are standardized (for more details, see Quinn et al. 2016). For the remaining, we work with unstandardized test scores.

3 An empirical example—controversy in findings

3.1 Study outline

This summer break study took place in Switzerland, based on a sample of 37 classrooms ($N_{\text{total}} = 664$ students) from the northwestern cantons of Basel-Landschaft and Aargau, which were assessed in the year of 2018 just before and just after their summer breaks. The sample is by no means meant to be representative of the entire student population. Rather, we have deliberately chosen students from school districts with particularly long and particularly short summer vacations. Because of students repeating a grade or leaving the school attendance areas and the exclusion of students with special needs, the final post-test sample size was reduced to $N_{\text{effective}} = 614$ students. Otherwise, we were not observing any noteworthy student drop out between the pre-test and the post-test as the participation in the post-test was mandatory for all students. Of these, 46% were girls and 54% boys. Two thirds of the students (66%) reported (Swiss) German as the language spoken at home, although only half of the sample (50%) had no history of family migration. The vast majority (93%) reported spending the most time during the summer break with their parents. Because of regional disparities in the start and length of the summer break, the pre-test assessment took place 1–3 weeks before the summer break in grade 5 of elementary education. Summer breaks vary between 4–6 weeks. The post-test assessment corresponded to the standardized educational evaluation (Check P6), which was mandatory for all students in grade 6 in these cantons and took place 4 weeks after the summer break.⁵ Grade 6 marks the final grade of elementary education in Switzerland, when the students are around the age of twelve years. Students' performance was assessed in the domains of mathematics and language (grammar and orthography). The standardized paper and pencil tests were developed to reflect the official school curriculum and included open response, multiple-choice, and multiple-response items. Additionally, a survey was conducted on individual attributes

⁵ See <https://www.check-dein-wissen.ch/>.

such as the students' gender and migrant status, their summer vacation activities, and socioeconomic variables of their familial environment. For more details on the summer break study see Tomasik and Gämperli (2019).

3.2 Measures and analyses

The pre- and post-tests were scaled according to the probabilistic 2 PL Item-Response Theory (IRT) model (Wu et al. 2016). Items with low item-total correlations and items showing bad model fit in view of the expected and observed item characteristic curves were excluded (Wu et al. 2016), so that the EAP reliability of the tests always exceeded $\rho = 0.80$. The test forms changed between pre- and post-tests, which means that the pre- and post-test consisted of different items. Due to previous calibration, the item parameters (difficulty and discrimination) of all items included in the tests were known in advance and we could link the students' scores from the pre- and post-test onto the same scale using an anchor-item approach (see Kolen and Brennan 2004). Students' abilities in math and language performance were estimated using weighted likelihood estimation (Warm 1989). These estimates were then linearly transformed to a scale of mean 600 (SD = 100) for the initial assessments in math and language. In math, the pretest sample mean is 600 (SD = 99) and the posttest sample mean is 599 (SD = 93). In language the pretest sample mean is 601 (SD = 97) and the posttest sample mean is 611 (SD = 77).

The socioeconomic status of students was operationalized as a composite measure based on children's reports on the highest educational level and occupational status of their parents (ISEI-08; Ganzeboom 2010a, 2010b) and the number of books available at home. Anything missing from the student's answers to these items was multiply imputed by predictive-mean-matching (PMM) (Robitzsch et al. 2016) with the approach of multiple imputation by chained equations (MICE; see van Buuren 2012) using the R-package *miceadds* (Robitzsch et al. 2018). Based on their socioeconomic status, students were then classified into four groups, which are given by the 25%-quantiles of this variable. Hence, in the study at hand, the group indicator included in the analyses consists of four SES groups. The division into four SES groups allows for a better illustration of the divergent findings by model-choice given the data at hand than would a more simple division into two groups. The regressor variable and difference score models were inputted in the software R. The analyses are based on multiply imputed data and final estimates were combined via Rubin's rules (Rubin 1987). Standard errors were further corrected for clustering due to the survey design, in which we assessed complete classrooms, using the R package *survey* (Lumley 2019). The simulations were conducted using the R-package *simstudy* (Goldfeld 2019).

3.3 Results

The descriptive results in Table 1 indicate, with a glance at the pre-test scores, that there are differences in pre-test achievement across SES groups in both math and language. The higher the SES group, the higher the pre-test scores on average. Looking at the difference in average scores between pre- and post-test (gap) by

Table 1 Descriptives: scores by SES group

	Pretest	Posttest	Gap
<i>Language</i>			
SES [1]	557	580	+23
SES [2]	595	606	+11
SES [3]	610	614	+4
SES [4]	644	643	−1
<i>Math</i>			
SES [1]	551	556	+5
SES [2]	591	592	+1
SES [3]	611	601	−10
SES [4]	648	649	+1

N = 163/164 students per SES group

Total *N* = 654

group, no summer setback is present for lower SES groups in either subject, language or math. If anything, the results even show that SES groups [1 and 2] (lowest quartiles), on average, gained in language and math scores between pre- and post-test, which contrasts with SES groups [3 and 4] (highest quartiles), who gained less or even experienced some loss. Our descriptive results hence contradict assumptions of disadvantageous effects of summer breaks for lower SES groups.

Estimating a **difference score model** [DM] (see Table 2), where individual differences in pre-post-test scores are regressed on SES group indicators, produces results very similar to the descriptive results. Because we have four SES group indicators *G*, the effect b_{DM} as measured by the difference score model is estimated in terms of group differences with respect to a reference category. Table 2 presents the results after running a difference score model, where the lowest SES group [1] is the reference category for the comparison of effects yielded for the other groups. In accordance with the descriptive statistics, the results for summer vacation effects in language suggest that lower SES groups gained in language skills, whereas for

Table 2 Difference score model: effects by SES group

	Effect	SE	Gap
<i>Language—difference score</i>			
Intercept—Ref. SES [1]	+23*	6	+23
SES [2]	−14	9	+9
SES [3]	−19*	9	+4
SES [4]	−24*	9	−1
<i>Math—difference score</i>			
Intercept—Ref. SES [1]	+5	6	+5
SES [2]	−1	9	+2
SES [3]	−14	9	−9
SES [4]	−3	8	+2

N = 654

Gap = Intercept (reference group) + effect of respective SES group

**p* < 0.05

Table 3 Regressor variable model: effects by SES group

	Effect	SE	Gap
<i>Language—posttest score</i>			
Intercept	+322	17	—
Pretest score	0.5	0.03	—
SES [2] ^a	+7	8	SES1 + 7
SES [3] ^a	+9	7	SES1 + 9
SES [4] ^a	+22*	7	SES1 + 22
<i>Math—posttest score</i>			
Intercept	+179	15	—
Pretest score	0.7*	0.03	—
SES [2] ^a	+9	8	+2
SES [3] ^a	+5	8	−9
SES [4] ^a	+28*	8	+2

N = 654

Gap = SES1 [Intercept + pretest score * effect] + effect SES[X]

^aRef. SES [1]* $p < 0.05$

the highest SES group no improvement is evident. The lowest SES group [1] gains significantly more in language over the summer break compared to the higher SES groups [3 and 4]. For math, similarly, lower SES groups do not lose skills in math over the summer break. It is mainly SES group [3] that scores lower in the post-test, though this decline is not significant in comparison to the small improvement of the lowest SES group [1] (reference category). All in all, no disadvantage is found for lower SES groups, rather the opposite if anything: On average, lower SES groups seem to experience a positive change in scores over the summer vacation.

When we estimate a **regressor variable model** [RM] (see Table 3), where post-test scores are regressed on the SES group indicators while controlling for pre-test scores, our results point towards a relative advantage for higher SES groups. Again, because we have four SES group indicators G , the effect b_{RM} as measured by the regressor variable model is estimated in terms of group differences with respect to a reference category. Controlling for pre-test scores, we find that the highest SES group [4] gained 22 points more in language and 28 points more in mathematics compared to the lowest SES group [1], which is a significant difference in improvement to the advantage of the higher SES group.

While the difference score model points towards a relative disadvantage for the lower SES groups, the regressor variable model points towards a relative disadvantage for the lower SES groups. So what do we make of these supposedly contradictory results, which may be interpreted as an example of Lord's paradox (Lord 1967)? This contradiction resolves itself when we see that for non-experimental data the two models answer different research questions. The parameterization by the difference score model [effect b_{DM}] shows the average learning progress by SES groups (independent of their pre-test levels), while the regressor variable model [effect b_{RM}] estimates average differences in learning progress by SES groups between students who achieved the same pre-test scores (that is, depending on their pre-test levels).

The difference score model answers the research question: On average, do lower SES students compared to higher SES students make different progress over the summer break? The regressor variable model answers the research question: On average, do lower SES students who achieved the same pre-test scores as higher SES students make different progress over the summer break? The data structure, which leads to these supposedly contradictory findings by model approach is explored in the following section.

Though it goes beyond the scope of this paper to discuss possible explanations of the counterintuitive effects in our real data example, we add a few tentative explanations here. Why do we find a general gain in skills over the summer, in particular a gain for lower SES groups? One explanation may lie in the study design and the preparation for high-stakes testing. Between the pre- and post-test there is some instruction time included, that is, between the end of the summer vacation and the post-test. Further, the post-test is a high-stakes test, the pre-test is low-stakes. After the summer vacation, teachers may provide tailored training in class and wrap-up the grade 5 school material in preparation for this high-stakes test (see Wiliam 2010; on the effects of high-stakes testing regimes). Especially lower ability students, which overlap with students of lower SES, may in total profit from such a wrap-up after the break. This may lead to the observed (unconditional) advantage for the lower SES group in learning progress (of course, there may additionally be some ceiling effects for high achievers present). Yet, comparing low and high SES students who started in the summer at the same pre-test levels, we, nevertheless, observe a conditional disadvantage for the lower SES students. Why is this? A tentative assumption explaining these dynamics may be that higher SES parents may expect more and generally provide more stimulating learning environments during the summer vacation and in preparation for the high-stakes test. This could explain the observation that students from higher SES families, who started in the summer at the same pre-test levels as their lower SES peers, achieve higher post-test scores. All in all, the models shed light on different dynamics of social disparities in (unconditional vs. conditional) progress over the summer break, which are simultaneously present and call for further exploration.

4 Constructed examples: a technical illustration of the problem

In the following, using simulation, we constructed two different examples to illustrate the correlational structure that may be present in the data that can lead to divergent findings depending on the statistical model chosen. The first scenario (1) sketches an example in which low SES students lose, on average, more during the summer break, and yet low SES students starting in the summer at the same pre-test levels as their higher SES peers experience smaller losses. Hence in example 1, unconditional disadvantage coincides with conditional advantage for low SES students. Intuitively, such a scenario may be present if, in general, lower achieving students lose more skills during the summer while the group of low SES students has a higher share of lower achievers. At the same time, the low SES group may be targeted by specific summer activities to counteract learning losses (e.g., educa-

tional vouchers for low SES group members), resulting in conditional advantages at respective pre-test levels through tailored support. The second scenario (2) sketches an example that mirrors the real data example outlined above, where on average low SES students gain in skills over the summer vacation, while compared to their peers of higher SES who started in the summer at the same pre-test levels they gain less. Here in example 2, unconditional advantage coincides with conditional disadvantage for low SES students. A tentative explanation for this scenario has been outlined above.

4.1 Scenario 1: summer loss

Higher total loss for the low SES group with relative advantages at respective pre-test levels. Simulating test scores, we assume a group of students ($N = 1000$) who achieved on average a mean score of $M = 500$ ($SD = 100$) in the pre-test ($Y1$) prior to their summer vacation. We further assume that in the absence of schooling, on average, the students experience a loss of 50 points ($SD = 50$) in skills over their vacation. In addition, we assume that there are two equal sized groups (G) of students (high vs. low SES students, coded 0 and 1) who achieved different mean score levels in the pre-test. Let's say the high SES group [1] achieved higher pre-test levels compared to the low SES [0] group. These differences in pre-test levels to the disadvantage of the low SES group can be captured by a *negative correlation between the grouping index (G) and the pre-test ($Y1$) score*. We assume this correlation to be $\text{corr}[Y1, G] = -0.6$, which results in pre-test group differences of about 120 points, where the low SES group scored on average $Y1_1 = 440$ compared to the high SES group scoring on average $Y1_0 = 560$. In scenario 1, the low SES group loses, on average, more skills over the summer. To model this, we assumed that there exists a *general positive correlation between the difference score $\Delta(Y2 - Y1)$ and the pre-test level ($Y1$)*, where students scoring higher in the pre-test lose (considerably) less skill over the vacation; higher pre-test levels coincide with lesser negative differences Δ . This trend is captured by a positive $\text{corr}[\Delta, Y1]$ of $+0.6$. Hence so far, we have a data structure where the SES groups diverge in pre-test levels and where loss clearly depends on pre-test levels, such that, on average, the low SES group loses more over the summer because this group comprises a higher share of lower achievers. Further, in scenario 1, we assume that over and beyond this general correlation between loss and pre-test levels, there is some group-specific component in loss. At respective pre-test levels, the lower SES students lose less compared to their higher SES peers. This shows in a (*minor*) *negative correlation between the group (G) and the difference score $\Delta(Y2 - Y1)$* , of $[\Delta, G] = -0.25$; indicating that though the low SES group loses comparatively more skill, this loss is not as big as we would expect it to be when focusing solely on the pre-test levels of this group. Hence all in all we assume a correlational structure for the constructed data whereby (i) the low and high SES groups achieved different pre-test levels (i.e., the groups encompass different shares of low vs. high achieving students), and (ii) skill decay depends on the pre-test levels of students (higher achievers lose less) while at the same time (iii) the skill decay of low SES students is less than expected given their pre-test achievement.

Fig. 1 Scenario 1: Higher total loss for the low SES group by relative advantages at pretest levels (*solid diagonal line*: overall relationship between pretest Y1 and learning progress ($Y2 - Y1$), *vertical lines*: average pretest levels ($Y1$) by SES group (*black*: high SES, *grey*: low SES), *horizontal lines* [*unconditional view*]: average learning progress ($Y2 - Y1$) by SES group (*black*: high SES, *grey*: low SES), *red markers* [*conditional view*]: learning progress ($Y2 - Y1$) by SES group at specific pretest levels (*cross*: high SES, *bullet*: low SES))

This correlational structure results in opposing inequality dynamics as shown in Fig. 1. The diagonal solid black line indicates the positive relationship between pretest achievement (x-axis) and progress (y-axis), where higher achieving students experience less skill decay, which is graphically displayed by less negative progress. The vertical lines indicate the pre-existing differences between SES groups in pretest achievement, where low SES students (grey dashed vertical line) achieved, on average, lower pre-test scores compared to the high SES students (black dashed vertical line). Because of these pre-existing group differences and the positive relationship between pre-test achievement and progress, the progress of the low-SES students (grey dashed horizontal line) is, on average, lower compared to the progress of the high-SES students (black dashed horizontal line). In other words, because progress is mainly negative, lower SES students lose more skills compared to higher SES students. However, at respective pretest levels, the average progress of students of low SES is higher compared to the average progress of the high-SES students with these respective pretest scores. This is depicted in Fig. 1 by the red markers. As example, within intervals of ± 25 around pretest scores of 400, 500 and 600 points, low SES students outperform the high SES students. Around pretest scores of 600, the progress of low and high SES students converges.

Such a correlational structure (see also the supplementary material) leads to the result that the difference score model shows, as assumed, that the lower SES group loses about 25 points more on average. Hence unconditionally, the low SES students are at a relative disadvantage because they comprise more lower achieving students and lower achieving students lose more skills over the summer. Yet conditionally they lose less than we may expect given their pre-test levels. Hence the regressor variable model recovers that at each pre-test level, low SES students do not lose as

much as one would expect by their pre-test levels; namely, they lose about 17 points less compared to higher SES peers who started in the summer with the same pre-test levels.

4.2 Scenario 2: summer gain

Slightly higher gains for the low SES group with relative disadvantages at respective pre-test levels. Simulating test scores for scenario 2, we again assume a group of students ($N = 1000$) who achieve on average a mean score of $M = 500$ ($SD = 100$) in the pre-test ($Y1$). Again we assume that there are two equal sized and pre-existing groups of students (G =high vs. low SES students, coded 0 and 1) who achieve different mean score levels in the pre-test. These differences in pre-test levels, to the disadvantage of the low SES group, are again captured by a *negative correlation between the grouping index (G) and the pre-test level ($Y1$)* of $\text{corr}[Y1, G] = -0.6$, which results in pre-test group differences of about 120 points, where the low SES group scored on average $Y1_1 = 440$ compared to the high SES group scoring on average $Y1_0 = 560$. In contrast to scenario 1, we assume that, on average, the students experience a minor gain of 20 points ($SD = 50$) over their vacation. The low SES group, on average, gains even more than the high SES group. This shows in a *negative correlation between the pre-test ($Y1$) and the difference score $\Delta(Y2 - Y1)$* of $\text{corr}[\Delta, Y1] = -0.4$. Higher achievers in the pre-test gain less over the summer, which coincides with the high SES group gaining less than the low SES group because it has a higher share of higher achievers. However at respective pre-test levels, scenario 2 assumes that the lower SES students gain less compared to their higher SES peers. This can be induced by a (*minor*) *positive correlation between the*

Fig. 2 Scenario 2: Slight gain for the low SES group by relative disadvantages at pretest levels (*solid diagonal line*: overall relationship between pretest $Y1$ and learning progress ($Y2 - Y1$), *vertical lines*: average pretest levels ($Y1$) by SES group (*black*: high SES, *grey*: low SES), *horizontal lines [unconditional view]*: average learning progress ($Y2 - Y1$) by SES group (*black*: high SES, *grey*: low SES), *red markers [conditional view]*: learning progress ($Y2 - Y1$) by SES group at specific pretest levels (*cross*: high SES, *bulet*: low SES))

group index (G) and the difference score $\Delta(Y_2 - Y_1)$ of $\text{corr}[\Delta, G] = +0.1$; indicating that the high SES group still gains comparatively more skill than expected, given their higher achievement in the pre-test. Hence all in all, we assumed a correlational structure for data construction, whereby (i) the low and high SES groups achieved different pre-test levels (i.e., the groups encompass different shares of low vs. high achieving students), and (ii) where skill gain depends on pre-test levels of students (higher achievers gain less) while at the same time (iii) the skill gain of high SES students is larger than expected given their pre-test achievement.

This correlational structure results in the opposing inequality dynamics as shown in Fig. 2. The diagonal solid black line indicates the negative relationship between pre-test achievement (x-axis) and progress (y-axis), where higher achieving students experience less progress. The vertical lines indicate the pre-existing differences between SES groups in pretest achievement, where low SES students (grey dashed vertical line) achieved, on average, lower pre-test scores compared to the high-SES students (black dashed vertical line). Because of these pre-existing group differences and the negative relationship between pre-test achievement and progress, the progress of the low-SES students (grey dashed horizontal line) is, on average, marginally larger compared to the progress of the high-SES students (black dashed horizontal line). However, at respective pre-test levels, the average progress of students of low SES is lower compared to the average progress of the high-SES students. This is depicted in Fig. 2 by the red markers. As an example, within intervals of ± 25 around pretest scores of 400, 500 and 600 points, the low-SES students gain less compared to the high-SES students.

This correlational structure in the data now leads to the finding that by using the difference score model, we see that on average the lower SES group gains slightly more (+10 points). Yet applying a regressor variable model reveals that lower SES students still gain a fair amount less over the summer at each respective pre-test level (about -22 points) compared to higher SES peers, who started in the summer at the same pre-test levels. Note that these effects can also be recovered using the formulas given in the supplementary material.

5 Discussion and conclusion

Summer break study designs have been widely used in educational research to disentangle school from non-school influences because summer breaks provide a natural experimental setting that allows for the measurement of learning progress when school is not in session. In line with others (see Dumont and Ready 2019; Quinn 2015), this paper calls for greater consideration to the modeling strategy (difference score vs. regressor variable model) used for the analysis of social disparities in progress over the summer, in particular among European seasonal comparative scholars who differ in their model preferences. Depending on how learning progress is modelled, which is as conditional or unconditional progress, the models reveal different insights into the dynamics in social inequality present in the data (see also Dumont and Ready 2019; Quinn 2015). Because we usually do not know the un-

derlying data generating structure, cross-model comparison of effects is prone to fallacy and premature conclusions.

Though by design the summer break pre- and post-tests may seem to mirror a natural experiment, the groups analyzed for differential progress over the summer most often are not randomized; rather, they occurred naturally and achieved different pre-test levels in the initial test. Because of this, the models can yield supposedly paradoxical findings, with coefficients differing in magnitude and direction across models (see Dumont and Ready 2019; Holland and Rubin 1982; Quinn 2015). The situation of conflicting results across models is known in the (quasi-)experimental literature as Lord's paradox (Lord 1967). For survey data and non-causal analyses, this paradox simply resolves in the differential model foci (Quinn 2015). Drawing on Quinn (2015) and Dumont and Ready (2019), this paper, by including a recent summer setback study conducted in Switzerland and some constructed examples, has illustrated in an intuitive fashion why and when the models reveal different dynamics in inequality in the data. The constructed examples showed that, technically, a correlational structure in the data, whereby there exists a (non-zero) *correlation between the grouping (G) and pre-test ($Y1$) variable (pre-existing groups)*, in combination with the *direction and magnitude of the correlation between the pre-test ($Y1$) and the difference score Δ ($Y2 - Y1$)* and the *correlation between the grouping (G) and the difference score Δ ($Y2 - Y1$)*, determines whether or not the two models (difference score vs. regressor variable) will yield opposing results. That is, whether or not the unconditional progress of the groups differs in direction and magnitude from the conditional progress of the groups (see for more details the formulas in the supplementary material). In essence, the difference score model explores progress between pre- and post-test measurements from an unconditional angle. This means that the model allows a comparison of whether socioeconomically distinct student groups make different progress over the summer, regardless of whether these groups started at different initial levels at the beginning of the summer. The regressor variable model focuses on the data from a conditional angle. In other words, the regressor or lagged variable model allows for a comparison of whether students of socioeconomically distinct backgrounds who started in the summer at the same pre-test levels make differential progress. Whereas in the former we compare average differences in progress across groups (i.e. across students of different baseline levels), in the latter we focus on differential progress between students of the same baseline levels across groups. Following Quinn (2015), we argue that the question of which model to use for the analysis of social disparities in learning progress may not simply be answerable methodologically but may also need to be considered in view of the content of the research. Is the research interested in revealing whether socioeconomically distinct student groups make the same progress when school is not in session, regardless of their prior achievement (unconditional equality)? Or is the research interested in comparing the progress of students of distinct socioeconomic backgrounds of similar prior achievement levels (conditional equality)?

Overall, there seems to be no tradition in model choice in the European seasonal comparative research, which is only at its beginning (Meyer et al. 2017; Paechter et al. 2015). The European studies often refer in their research body to the U.S. seasonal studies and to other European findings. Model choice, though, has not

been taken into consideration when drawing lines between similar and divergent findings across studies (see Becker et al. 2008; Coelen and Siewert 2008; Lindahl 2001; Meyer et al. 2017; Paechter et al. 2015). Future European seasonal comparative research should be more precise and sensitive towards the (unconditional or conditional) parameterization of progress modelled in their interpretation and in the comparison of their findings as there are substantial differences in what they actually mirror and what their potential implications might be. A thorough examination of both unconditional and conditional dynamics in inequality may also be fruitful as it allows researchers to gain a more in-depth understanding of the social disparities in learning progress in their data.

This paper focused on summer break research, which primarily involves the investigation of socioeconomically distinct progress between pre- and post-test measurements. Yet the main conclusions are not limited to summer break research. Sensitivity towards the different perspectives on change (conditional and unconditional) is relevant for an accurate interpretation of any pre- or post-test research based on non-randomized survey data. Thus, awareness of the differential foci of modelling unconditional versus conditional change is of relevance to the field of social sciences in general.

Supplementary Information The online version of this article (<https://doi.org/10.1007/s11618-021-01009-4>) contains supplementary material, which is available to authorized users.

Acknowledgements We want to thank Marianne Gämperli for the project management, data preparation, and initial analyses of the summer setback study as basis for her master's thesis.

Funding This work was supported by the foundation for research in science and the humanities at the University of Zurich.

Funding Open access funding provided by University of Zurich

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ainsworth, J. W. (2002). Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces*, 81, 117–152.
- Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: a seasonal perspective. *Educational Evaluation and Policy Analysis*, 23, 171–191.
- Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. *American Sociological Review*, 72, 167–180.
- Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, 20, 93–114.
- Becker, M., Stanat, P., Baumert, J., & Lehmann, R. (2008). Lernen ohne Schule: Differenzielle Entwicklung der Leseleistungen von Kindern mit und ohne Migrationshintergrund während der Sommerferien. In F. Kalter (Ed.), *Migration und Integration* (pp. 252–276). Wiesbaden: VS.
- Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1, 289–328.

- Boudon, R. (1974). *Education, opportunity, and social inequality: changing prospects in western societies*. New York: Wiley.
- Bourdieu, P. (1971). Bildungsprivileg und Bildungschancen: Auslese und Gnadenwahl. In P. Bourdieu (Ed.), *Die Illusion der Chancengleichheit* (pp. 19–45). Stuttgart: Klett.
- Bourdieu, P. (1985). Das Bildungswesen ein maxwellischer Dämon? In P. Bourdieu (Ed.), *Praktische Vernunft* (pp. 36–41). Frankfurt a. M.: Suhrkamp.
- Bourdieu, P., & Passeron, J.-C. (1977). *Reproduction in education, society, and culture*. London: SAGE.
- Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, 53, 371–399.
- van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: the difference. *Multivariate Behavioral Research*, 48, 895–922.
- Burkam, D. T., Ready, D. D., Lee, V. E., & LoGerfo, L. F. (2004). Social class differences in summer learning between kindergarten and first grade: model specification and estimation. *Sociology of Education*, 77, 1–31. <https://doi.org/10.1177/003804070407700101>.
- van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton: CRC.
- Cameron, C. E., Grimm, K. J., Steele, J. S., Castro-Schilo, L., & Grissmer, D. W. (2015). Nonlinear Gompertz curve models of achievement gaps in mathematics and reading. *Journal of Educational Psychology*, 107, 789–804.
- Caro, D. H., McDonald, J. T., & Willms, J. D. (2009). Socioeconomic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education*, 32, 558–590.
- Coelen, H., & Siewert, J. (2008). Ferieneffekte. In T. Coelen & H.-U. Otto (Eds.), *Grundbegriffe Ganztagsbildung* (pp. 411–440). Heidelberg: Springer.
- Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: a narrative and meta-analytic review. *Review of Educational Research*, 66, 227–268.
- Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, 69, 613–635.
- Dumont, H., & Ready, D. D. (2019). Do schools reduce or exacerbate inequality? How the associations between student achievement and achievement growth influence our understanding of the role of schooling. *American Educational Research Journal*. <https://doi.org/10.3102/0002831219868182>.
- Entwisle, D. R., & Alexander, K. L. (1992). Summer setback: race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review*, 57, 72–84.
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2000). Summer learning and home environment. In R. D. Kahlenberg (Ed.), *A notion at risk: preserving public education as an engine for social mobility* (pp. 9–30). New York: Century Foundation Press.
- van Ewijk, R., & Slegers, P. (2010). The effect of peer socioeconomic status on student achievement: a meta-analysis. *Educational Research Review*, 5, 134–150.
- Fairchild, R., & Noam, G. G. (2007). *Summertime: confronting risks, exploring solutions* (New Directions for Youth Development, Number 114). San Francisco: Jossey-Bass.
- Ganzeboom, H. B. (2010a). International standard classification of occupations ISCO-08 with ISEI-08 scores. http://www.harryganzeboom.nl/isco08/isco08_with_isei.pdf. Accessed: 12 July 2018.
- Ganzeboom, H. B. (2010b). A new international socioeconomic index (ISEI) of occupational status for the international standard classification of occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007. In *Annual Conference of International Social Survey*. Lisbon.
- Gastro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference score for longitudinal research. *Journal of Social and Personal Relationships*, 35, 32–58.
- Gershenson, S. (2013). Do summer time-use gaps vary by socioeconomic status? *American Educational Research Journal*, 50, 1219–1248.
- Goldfeld, K. (2019). Simstudy: simulation of study data. R package version 0.1.13. <https://cran.r-project.org/web/packages/simstudy/simstudy.pdf>. Accessed: 2 March 2019.
- Helbling, L. A., Tomasik, M., & Moser, U. (2019). Long-term trajectories of academic performance in the context of social disparities longitudinal findings from Switzerland. *Journal of Educational Psychology*. <https://doi.org/10.1037/edu0000341>.
- Hippel, P., & Hamrock, C. (2019). Do test score gaps grow before, during, or between school years? Measurement artifacts and what we can know in spite of them? *Sociological Science*, 6, 43–80.
- Holland, P. W., & Rubin, D. B. (1982). *On Lord's paradox*. Princeton: Educational Testing Service.
- Jennings, J. L., Deming, D., Jencks, C., Lopuch, M., & Schueler, B. E. (2015). Do differences in school quality matter more than we thought? New evidence on educational opportunity in the twenty-first century. *Sociology of Education*, 88, 56–82.

- Jencks, C., & Mayer, S. E. (1990). The social consequences of growing up in a poor neighborhood. In L. E. Lynn & M. G. H. McGeary (Eds.), *Inner-city poverty in the United States* (pp. 111–186). Washington: National Academy Press.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd edn.). New York: Springer.
- Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.
- Lindahl, M. (2001). *Summer learning and the effect of schooling: evidence from Sweden* (IZA Discussion Paper No. 262). Bonn: IZA.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304–305.
- Lumley, T. (2019). Survey: analysis of complex survey data. R package version 3.36. <https://cran.r-project.org/web/packages/survey/survey.pdf>. Accessed: 2 March 2019.
- Meyer, F., Meissel, K., & McNaughton, S. (2017). Patterns of literacy learning in German primary schools over the summer and the influence of home literacy practices. *Journal of Research in Reading*, 40, 1–21. <https://doi.org/10.1111/1467-9817.12061>.
- Paechter, M., Luttenberger, S., Macher, D., Berding, F., Papousek, I., Weiss, E. M., & Fink, A. (2015). The effects of nine-week summer vacation: losses in mathematics and gains in reading. *Eurasia Journal of Mathematics, Science and Technology Education*, 11, 1339–1413.
- Patterson, M. V., & Rensselaer, N. Y. (1925). The effect of summer vacation on children's mental ability and on their retention of arithmetic and reading. *Education*, 46, 222–228.
- Quinn, D. M. (2015). Black-white summer learning gaps: interpreting the variability of estimates across representations. *Educational Evaluation and Policy Analysis*, 37, 50–69. <https://doi.org/10.3102/0162373714534522>.
- Quinn, D. M., Cooc, N., McIntyre, J., & Gomez, C. J. (2016). Seasonal dynamics of academic achievement inequality by socioeconomic status and race/ethnicity: updating and extending past research with new national data. *Educational Researcher*, 45, 443–453.
- Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (pp. 259–294). Vienna: facultas.
- Robitzsch, A., Grund, S., & Henke, T. (2018). Miceadds: Some additional multiple imputation functions, especially for 'mice'. R package version 2.15-22. <https://cran.r-project.org/web/packages/miceadds/miceadds.pdf>. Accessed: 2 March 2019.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Shinwell, J., & Defeyter, M. A. (2017). Investigation of summer learning loss in the UK—implications for holiday club provision. *Frontiers in Public Health*. <https://doi.org/10.3389/fpubh.2017.00270>.
- Siewert, J. (2013). *Herkunftsspezifische Unterschiede in der Kompetenzentwicklung in der Kompetenzerwartung: weil die Schule versagt? Untersuchungen zum Ferieneffekt in Deutschland*. Münster: Waxmann.
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: a meta-analytic review of research. *Review of Educational Research*, 75, 417–453.
- Stanat, P., Baumert, J., & Müller, A. G. (2005). Förderung von deutschen Sprachkompetenzen bei Kindern aus zugewanderten und sozial benachteiligten Familien. Evaluationskonzeption für das Jacobs-Sommerscamp Projekt. *Zeitschrift für Pädagogik*, 51, 856–875.
- Tomasik, M. J., & Gämperli, M. (2019). *Soziale Herkunftseffekte und regionale Unterschiede in den Veränderungen von Schulleistungen von Schülerinnen und Schülern während der schulfreien Zeit: Kurzbericht zuhanden der Stiftung für wissenschaftliche Forschung*. Zurich: Institut für Bildungsevaluation.
- Verachtert, P., Damme, J. V., Onghena, P., & Ghesquière, P. (2009). A seasonal perspective on school effectiveness: evidence from a Flemish longitudinal study in kindergarten and first grade. *School Effectiveness and School Improvement*, 20, 215–233.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450.
- Werts, C. E., & Linn, R. L. (1979). A general linear model for studying growth. *Psychological Bulletin*, 73, 17–22.
- Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist*, 45, 107–122. <https://doi.org/10.1080/00461521003703060>.
- Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers. Theory into practice*. Singapore: Springer.