School start times and academic achievement - a systematic review on grades and test scores

Authors:

Anna M. Biller[1,2*], Karin Meissner[1,3], Eva C. Winnebeck[1#] & Giulia Zerbini[4*]

*corresponding authors

Affiliation:

[1] Institute of Medical Psychology, Ludwig Maximilian University Munich, Munich, Germany

[2] Graduate School of Systemic Neuroscience, LMU Munich, Germany

[3] Division of Health Promotion, Hochschule Coburg, University of Applied Sciences & Arts, Coburg, Germany

[4] Department of Medical Psychology and Sociology, University of Augsburg, Augsburg, Germany

current affiliation: Neurogenetics, Technical University Munich, and Institute for Neurogenomics, Helmholtz Center Munich, Munich, Germany

Contact information:

anna.biller@med.uni-muenchen.de
Institute of Medical Psychology, Goethestrasse 31, 81373 Munich, Germany

giulia.zerbini@med.uni-augsburg.de
Department of Medical Psychology and Sociology, Stenglinstraße 2, 86156 Augsburg, Germany

Abstract

School start times have been at the centre of many scientific and political debates given the accumulating evidence that bell times are generally too early, and thus lead to an epidemic of sleep restriction in the student population. Recent media attention has conveyed the message that later school starts not only improve sleep but also result in better academic achievement. Several studies have been recently published on this topic requiring a comprehensive review of the results to clarify the relationship between later school start times and academic achievement to inform the general public and policy makers.

To this end, we conducted a systematic review of the current literature on school starting times and academic achievement in middle and high school students, considering grades and standardised test scores as achievement measures. We followed the PRISMA guidelines for searching, including, and reporting relevant literature and identified 21 studies for detailed analysis. Evidence quality of included studies was assessed with a pre-defined risk of bias assessment using modified items from the GRADE scheme and ROBINS-I tool.

About half of the reviewed studies reported no (positive or negative) effect of delaying school times on grades and test scores, while the other half reported either mixed or positive results. Given the strong heterogeneity of included studies, we grouped them according to various characteristics, such as academic outcomes, dose of delay, evidence quality, or study design to identify potential hidden effects. Despite this, we could not identify any generalisable effect beyond single studies as to whether delaying school times has clear beneficial effects on academic performance.

Given that grades and scores determine future career trajectories and predict future success, the question whether school start times contribute to academic achievement is of great interest for the general public and needs to be further clarified. Mechanistically, it is very likely that improved sleep leads to or mediates improved cognitive performance and learning, but definitive conclusions on whether this also translates into better grades and scores across all students requires better evidence at this stage. Importantly, this does not preclude other positive outcomes of later start times such as improved sleep (quality), motivation or learning but draws attention on current gaps and shortcomings. To this end, we also highlight critical methodological aspects and provide suggestions to increase the evidence-level and to guide the direction of research in future studies.

Keywords: *school starting times, academic performance, grades, scores, adolescence, sleep*

Introduction

Early school start times (SSTs) have been recognized as one of the leading causes of inadequate sleep in teenagers worldwide. They clash with the longer and later sleep needs of teenagers[e.g. 1–4], leading to wide-spread, chronic sleep restrictions in the student population[e.g. 5–8]. Because of the accumulating evidence that sleep restriction is detrimental for psychological[9–11] and physical health[12,13], some schools (mainly in the US) have delayed their SSTs during the past decades.

Several studies - although mostly short-term and cross-sectional - have documented the beneficial effects of delaying SSTs on sleep duration and daytime sleepiness (as reviewed in[14–16]). More recently, other outcomes with regards to SSTs have been investigated, such as cognitive and academic performance. Since short sleep has been linked to detrimental effects on learning, memory, and cognition[17–23], it is fair to hypothesize that delaying SSTs could result in better academic achievement (e.g. as measured in grades or scores) mediated by longer sleep duration, improved sleep quality or better circadian alignment.

However, early findings from field studies on this topic are very heterogenous, likely due to methodological differences in outcome variables and study designs[24,25]. For instance, academic achievement has been operationalised in different ways (e.g. self-reported grades, single final grades, grade point averages, standardised test scores) and with different scales. In addition, study designs vary considerably across studies, and achievement is influenced by many student- and school-level factors[e.g. 26–30].

Previous reviews have mostly summarized the effects of delaying SSTs on several different variables (e.g., sleep, tardiness rates, absences, motor vehicle accidents and health[14,16,25]). We identified a total of 12 peer-reviewed reviews[15,16,37,38,24,25,31–36] – only 3 of them systematic reviews[15,16,34] – that discuss SSTs in relation to academic achievement. However, all of them cover the topic only broadly or on the side, so that no unifying conclusion can be drawn from the existing reviews to date. Despite this lack in systematic reviews and meta analyses, newspaper articles often purport it as established scientific fact that later SSTs improve academic achievement[39–41], while some public outreach programs also convey this message[42], mostly referring to single studies that found positive associations.

Since academic achievement shapes future career trajectories[43–45], answering the question whether delaying SSTs improves achievement goes beyond simple and genuine scientific curiosity - a rigorous and up-to-date analysis of the accumulating evidence is warranted. Following the PRISMA guidelines for systematic reviews and including a detailed risk-of-bias assessment based on items from the GRADE scheme[46] and the ROBINS-I tool[47], we addressed the specific gaps in the review literature to date, such as a particular need for discussion of the quality of evidence, a detailed description of the outcome variables and statistical analyses, and a distinction between middle/high school and college students, who differ considerably in their sleep characteristics and class schedules. In our review, we thus systematically assessed the existing evidence on SST effects on academic achievement via studies on course grades or standardised test scores in middle and high school students. We provide both a summary as well as detailed descriptions of each included study, assess the overall and individual evidence level and highlight critical points for future research.

Methods and Materials

Literature search

Our focused question was whether changes in school start times in middle or high schools (or international equivalents) have any effect on academic achievement as measured in (standardised) test scores or course grades (both subjectively and objectively reported). Therefore, we conducted a systematic electronic literature search in Web of Science and PubMed via Endnote (version 9.3.1), and an online search on SCOPUS in August 2020, which was updated in November 2020. All languages, article types or year of publications were allowed. The following search terms were used (in title, abstract or keywords):

school start times OR *school start time* OR *school starting times* OR *school start delay* OR *start late* OR *start early*

AND

grades OR *school performance* OR *academic performance* OR *test scores* OR *standardised scores* OR *achievement*

Additionally, reference lists of previous reviews and articles were scanned to ensure complete retrieval. We included two unpublished articles that are currently under review in peer-reviewed journals[48,49]. The PRISMA flowchart (Fig. 1) was followed to adhere to preferred reporting guidelines for systematic reviews[50].

Study selection criteria

All duplicate retrievals were removed via the Endnote duplicate function followed by a manual search and deletion round. All remaining titles and abstracts were subsequently screened on relevance with regards to the focused question. Full articles were then searched if the following study selection criteria were fulfilled: academic achievement was assessed as grades or (standardised) test scores; participants were middle school or high school students; articles included both a change/variation in SSTs and a measure of academic achievement (course grades or standardised test scores).

Data abstraction and analysis

The recommended PRISMA guidelines for data synthesis and systematic reviews were followed[50]. AMB and GZ independently and systematically extracted pre-defined study characteristics as per Figure 1.
The 21 studies included in the final qualitative synthesis were grouped by their study design. Please note that we grouped according to the design underlying the grade or test score outcomes, which can differ from the design for other outcomes investigated in the respective study such as sleep duration. Identified designs were: longitudinal designs with control group, longitudinal designs without control group, and cross-sectional designs. Note that a longitudinal design means that the *same students* were followed over several time points (within-subject comparisons) whereas a cross-sectional design compares *different students* either at one time point or between time points (between-subject comparisons). It was noticed that several cross-sectional studies described their design as longitudinal because they followed the same *schools or districts* over several time points (which might or might not include the same students); we considered these studies as *repeated cross-sectional studies* based on

their study design and statistical analysis. Authors were contacted when information was missing, not clearly defined or further analyses were available upon request. If authors responded, information was updated accordingly. If authors did not answer or failed to provide necessary information in the original article, this was marked as "not available" ("NA") in Tab. S1 and, and flagged orange or red (depending on the severity) in the reporting bias category of the risk of bias assessment (Tab. 1).

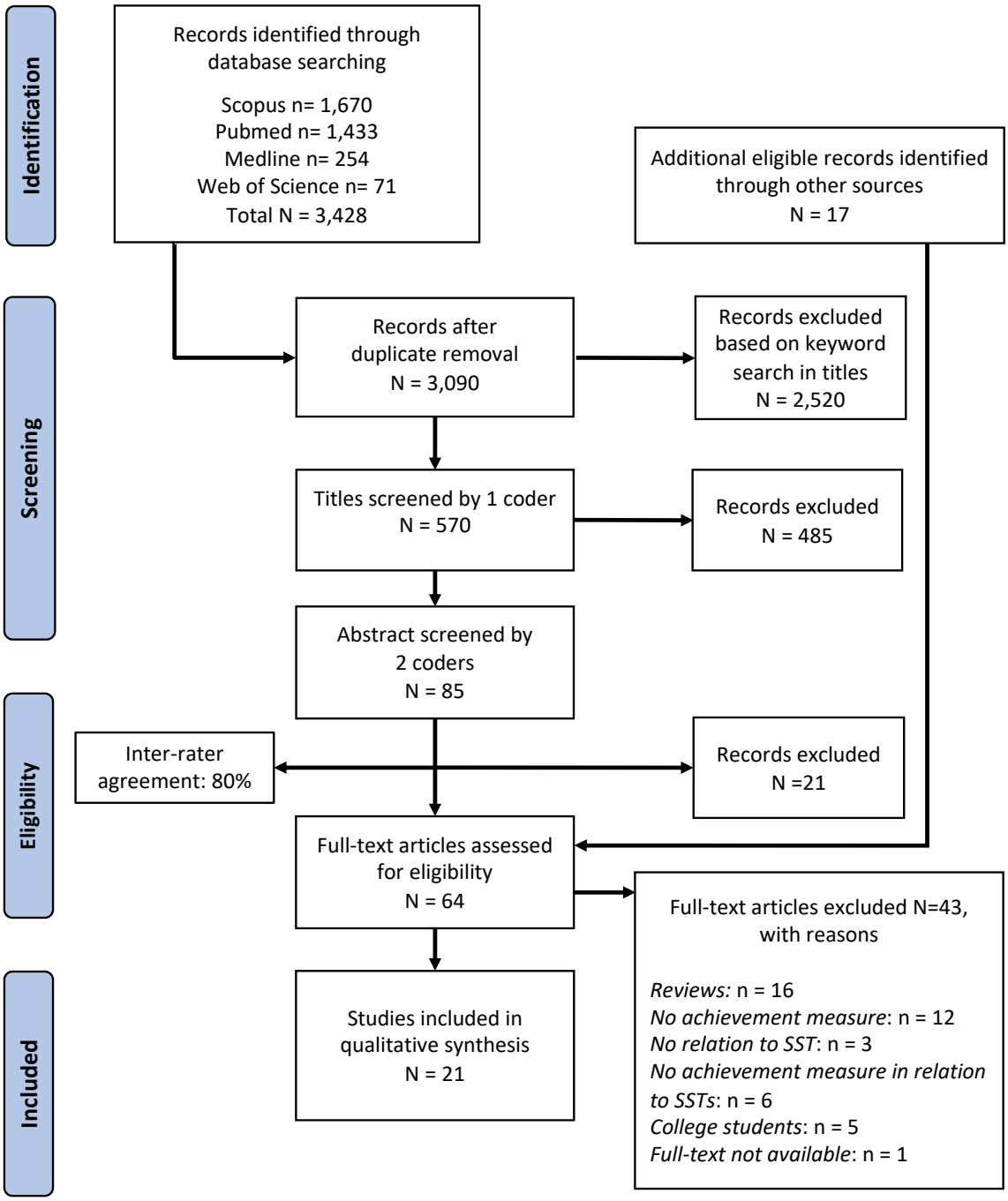

Fig. 1 | PRISMA flowchart. The PRISMA flow diagram for our systematic review process detailing the database searches, the number of identified records, titles and abstracts screened, the final studies included in qualitative synthesis and reasons for exclusion of studies.

Risk of bias assessment

A pre-defined risk of bias assessment was conducted independently by AMB and GZ (Tab 1). Given that there were no randomised controlled trials (RCTs) in the final sample and the large methodological differences between studies, bias assessment guidelines were adapted as there are no standard guidelines for non-RTCs. To this end, items from the GRADE scheme[46] and ROBINS-I tool[47] used for non-RTCs were included and modified. Each study was evaluated on the following bias categories and flagged green (low risk), orange (intermediate risk) or red (high risk):

Selection bias (randomisation): Participants were not randomly assigned to the control group or the treatment group. Non-RTC are high risk by definition.

Allocation concealment: Researchers did not know the sequence or method of randomisation and hence could not predict the next allocation. Non-RTC are high risk by definition.

Reporting bias on author level: Authors did not report or only partially reported all outcome variables, sources of outcomes, statistical analyses or general information necessary to judge the study. When a publication stated that information was available upon request, the authors were contacted.

Responder bias on student level: Students could be biased when self-reporting, which is not the case for objectively reported grades or scores provided by official sources (e.g. the registry or state level administrations).

Performance bias (blinding of participants/personnel): Participants who knew that they took part in a study are prone to behavioural changes (Hawthorne effect). If informed consent was obtained, students were considered unblinded, else they were blinded. This also covers a potential self-selection bias towards taking part in a study.

(Dis)similarity of baseline characteristics: Authors checked and reported the (dis)similarity of baseline characteristics between cross-sectional groups or between control and treatment groups.

Appropriate statistical models: Statistical analyses accounted for confounders and were appropriate for the given study design.

Cohort bias (control group present): Longitudinal changes might be due to cohort characteristics and not due to an intervention when no control group was present. Only applies to longitudinal studies.

Tab. S1 lists the decision criteria underlying the risk of bias assessment. In cases where the assessment differed between AMB and GZ, mutual agreement was sought after discussion of critical points. In case no agreement could be reached, two independent scorers (ECW and KM) evaluated the respective studies and a consensus was found across all scorers. From the assessment, a total quality-of-evidence score was calculated as follows: scores for each bias category were added up (green contributed 1 point, orange 0.5 points and red 0 points) and then divided by the maximal possible score (8 for the longitudinal studies with control group, 7 for the longitudinal studies without control group, and 7 points for cross-sectional studies). The quality-of-evidence score was the proportion (%) of the maximum score (e.g. 6 out of max 8 points = 75%). The different bias categories were not weighted. We defined scores <25% as low, ≥25% and <75% as moderate and ≥75% as good.

Results

Literature search

A total of 3,428 articles were identified based on the automated search in title, abstract and keywords, of which 3,090 remained after duplicate removal (Fig. 1). Due to this large number, a second automated search was carried out on titles only, resulting in 570 articles. One coder (AMB) then screened titles excluding 485 manually due to irrelevant titles. The abstracts of the remaining 85 studies were screened by both coders (AMB and GZ), who agreed on 47 studies (80% inter-rater agreement) and additionally identified 17 studies through reference lists of included studies. The identified 64 articles were subsequently screened in full by both coders, 43 excluded based on the pre-defined exclusion criteria and 21 ultimately included in the qualitative synthesis (Fig. 1).

Study characteristics and quality

In the following paragraphs, summary information concerning all included studies are reported (see also Fig. 2 and Tab. 2).

School type and cohort characteristics

The majority of studies collected data in high schools (>900 schools), of which 2 were also boarding-schools[51,52], 2 grammar schools and 2 vocational schools[53]. Other school types were middle schools (>140) and elementary schools (85, not considered here). In one study, school type was not specified[54]. The sample sizes varied drastically between 157 to >770,000 individual students and up to >1 Mio number of observations (e.g. individual grades). However, some authors did not distinguish between number of individuals, number of schools and number of observations. In 13 studies, age of participants was reported and ranged approximately between 11-19. Most studies were conducted in the US (13)[51,52,62–64,54–61], followed by South Korea (4)[48,65–67], Germany[49], Croatia[53], England[68], and one unknown location[69] (Fig. 2a and Tab. 2). Gender ratios, ethnicity/race and a proxy for socioeconomic status (SES; free or reduced lunch eligibility) were not consistently reported.

Study designs

We identified longitudinal (within-subject) and cross-sectional (between-subject) studies. The 11 longitudinal studies all included a change in SSTs and hence had an intervention group[48,49,67,51,52,55–57,64–66]. However, only 6 studies had an additional control group with no change[48,56,65–67] or advance of SSTs[55] (Fig. 2b). Of the cross-sectional studies, 4 studies compared schools in various districts without an intervention but based on their different school start times[53,58,61,69]. The rest included a change in SST, providing mostly repeated cross-sectional comparisons of schools or districts over roughly one[60] or several years[54,59], or at one time point after the change[62,63]. One cross-sectional study also had an A-B-A design, in which the school start delay during phase B was abolished to return to baseline start time (A) after 2 years[68].

Statistical analyses

A vast range of different statistical analyses was reported (Tab. S1 and Fig. 2c). Notably, regressions were the dominant analysis method, ranging from general OLS regressions[54,55,58,59,65], quantile

regression[55], difference-in-difference methods[48,56,65,67] and binomial regression[60,66] to linear mixed models[49,57] and path analysis with probit regression[69]. One study reported Oster models with bounded effects and instrumental estimates[58]. Another study used MANOVA[61], while several simpler analysis methods not controlling for covariates were also used. These were t-tests[57,63,68], X^2-tests[52], Mann-Whitney Test[53] and correlations[63]. Notably, several studies did not report statistical analyses[51,62,64].

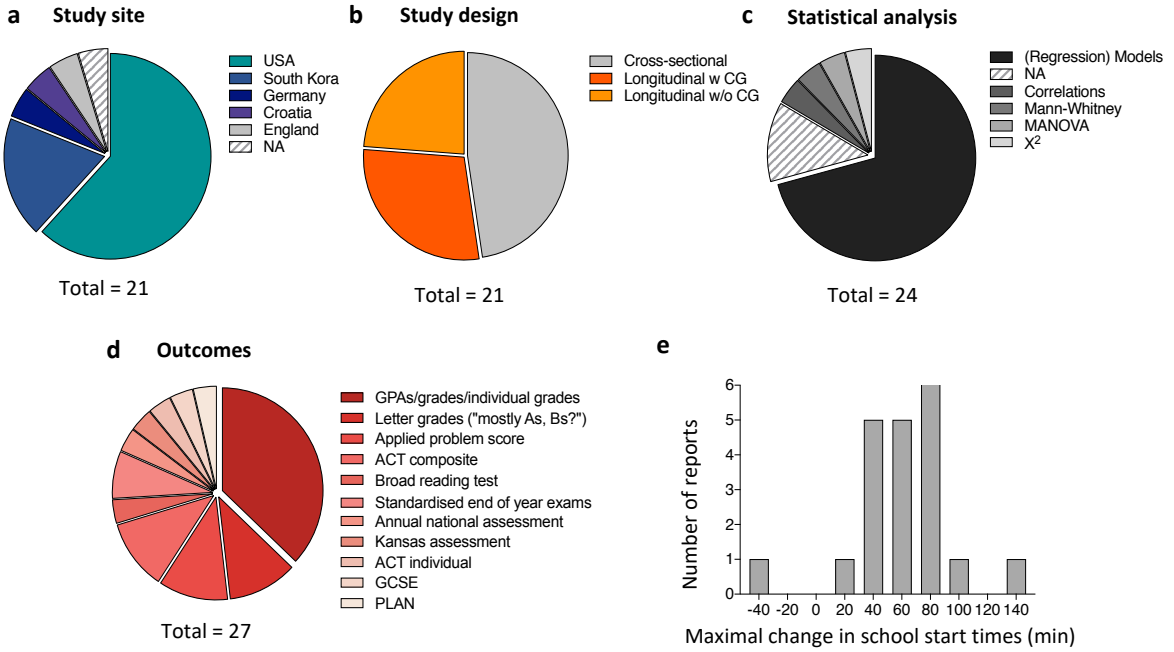

Fig. 2 | Characteristics of included studies. a-d, Pie charts depicting key characteristics of the 21 studies included in the final review. Since several studies used multiple types of analysis or assessed multiple outcomes, the total number in c,d is >21. **e,** Histogram displaying the magnitude of the school start changes reported in the 21 studies. When a study reported ranges, the maximum of the range was taken. Please note that these numbers therefore just provide a rough overview and are not precise. Abbreviations: NA, not available; w, with; w/o, without; CG, control group; GPAs, grade point average; ACT, American College Test; GCSE, General Certificate of Secondary Education; PLAN, a preliminary ACT test discontinued in 2014.

Study outcome measures

About half of the studies provided grades as outcome measures, while the other half provided (standardised) test scores. However, since some studies did not provide explanations whether scores originated from standardised tests, a clear distinction between course grades and test scores was not always possible. Clearly defined scores were ACT scores (American College Test)[54,56,59], national achievement scores or PLAN scores[63], standardised test scores from Regents Exams[57], standardised end-of-course exams[59], annual national assessment of achievement in South Korea[67], GCSE in the UK (General Certificate of Secondary Education)[68], and Woodcock-Johnson Revised Test of Basic Achievement scores[58], all of which were objectively reported (except for Groen *et al.*, for which the source was unclear[58]) (Fig. 2d). The remaining studies analysed other types of objective scores or grades [48,49,55,59,61,64,65], subjective grades[51,52,60,62,63,66,69], and in one study the outcome was unclear[53]. Sampling resolution was mostly once per year, the highest reported resolution was once per academic quarter[49].

Amount of school start time change and duration of exposure to the new start time

The SST delay reported was on average 64min (median=60, SD=26) with a range of 25-135 min (Fig. 2e). This average is based on the maximal delay reported by each study (several reported multiple amounts) and thus an approximation. Since some studies only provided SST ranges or a minimal start delay, the numbers are not precise. In 2 studies, SSTs were actually advanced by 40 min and 25-45 min respectively[55,56]. One study changed to a flexible SSTs in which students could choose daily whether to attend school at 8:00h or 8:50h[49]. Exposure duration to the (new) start time ranged from as little as 3 months to 7 years (Tab. 2). However, several studies did not clearly state the timeframe (so we inferred where possible), or did not test a change but a difference in start times across schools.

Summary of individual study results

In the following, we shortly report findings of individual studies, grouped by study design. In summary, 5 studies found clear positive effects of delayed school starts on academic achievement[48,55,62,68,69], 5 reported mixed effects[58,59,61,63,67], 9 did not detect significant effects[49,51,52,54,56,57,64–66], 1 reported negative effects[53], and one study's finding was unclear[60] (Fig. 3b). Regarding SST *advances*, 1 study reported negative effects[69], while another did not[56].

Notably, of the 21 studies, 4 studies investigated the same 9 o'clock policy in South Korea[48,65–67]. Although they considered partly different outcomes and schools (middle vs high schools), the Korean studies likely analysed data from overlapping students, hence this cannot be regarded as entirely independent evidence. The same may apply to 2 studies by Wahlstrom *et al.* conducted in the same district: the report in 2002[64] might be a longitudinal follow-up of the report from 1997[62], but we were unable to confirm this.

Longitudinal studies with control group

Edwards (2012)[55] followed several middle schools in Wake County, North Carolina (USA), over 8 years (up to $N_{observations}$>102,000) during which 9 schools delayed, 4 advanced and 11 did not change their SSTs. The authors analysed objective standardised end-of-year test scores in reading and math via regression models with pooled OLS models and accounted for various covariates both on the student and school level. They found that a 1h later school start corresponded to a 1.8-2.9 percentile increase in math (0.06-0.07 SD) and 1.0-3.6 increase in reading (0.04-0.05 SD) when adjusted for covariates (both significant), and that the effect was stronger for lower achieving students.

Jung (2018)[65] followed 85 elementary and 63 middle schools ($N_{students}$>4,000) in South Korea 3 years prior to and 2 years after a delay from 8:00h-8:20h to 9:00h. Participants were recruited as part of the Gyeonggi Education Panel Study and their objective Korean, English and math course grades were reported. The author found no effect of delaying SSTs on grades in the longitudinal within-subject comparison with the control group (difference-in-difference estimation/OLS estimation). Cross-sectional analyses as robustness check confirmed the longitudinal results. Similar to Kim [67] and Biller *et al.*[49], the author also found that when not controlling for covariates, test scores increased, while the effect became statistically non-significant when covariates were added.

Kim (2018)[67] also compared high schools from two districts in South Korea ($N_{students}$>2,000), of which Gyeonggi adopted a 9 o'clock start time policy. Pre-change SSTs in this district ranged from 7:40h-9:00h and were delayed to 9:00h post-change, while Seoul did not change (control group). The author used the difference-in-difference method and mixed within-between regression models to estimate the influence of the 9 o'clock policy on the objective Annual National Assessment of Educational Achievement for 9th and 11th graders, and the College Scholastic Ability Test (CSAT) for 12th graders (data cover 5 years pre and 2 years after the change). Only male 11th graders showed a significant increase of 0.06-0.08 SD for math, even after adjusting for confounders. CSAT scores did not increase significantly with the 9 o'clock policy.

Similarly, **Rhie and Chae (2018)**[66] studied middle and high schools in 4 South Korean districts, of which Gyeonggi delayed SSTs (baseline from a range of 7:30h-8:10h) to 9:00h, while Daegu, Gyeongbuk and Ulsan did not change (SSTs ranged from 7:30h to 8:00h; control group). Based on logistic regression analysis in their large sample ($N_{students}$>42,000), they found that self-reported GPAs increased year by year in both the intervention and the control group (data cover 2 years pre and after the change; adjustments for covariates not reported).

Shin (2018)[48] is the fourth study that investigated the South Korean 9 o'clock policy effects in Gyeonggi (change in SST from around 8:20h to 9:00h), compared to Seoul (control group), but the author used objective semester grades as outcome and focused on middle schools ($N_{observations}$>33,000). The data span 2 years and was analysed using the difference-in-differenced method, which accounted for various individual and school-level variables. Shin reported a 0.03 SD increase in math and 0.02 SD increase in reading grades when adjusted for time trending (both significant).

Lenard *et al.* (2020)[56] looked at 19 high schools in Wake Country, North Carolina, USA, of which 5 had advanced their SSTs from 8:05h to 7:25h, while the control group (14 high schools) kept their start at 7:25h. They found no significant change in objective standardised American College Test (ACT) scores, neither in their longitudinal nor their cross-sectional comparison of about $N_{students}$~10,000 in 8 cohorts. The authors used a difference-in-difference approach and comparative interrupted series controlling for various individual and school-level variables. Their data spanned 4 years prior and 7 years after the change.

Longitudinal studies without control group

Biller *et al.* (2021)[49] investigated in a German secondary school the effects of a unique SST change, the introduction of *flexible* SSTs, on objective, quarterly grades for up to 2.5 years prior to and 1.5 years after the change. In the flexible system, students chose daily whether to attend school at 8:00h or 8:50h after starting predominantly at 8:00h in the conventional system before. Longitudinal linear mixed model analyses of 16,724 grades in 12 academic subjects ($N_{students}$=157) indicated that the flexible system did not affect grades when accounting for several student and school-level factors.

Boergers *et al.* (2014)[51] studied an independent high school (boarding school) in Rhode Island, USA, that delayed its start time from 8:00h to 8:25h ($N_{students}$=197). The percentage of students who reported to obtain "mostly Bs or better" changed from 93% to 91% after 2 months, however statistical analyses were not reported.

Owens *et al.* (2010)[52] used the same outcome variable as Boergers *et al.*[51] in their study of $N_{students}$=201 from an independent high school (boarding and day school) in Rhode Island, USA, over 6 months (3 time points of assessment). They found that a school start delay from 8:00h to 8:30h was associated with a non-significant increase of students reporting to obtain "mostly Bs or better" (82% pre vs 87.1% post, using a χ^2 test). Adjustment for covariates was not reported.

Thacher & Onyper (2016)[57] studied $N_{students}$~800 from one public high school in Glen Falls, NY, USA, which delayed their SSTs from 7:45h to 8:30h. They used mixed effect analyses to analyse longitudinal effects (2 years before and 2 after the change), adjusting for multiple covariates and including moderator effects. This analysis indicated no systematic effect on subjectively reported GPAs (0-100%) nor subject-specific GPAs or standardised test scores (Regents exam). They did find positive effects for 11th graders' overall GPAs, however, only when they ran cross-sectional comparisons (increase from 78.79% to 81.34%). In contrast, no systematic effects on individual academic subjects were found in this cross-sectional analysis. In fact, 2 out of 20 subjects were significantly worse after the change and also Regents exam scores decreased significantly.

Wahlstrom (2002)[64] investigated the effect of later SSTs in 7 high schools in Minneapolis, Minnesota, USA, for 3 years before and after the change from a 7:15h to an 8:40h-start. The study analysed objective letters grades and found small improvements that were not statistically significant. However, no actual numbers (or the letter grade scale), nor any statistical test were reported.

Cross-sectional studies

Groen and Pabilonia (2019)[58] studied $N_{students}$=1200 from a sample of 790 U.S. high schools and reported that a 1h-delay in high school start times was associated with significantly increased reading scores (but not math scores) by 0.16 SD for females, while no significant effect was found for males. The authors used OLS models, including many covariates (individual, family, high school, and community characteristics) that were added sequentially to the models. Data were collected for 2 years, sampled once per year.

Hinrichs (2011)[54] found no association between SSTs and ACT scores ($N_{students}$> 196,000) after a delay of 85 minutes from 7:15 to 8:40 AM in 73 schools in Minneapolis, USA, when accounting for various student-level and district level covariates and the length of the school day using OLS regression models (9 years of data). In a similar analysis, the author also found no effect on Kansas assessment scores in reading, maths, science, and social disciplines including all public high schools in Kansas (1,666 schools; up to 5 years of data). In another sample of 75 schools in 19 districts in Virginia, USA, again no association was found between delayed SST and test scores in standardised end-of-course exams (8 years of data).

Bastian and Fuller (2018)[59] analysed 4 years of data from $N_{students}$>770,000 in 410 high schools in North Carolina, USA, of which 23 changed their start times (9 schools by ≥30 min). The authors tested both the influence of a linear SST delay per 1h and a categorised school start depending on actual start time i) on overall and 1st period course grades, ii) standardised end-of-course exams, and iii) ACT scores. Linear regression models adjusting for several student and school-level covariates showed that only a start at 8:30h or later was associated with significant improvements of 0.05 SD in 1st period course

grades. Importantly, this was one of the only studies that particularly focused on specific subgroups of students: especially low-performers, students with a minority background and with a low SES benefitted from later starts (0.05-0.07 SD in course grades and up to 0.28 SD in ACT composite scores per 1h).

Dunster *et al.* (2018)[60] reported results from a cross-sectional pre-post comparison of a one-semester biology course grade from 2 high schools in Seattle, USA, which delayed their starts from 7:50h to 8:45h. Median grade was 77.5% in the year before the delay and 82% in year after the delay ($N_{students}$=178, ~7-month exposure). In logistic regression, where *grade* was used as a predictor variable (not outcome), an increase in grades significantly increased the odds of a student stemming from the after-delay-cohort rather than from the before-delay-cohort (quantitative results not reported; covariates: school, mood, chronotype, sleepiness and sleep offset time on schooldays). Interpretation of the results is limited by the use of grades as independent variable and adjustment for sleepiness and sleep offset in the analysis.

Milić *et al.* (2014)[53] analysed in a one-off assessment the final semester grades of 4 Croatian schools (grammar and vocational schools) with morning and afternoon schedules: 2 schools followed early schedules (alternating between 7:00h and 13:00h), while 2 schools had later schedules (8:00h and 14:00h). Based on the sample of $N_{students}$=821 and Mann-Whitney Test (no covariates), it was concluded that students attending the early schedules got significantly better grades (72.0% vs 65.6% in the later-scheduled schools). An extra caveat of the unadjusted analysis is that the sample in the early scheduled schools consisted of three times more boys and grades often associate with gender (see e.g. Fig 3).

Wolfson *et al.* (2007)[61] compared the average fall-quarter grade (0-100%) of a total of $N_{students}$=205 attending either an early middle school (starting at 7:15h) or later middle school (starting at 8:37 AM) in New England, USA. MANOVA results with school, grade and gender as covariates indicated that, after half a year of exposure, 8th graders in the later school obtained significantly better objectively reported grades than their early school peers (83.79% vs 76.85%) while no difference was found for 7th graders.

Lewin *et al.* (2017)[69] compared 26 middle schools (unknown location) clustered into 3 groups depending on their SSTs (earliest, early, late). The authors obtained self-reported grades ("mainly As", "mainly Bs", "mainly Cs", "mainly Ds/Fs") and sleep duration from $N_{students}$>32,000 in 3 years. Path-analyses with probit regression with grades as outcome, sleep duration as mediator and inclusion of several covariates showed significantly better grade estimates in the late group compared to the earliest group but not the early group. This effect was in a similar magnitude as that of gender (females better than males) and ethnicity (non-whites worse than whites). Free lunch status as a proxy for SES had the greatest predictive value for grades, while the influence of sleep duration as a mediator was smaller but still significant.

Wahlstrom *et al.* (1997)[62] compared middle and high schools in three districts in Minnesota, USA, of which District A delayed its start time to 8:30 and Districts B and C stayed with their earlier starts of 7:25h and 7:15h, respectively. The study reports that mean self-reported grades in high schools in district A were highest compared to the other 2 districts, however the scale, statistical analyses and the use of covariates were not reported. Results for middle schools (7-8th graders) were comparable but again no statistics were given and absolute differences were marginal.

Kelley et al. (2017)[68] followed an English high school over 4 years across two changes in SST (A-B-A design): from the standard school start at 8:50h in year 0, to a delayed start at 10:00h for two years,

and for another year back to the original start time. For each year, national examination results (GCSE exams, achievement over the past 2 years plus final examination) of a different cohort of students was assessed ($N_{students}$>2,000). As the only study in our selection, the study analysed the achievement of the cohorts not only in comparison to each other but also in comparison to a national benchmark and to an indicator of predicted progress (value-added prediction) based on student cohorts' past achievements. Based on t-tests (no covariates), the study found that delayed SST for two years were significantly associated with an increased percentage of students making good academic progress (*i.e.,* achieving ≥5 GCSE grades of ≥C in English, math and ≥3 other subjects) (from 34% in year 1 to 52% in year 3) and with a 12-percentage point increase in the value-added number of students. Both improvements were partly reversed after return to the earlier start time.

Wahlstrom *et al.* (2014)[63] analysed self-reported grades, objectively reported GPAs, and standardised test scores (state-wide achievement tests or PLAN) from 9-12th graders after a high school start delay from 7:35h-7:50h to 8:00h-8:55h in Minnesota, Colorado, and Wyoming, USA. The study yielded mixed and mostly not-significant effects using t-tests and correlations without considering covariates.

Risk of bias assessment

To judge the evidence quality of the included studies, we performed a risk of bias assessment (Tab 1). Overall, since none of the studies were RCTs, selection bias was high by definition for all studies. Furthermore, in many studies, basic reporting standards were only partially met (reporting bias), blinding was a high concern in over half of the studies (performance bias), and appropriate statistical models that control for confounders were not used in 7 out of 21 studies. This meant that over half of the studies stayed below 75% of the good-evidence-score within their respective study design category. Therefore, the quality of the evidence can be deemed only moderate.

On the positive side, especially the longitudinal studies with a control group showed a high evidence quality with 2[56,65] out of 6 studies reaching at least a 75%-score and 3 more studies[48,55,67] >50%. Two studies[48,67] could have improved their score to 75% simply by ensuring sufficient reporting of outcomes and statistical analyses. Furthermore, all included studies had appropriately large sample sizes (and/or high resolution) and were therefore very likely suited to detect a true effect (sufficient statistical power).

Tab. 1 | Risk of bias assessment. Included studies are ordered by their study design (used for grade or score analyses) and assessed in different bias categories. Cell colour shows the risk status for the respective bias category (red=high risk; orange=intermediate; green=low risk). Question marks indicate ambiguous information (more details given in Tab. S1). For the final study result based on the obtained evidence score, an upward arrow indicates a positive finding for later school start times on academic achievement, a right arrow indicates mixed findings. Longitudinal studies assess the same individual over time, while cross-sectional studies follow different students. NA, not applicable.

	Longitudinal studies (within-subjects) with control group						Longitudinal studies (within-subjects) without control group					Cross-sectional studies (between-subjects)									
	Jung 2018[1]	Lenard 2020	Edwards 2012	Shin 2018	Kim 2018	Rhie 2018	Biller 2021	Thacher 2016[1]	Wahlstrom 2002	Owen 2010	Boergers 2014	Groen 2019	Hinrichs 2011	Bastian 2018	Lewin 2017	Kelley 2017	Wolfson 2007	Dunster 2018	Milić 2018	Wahlstrom 2014	Wahlstrom 1997
Randomisation (selection bias): non-RCT are high risk by definition	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Allocation concealment (selection bias): non-RCT are high risk by definition	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Reporting bias on author level: selective reporting of outcomes and statistical analyses	+	+	+				+	+	-	+	-	+	+	+	+		+	?	+	-	-
Responder bias on student level: Subjective vs. objective grades or scores[2]	+	+	+	+	+	-	+	+	+	-	-	+	+	+	-	+	+	+		+	-
Blinding of participants/personnel (performance bias)[3]	+	+	+	+	+	-	-	-	?	-	-	+	+	+	-	+	-	-	-	-	-
(Dis)similarity of baseline characteristics reported/checked	+	+	+	+	+	+	N.A.	N.A.	N.A.	N.A.	N.A.	+	+	+	+	-				?	
Appropriate statistical models which control for confounders	+	+	+	+	+		+		-	-	-	+	+	+	+	-		?	-	-	-
Control group present and used for statistical comparisons (cohort bias)	+	+		+	+	+	-	-	-	-	-	N.A.	N.A.	N.A.	N.A.		N.A.	N.A.	N.A.	N.A.	N.A.
Total score[4]	6/8	6/8	5.5/8	5.5/8	5.5/8	2.5/8	3/7	2.5/7	1.5/7	1/7	0/7	5/7	5/7	5/7	3/7	3/8	3/7	3/7	2/7	1/7	0.5/7
At least 75% good evidence score	✓	✓										✓	✓	✓							
At least 50% good evidence score	✓	✓	✓	✓	✓							✓	✓	✓							
Results	⇨	⇨	⇧	⇧	⇧⇨							⇧⇨	⇨	⇧⇨							

[1]These studies also included a cross-sectional analysis (between-subjects comparison) for their grades or scores analyses; either as a robustness check or as secondary analysis. For results on these see the result section and Tab.2.

[2]Subjective if students themselves reported their grades or scores; objective if the school, registry or any other administration reported the grades or scores.

[3]Blinding refers to informed consent; yes(unblinded), no (blinded). If data are solely obtained from archives, students are considered to be blinded. This also covers a potential self-selection bias towards taking part in a study which is eliminated in archive studies.

[4]Total score is constructed from the maximal number of available bias categories within a study type. Green=1 point; orange=0.5 points; red=0 points.

Magnitude of effects

Given that a meta-analysis was not indicated because of the great variation in outcome measures, study designs and analyses, we identified a subgroup of studies that provided standardised beta coefficients to compare the magnitude of grade or score changes across studies and with those of covariates. The statistically significant beta coefficients of the identified 8 studies are plotted in Figure 3a. Overall, the magnitude of the influence of school start times is smaller than SES or ethnical/racial background. In line with this, studies from Edwards[55] and Bastian and Fuller[59] demonstrated that it was the disadvantaged and minority students that particularly benefitted from later starts. Importantly, Figure 3a purports a biased picture towards positive results of school start times on achievement given the selection for significant standardised coefficients. Figure 3b puts findings from 3a into perspective with all included studies and paints a very different picture.

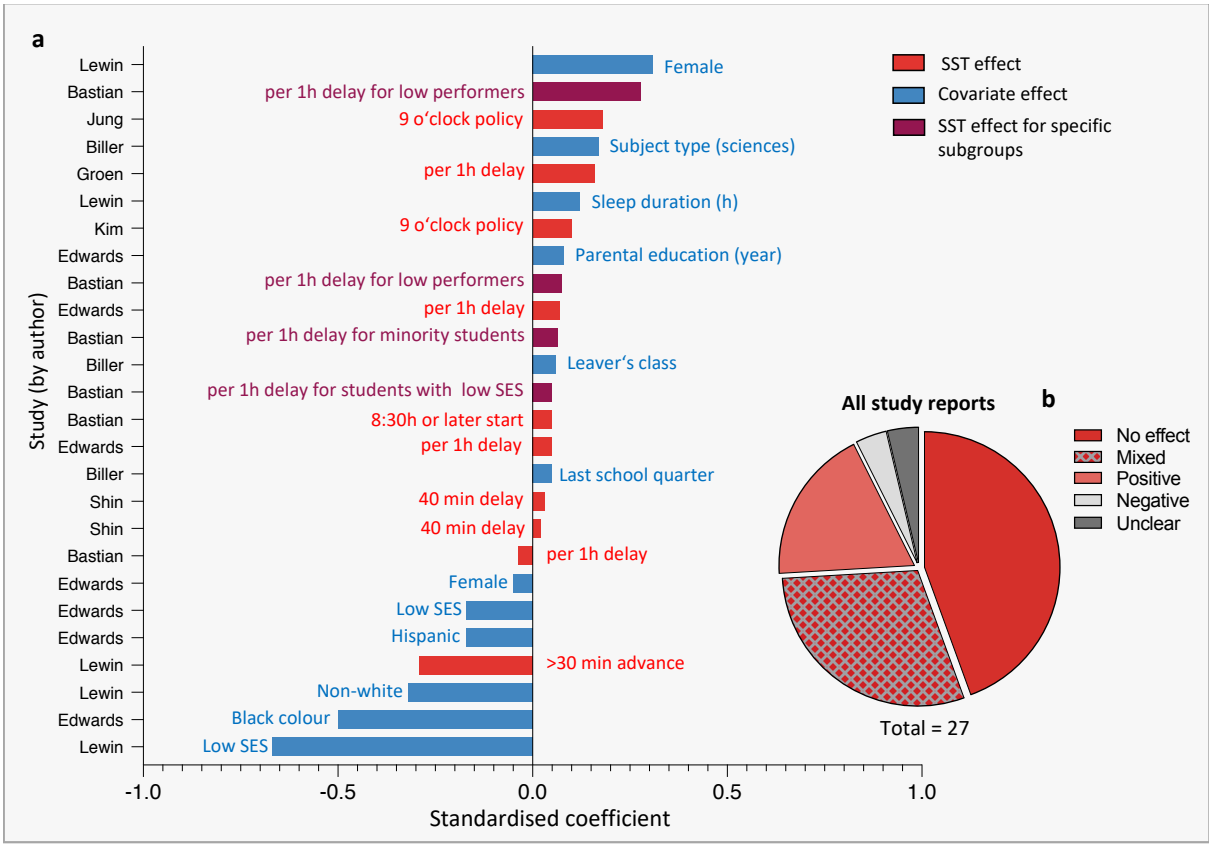

Fig. 3 | Overall study results and effects sizes. a, Standardised beta coefficients ordered by magnitude and study author from the subset of studies that reported standardised coefficients and statistically significant effects (n=8 of 21). Only these statistically significant effects are depicted, non-significant ones were left out. Standardised coefficients are in units of standard deviation of the outcome variable. Quarter refers to the academic quarter of a school year in Germany. Low socioeconomic status was measured as free lunch status. For exact study references see Tab. 1. **b,** Summary of simplified findings from all included studies (N=21). The total is >21 since several studies reported multiple outcomes. Abbreviations: SST, school start time; SES, socio-economic status.

Discussion

Chronic sleep restriction in teenagers has become a serious health concern worldwide[e.g. 8,70]. The widespread sleep restriction is largely a result of the conflict between the late sleep times typical of adolescence and the early SSTs imposed by society[e.g. 3,71,72]. Delaying school start times has the great potential of improving cognitive functioning, physical health and well-being of students mediated by improving sleep (as reviewed elsewhere[16,25,31]) with possibly relatively little costs[73,74]. But does a delay in SSTs also translate into improved academic achievement in middle and high school students? Our systematic literature search identified 21 studies that investigated whether SSTs have any systematic effect on course grades or (standardised test) scores. The analyses revealed that about half of the studies did not find any effect (neither positive nor negative) of delaying school times on academic performance, while the other half found either mixed or positive results. Just one study found better grades associated with earlier SSTs[53]. Given the high risk of biases observed in most of the studies and the great heterogeneity in school settings, there is a need for more high-quality evidence to draw sound conclusions (see Fig. 4 for suggested improvements).

Methodological considerations

Our systematic risk of bias assessment showed that the evidence level was mostly moderate (only 5 out of 21 studies achieved a score of ≥75% within their category). Specifically, we did not identify any randomised controlled trials, which is not surprising considering the circumstances of educational research and the hesitation of many schools to participate in such complex and time-consuming study designs[75]. In many studies, basic reporting standards were only partially met, blinding was a high concern in over half of the studies (i.e. high performance bias), and appropriate statistical models which control for confounders were not used in 7 out of 21 studies.

The study design we identified as one of the most commonly used was longitudinal studies with a pre-post design that followed a specific cohort of students over time i) including a control group that did not change start times, and ii) without a control group. A second common design was cross-sectional studies that compared different, independent groups of students (either at one specific time point or over several years) with varying start times. Studies that performed best in the risk of bias assessment were mostly longitudinal studies with a control group, a large sample size and with appropriate and advanced statistical analyses that controlled for possible confounders.

Well-designed studies also reveal no clear picture

What do studies with low risk (*i.e.,* a ≥75% good-evidence-score) conclude about the influence of SSTs on academic achievement? Lenard *et al*. (2020)[56] found that advancing SSTs by 40 minutes did not affect ACT scores, while Jung (2018)[65] showed that delaying start times by 40-60 min also did not affect grades when personal covariates were controlled for. If studies with a good-evidence-score of 50% are also considered, the picture is more complex: two studies report small gains in math and reading[48,55], and one reports small effects on math but not on Korean nor English[67]. Three cross-sectional studies also achieved a good evidence score of over 75%[54,58,59]. The associations found between SSTs and academic achievement again did not point in one direction: Groen and Pabilonia (2019) considered a range of different start times and reported small increases on the Woodcock-Test but only for females and reading[58], while Hinrichs did not find any positive association of a delay of 85 min on either ACT scores,

Kansas assessment scores, or end of course exams[54]. Bastian and Fuller (2018) reported that a 8:30h or later start was necessary for positive associations with 1st period grades[59]. Furthermore, the authors showed that especially lower achievers, minority students and students with a low SES benefit from later starts. In summary, good evidence studies report either no, relatively small, or not generalisable effects of changing SSTs.

Do results for course grades and standardised test scores differ?

Since course grades and standardised scores possibly measure different underlying skills and knowledge, they might also differ in their sensitivity to SST changes. For instance, standardised test scores seem to be sensitive enough to reflect effects of other school policies, *e.g.* reducing classroom size[76] or racial segregation[77]. However, general test scores might be less sensitive to acute changes in SSTs because they measure the accumulated knowledge over several schooling years[56] and are taken predominantly by high-achieving students, who are prone to ceiling effects. Moreover, they are often scheduled in the morning[54] and therefore confounded by time-of-day effects on attention and fluid intelligence (*e.g.* logic, reasoning, problem solving)[56,78–80]. Moreover, in the case of ACT or PLAN scores, tests are usually only taken by high-achieving students applying for admission to college – a specific student population which is prone to ceiling effects, making these students less likely to benefit from later SSTs compared to lower-achieving students as two other studies also confirmed[55,59].

Course grades, on the contrary, derive from exams taken by all students. If collected with high temporal resolution (*i.e.* more than once per year), they are potentially more sensitive to acute SSTs changes and less influenced by time-of-day effects if distributed evenly across the day. However, grades might be more influenced by certain student characteristics, such as conscientiousness or perseverance[81]. A "teacher bias" could particularly influence the results of interventional studies if not controlled for. Altogether, both standardised test scores and course grades have their pros and cons, which might be the reason why no clear answer emerges even when results are grouped by outcomes: there was no tendency or differential effect on either objective test scores (2 positive[55,68], 3 null findings[54,56,57], and 4 mixed results[58,59,63,67]) or objective grades (2 positive[48,60], 4 null findings[49,57,64,65] and 2 mixed results[59,61]) or self-reported grades (2 positive[62,69], 3 null findings[51,52,66], and 1 each for mixed[63] and negative[53]).

Considerations of power and dose

An alternative explanation for these mixed results could be a lack of statistical power. However, almost all studies had very large sample sizes or number of observations and were able to detect other influences such as gender differences and achievement gaps between whites and non-whites. The effect sizes of these factors tended to be of larger magnitude than effect sizes for changes in school starts (Fig. 3a).

Another interesting consideration is that effects of changed SSTs on achievement might not be linear. When exactly should schools start? How much should schools delay their start times? How long do students need to be exposed to later starts until effects become visible? These are important practical questions that are, however, difficult to answer. Intuitively, one would expect that small delays are not enough to produce robust effects. However, it is not clear whether further delays would be beneficial or even harmful. Hinrichs[54] tried to model this hypothesis using spline regressions but found no clear answer. Furthermore, the latest start time in the studies reviewed here was 10:00h and the largest delay

was 135 min (Fig. 2e). Despite a great variation in delays and SSTs, we were not able to detect any clear dose response curve, *i.e.* positive effects only appearing with the largest delay. Further studies should clarify this question. Nevertheless, the American Association of Pediatrics recommends to start schools not earlier than 8:30h[82], which is supported by Bastian and Fuller[59] who found that only when school started at 8:30h or later, significant positive effects were detected on 1st period grades, although overall grades were unaffected.

A second consideration about dose is how long the school has already operated in a delayed system – the longer the delay has been in place, the longer students were exposed. Several studies analysed time trends for several years before and after a change but no unifying results emerge from these studies.

Factors influencing academic achievement

A very likely reason for inconclusive results derives from the many variables affecting course grades and test scores. Whether these variables are assessed, considered, and controlled for can drastically change the conclusions of a study. These influences range from student-level factors (*e.g.* chronotype[83], ethnic or racial background[59], conscientiousness[81] or prior knowledge[84]) to family-level factors (*e.g.* parental involvement[85], parental education[65], or SES[86]), and to classroom- and school-level factors (*e.g.* classroom size[76] and atmosphere[84], teacher quality[87]). Indeed, we also observed here that SES and race/ethnicity influence achievement (Fig. 3a). Moreover, there are sleep variables, such as sleep duration and daytime sleepiness that play an important role for health, cognition and learning and are often connected to demographic variables, such that students with difficult social backgrounds are also prone to reduced and poorer sleep than their more advantaged peers[88,89]. It is therefore likely that SST delays potentially only translate into meaningful grade or test score benefits in a specific subset of students. Stratified analyses could answer this question but have rarely been done (for notable exceptions see[55,59] which confirm such tendencies). In general, reflecting on confounders, their influence on academic achievement and on how they might also be affected by changes in SSTs is important for designing future studies.

Limitations of the review

Although an extensive search across different databases was carried out, an incomplete retrieval of all published articles on the topic cannot be excluded. A total of 21 studies were included, which is far more than in previous reviews (2-12 included studies). We also chose to report grey literature to reduce a possible publication bias in favour of positive results. Previous reviews[16] decided otherwise to ensure a good quality of the findings reported. However, the included risk of bias assessment allowed for critical reporting of both peer and non-peer-reviewed articles. Since the studied population was restricted to middle and high school students, several studies which used valuable randomisation at the class-level had to be excluded because they included college students (for a review see[24]). However, life-style and sleep characteristics widely differ between high school and college students, which is why we focused only on adolescents. We included middle schools, since sleep changes tend to start with the onset of puberty [90,91].

Final Conclusions

Our systematic research and analysis of the literature shows that the current evidence does not allow to draw sound conclusions as to whether delaying SSTs improves or is associated with increased achievement at the grade and test score level across all students. This is mostly due to the heterogeneity in school settings and the vast differences between studies with regards to study design, quality and chosen outcome measure and consequently a lack of generalisability of individual study results that also prevented conducting a meta-analysis (see Fig. 4 for suggested improvements). Importantly, as much as course grades and test scores do not *systematically* or greatly improve across the majority of studies, all included studies (except for one) showed no worsening after a SST delay. This means that SSTs could be delayed, while academic achievement is maintained at the same level (or improved in sub-groups or individuals) and possibly achieved with less cognitive effort or time spent on studying and homework (students are likely better rested and therefore cognitively more capable and efficient). In combination with other reported positive outcomes on sleep, daytime sleepiness, mood and motivation, computer gaming, attendance rates, or tardies and suspensions[e.g.14,15,24,25,33], this remains a valid argument in favour of delaying SSTs.

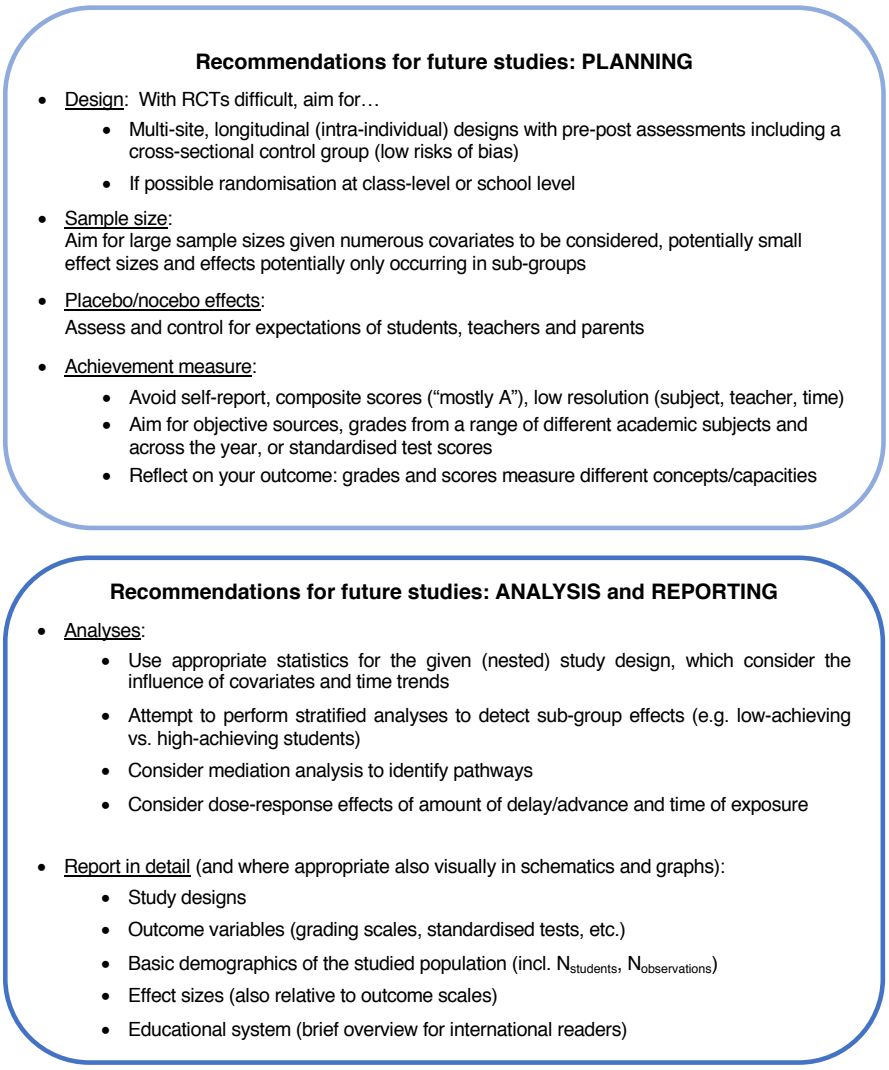

Fig. 4 | Suggestions for future studies.

Acknowledgements

We thank all contacted authors who replied and helped us to report their findings as accurately as possible. AMB thanks the Graduate School for Systemic Neurosciences, Munich, for financial support.

Author contributions (CRedIT Taxonomy)

Conceptualisation: AMB, GZ, ECW

Methodology: AMB, GZ, ECW, KM

Investigation: AMB, GZ

Data curation: AMB, GZ

Formal analysis: AMB, GZ

Validation: ECW, KM

Supervision: ECW

Visualization: AMB, GZ

Writing – original draft: AMB, GZ

Writing – review and editing: AMB, GZ, ECW, KM

Conflict of interest

AMB received research and travel funds from the Graduate School of Systemic Neurosciences Munich. GZ and KM report no funding in relation to the study and outside the submitted work. ECW reports receiving funds from the German Research Foundation (DFG) during the conduit of this study but outside of the submitted work.

References

1. Crowley, S. J. *et al.* A longitudinal assessment of sleep timing, circadian phase, and phase angle of entrainment across human adolescence. *PLoS One* **9**, (2014).

2. Crowley, S. J., Acebo, C. & Carskadon, M. A. Sleep, circadian rhythms, and delayed phase in adolescence. *Sleep Med.* **8**, 602–612 (2007).

3. Crowley, S. J., Wolfson, A. R., Tarokh, L. & Carskadon, M. A. An Update on Adolescent Sleep: New Evidence Informing the Perfect Storm Model. *J Adolesc.* 55–65 (2018). doi:10.1016/j.adolescence.2018.06.001

4. Carskadon, M. A., Acebo, C. & Jenni, O. G. Regulation of adolescent sleep: Implications for behavior. *Ann. N. Y. Acad. Sci.* **1021**, 276–291 (2004).

5. Gibson, E. S. *et al.* 'Sleepiness' is serious in adolescence: Two surveys of 3235 Canadian students. *BMC Public Health* **6**, 116 (2006).

6. Matricciani, L., Olds, T. & Petkov, J. In search of lost sleep: Secular trends in the sleep time of school-aged children and adolescents. *Sleep Medicine Reviews* (2012). doi:10.1016/j.smrv.2011.03.005

7. Keyes, K. M., Maslowsky, J., Hamilton, A. & Schulenberg, J. The Great Sleep Recession: Changes in Sleep Duration Among US Adolescents, 1991-2012. *Pediatrics* **135**, 460–468 (2015).

8. Gradisar, M., Gardner, G. & Dohnt, H. Recent worldwide sleep patterns and problems during adolescence: A review and meta-analysis of age, region, and sleep. *Sleep Medicine* (2011). doi:10.1016/j.sleep.2010.11.008

9. Raniti, M. B. *et al.* Sleep Duration and Sleep Quality: Associations With Depressive Symptoms Across Adolescence. *Behav. Sleep Med.* (2017). doi:10.1080/15402002.2015.1120198

10. Baum, K. T. *et al.* Sleep restriction worsens mood and emotion regulation in adolescents. *J. Child Psychol. Psychiatry Allied Discip.* **55**, (2014).

11. Short, M. A., Gradisar, M., Lack, L. C. & Wright, H. R. The impact of sleep on adolescent depressed mood, alertness and academic performance. *J. Adolesc.* **36**, 1025–1033 (2013).

12. Garaulet, M. *et al.* Short sleep duration is associated with increased obesity markers in European adolescents: Effect of physical activity and dietary habits. The HELENA study. *Int. J. Obes.* **35**, 1308–1317 (2011).

13. Mullington, J. M., Haack, M., Toth, M., Serrador, J. M. & Meier-Ewert, H. K. Cardiovascular, inflammatory, and metabolic consequences of sleep deprivation. *Prog. Cardiovasc. Dis.* **51**, 294–302 (2009).

14. Bowers, J. M. & Moyer, A. Effects of school start time on students' sleep duration, daytime sleepiness, and attendance: a meta-analysis. *Sleep Heal.* **3**, 423–431 (2017).

15. Marx, R. *et al.* Later school start times for supporting the education, health, and well-being of high school students. *Cochrane Database Syst. Rev.* **2017**, (2017).

16. Minges, K. E. & Redeker, N. S. Delayed school start times and adolescent sleep: A systematic review of the experimental evidence. *Sleep Med. Rev.* **28**, 82–91 (2016).

17. Beebe, D. W., Rose, D. & Amin, R. Attention, learning, and arousal of experimentally sleep-restricted adolescents in a simulated classroom. *J. Adolesc. Heal.* (2010). doi:10.1016/j.jadohealth.2010.03.005

18. Killgore, W. D. S. *et al.* Sleep deprivation reduces perceived emotional intelligence and

constructive thinking skills. *Sleep Med.* **9**, 517–526 (2008).

19. Hysing, M., Haugland, S., Bøe, T., Stormark, K. M. & Sivertsen, B. Sleep and school attendance in adolescence: Results from a large population-based study. *Scand. J. Public Health* (2015). doi:10.1177/1403494814556647

20. Walker, M. P. & Stickgold, R. Sleep, memory, and plasticity. *Annu. Rev. Psychol.* **57**, 139–166 (2006).

21. Stickgold, R. Sleep-dependent memory consolidation. *Nature* **437**, 1272–1278 (2005).

22. Maquet, P. The role of sleep in learning and memory. *Science (80-.).* **294**, 1048–1052 (2001).

23. Alhola, P. & Polo-Kantola, P. Sleep deprivation: Impact on cognitive performance. *Neuropsychiatric Disease and Treatment* **3**, 553–567 (2007).

24. Fuller, S. C. & Bastian, K. C. The Relationship Between School Start Times and Educational Outcomes. *Curr. Sleep Med. Reports* 18–19 (2020). doi:10.1007/s40675-020-00198-4

25. Wheaton, A. G., Chapman, D. P., Croft, J. B., Chief, B. & Branch, S. School start times, sleep, behavioral, health and academic outcomes: a review of literature. *J Sch Heal.* **86**, 363–381 (2017).

26. Hofer, M., Kuhnle, C., Kilian, B. & Fries, S. Cognitive ability and personality variables as predictors of school grades and test scores in adolescents. *Learning and Instruction* **22**, 368–4752 (2012).

27. Fehrmann, P. G., Keith, T. Z. & Reimers, T. M. Home influence on school learning: Direct and indirect effects of parental involvement on high school grades. *J. Educ. Res.* **80**, 330–671 (1987).

28. Keith, T. Z. & Benson, M. J. Effects of manipulable influences on high school grades across five ethnic groups. *J. Educ. Res.* **86**, 85–671 (1992).

29. Lekholm, A. K. & Cliffordson, C. Discrepancies between school grades and test scores at individual and school level: effects of gender and family background. *Educ. Res. Eval.* **14**, 181–3611 (2008).

30. Caprara, G. V., Barbaranelli, C., Steca, P. & Malone, P. S. Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *J. Sch. Psychol.* **44**, 473–490 (2006).

31. Alfonsi, V. *et al.* Later school start time: The impact of sleep on academic performance and health in the adolescent population. *Int. J. Environ. Res. Public Health* **17**, (2020).

32. Berger, A. T., Widome, R. & Troxel, W. M. *Delayed school start times and adolescent health*. *Sleep and Health* (Elsevier Inc., 2019). doi:10.1016/B978-0-12-815373-4.00033-2

33. Hershner, S. Sleep and academic performance: measuring the impact of sleep. *Curr. Opin. Behav. Sci.* **33**, 51–56 (2020).

34. Morgenthaler, T. I. *et al.* High school start times and the impact on high school students: What we know, and what we hope to learn. *J. Clin. Sleep Med.* **12**, 1681–1689 (2016).

35. Wahlstrom, K. L. & Owens, J. A. School start time effects on adolescent learning and academic performance, emotional health and behaviour. *Curr. Opin. Psychiatry* **30**, 485–490 (2017).

36. Wolfson, A. R. & Ziporyn, T. Adolescent sleep and later school start times. in *Sleep, Health, and Society: From Aetiology to Public Health* 215–223 (2018). doi:10.1093/oso/9780198778240.003.0024

37. Gomez Fonseca, A. & Genzel, L. Sleep and academic performance: considering amount, quality and timing. *Curr. Opin. Behav. Sci.* **33**, 65–71 (2020).

38. Wolfson, A. R. & Carskadon, M. A. Understanding adolescents' sleep patterns and school performance: a critical appraisal. *Sleep Med. Rev.* **7**, 491–506 (2003).

39. Schmidt, S. Later school starts linked to better teen grades. (2019). Available at: https://www.sciencenewsforstudents.org/article/later-school-starts-linked-better-teen-grades. (Accessed: 21st December 2020)

40. Urton, J. Teens get more sleep, show improved grades and attendance with later school start time, researchers find. (2018). Available at: https://www.washington.edu/news/2018/12/12/high-school-start-times-study/#:~:text=12 in the journal Science,minutes of sleep each night. (Accessed: 21st December 2020)

41. Lee, K. More Evidence Finds That Delaying School Start Times Improves Students' Performance, Attendance, and Sleep. (2018). Available at: https://www.everydayhealth.com/kids-health/delaying-school-start-times-improves-students-performance-health/. (Accessed: 21st December 2020)

42. Ackerman, X. *et al.* School Start Times. (2019). Available at: https://ccb.ucsd.edu/_files/bioclock/Infographic PDF, School Start Times 2019, Ackerman, Phan, Gee, Kim, Imani, Welkie, Golden.pdf. (Accessed: 21st December 2020)

43. French, M. T., Homer, J. F., Popovici, I. & Robins, P. K. What you do in high school matters: High School GPA, educational attainment, and labor market earnings as a young adult. *East. Econ. J.* **41**, 370–386 (2015).

44. Geiser, S. & Santelices, M. V. Validity of high-school grades in predicting student success beyond the freshman year: High school record vs. standardized tests as indicators of four-year college outcomes. *CSHE Res. Occas. Pap. Ser.* 35 (2007).

45. Ma, J., Pender, M. & Welch, M. Education Pays 2016. *Coll. Board Trends High. Educ. Ser.* 1–44 (2016).

46. Guyatt, G. H. *et al.* GRADE guidelines: 4. Rating the quality of evidence - Study limitations (risk of bias). *J. Clin. Epidemiol.* **64**, 407–415 (2011).

47. Sterne, J. A. *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* **355**, i4919 (2016).

48. Shin, J. Sleep More, Study Less ? The Impact of Delayed School Start Time on Sleep and Academic Performance. (2018). Available at: https://aefpweb.org/sites/default/files/JShin_Poster_AEFP2020.pdf.

49. Biller, A. M. *et al.* One year later: longitudinal effects of flexible school start times on teenage sleep, psychological benefits, and academic grades. (2021). Available at: https://syncandshare.lrz.de/public?folderID=MkFWcEdlcUtwU05QNGIxeHNENEU5.

50. Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. Preferred Reporting Items for Systematic Reviews and MetaAnalyses: The PRISMA Statement. *PLoS Med.* **6**, e1000097 (2009).

51. Boergers, J., Gable, C. J. & Owens, J. A. Later school start time is associated with improved sleep and daytime functioning in adolescents. *J. Dev. Behav. Pediatr.* **35**, 11–17 (2014).

52. Owens, J. A., Belon, K. & Moss, P. Impact of delaying school start time on adolescent sleep, mood, and behavior. *Arch Pediatr Adolesc Med* **164**, 608–614 (2010).

53. Milić, J. *et al.* Are there differences in students' school success, biorhythm, and daytime sleepiness depending on their school starting times? *Coll. Antropol.* **38**, 889–894 (2014).

54. Hinrichs, P. When the bell Tolls: The effects of school starting times on academic achievement.

Educ. Financ. Policy **6**, 486–507 (2011).

55. Edwards, F. Early to rise? The effect of daily start times on academic performance. *Econ. Educ. Rev.* **31**, 970–983 (2012).

56. Lenard, M., Morrill, M. S. & Westall, J. High school start times and student achievement: Looking beyond test scores. *Econ. Educ. Rev.* **76**, (2020).

57. Thacher, P. V & Onyper, S. V. Longitudinal Outcomes of start time delay on sleep, behavior, and achievement in high school. *Sleep* **39**, 271–281 (2016).

58. Groen, J. A. & Pabilonia, S. W. Snooze or lose: High school start times and academic achievement. *Econ. Educ. Rev.* **72**, 204–218 (2019).

59. Bastian, K. C. & Fuller, S. C. Answering the Bell: High School Start Times and Student Academic Outcomes. *AERA Open* **4**, 233285841881242 (2018).

60. Dunster, G. P. *et al.* Sleepmore in Seattle: Later school start times are associated with more sleep and better performance in high school students. *Sci Adv* **4**, eaau6200 (2018).

61. Wolfson, A. R., Spaulding, N. L., Dandrow, C. & Baroni, E. M. Middle school start times: The importance of a good night's sleep for young adolescents. *Behav. Sleep Med.* **5**, 194–209 (2007).

62. Wahlstrom, K. L., Frederickson, J. & Wrobel, G. *School Start Time Study : Technical Report, Volume II Analysis of Student Survey Data*. (1997).

63. Wahlstrom, K. L. *et al.* Examining the Impact of Later High School Start Times on the Health and Academic Performance of High School Students: A Multi-Site Study. (2014).

64. Wahlstrom, K. Changing Times: Findings From the First Longitudinal Study of Later High School Start Times. *NASSP Bull.* **86**, 3–21 (2002).

65. Jung, H. A late bird or a good bird? The effect of 9 o'clock attendance policy on student's achievement. *Asia Pacific Educ. Rev.* **19**, 511–529 (2018).

66. Rhie, S. & Chae, K. Y. Effects of school time on sleep duration and sleepiness in adolescents. *PLoS One* **13**, e0203318 (2018).

67. Kim, T. The Effects of School Start Time on Educational Outcomes: Evidence From the 9 O'Clock Attendance Policy in South Korea. *SSRN Electron. J.* 1–26 (2018). doi:10.2139/ssrn.3160037

68. Kelley, P., Lockley, S. W., Kelley, J. & Evans, M. D. R. R. Is 8:30 a.m. still too early to start school? A 10:00 a.m. school start time improves health and performance of students aged 13-16. *Front. Hum. Neurosci.* **11**, (2017).

69. Lewin, D. S. *et al.* Variable School Start Times and Middle School Student's Sleep Health and Academic Performance. *J. Adolesc. Heal.* **61**, 205–211 (2017).

70. Chattu, V. *et al.* The Global Problem of Insufficient Sleep and Its Serious Public Health Implications. *Healthcare* **7**, 1 (2018).

71. Carskadon, M. A. Factors influencing sleep patterns of adolescents. in *Adolescent Sleep Patterns: Biological, Social, and Psychological Influences.* (Cambridge University Press, 2002).

72. Wittmann, M., Dinich, J., Merrow, M. & Roenneberg, T. Social Jetlag: Misalignment of Biological and Social Time. *Chronobiol. Int.* **23**, 497–509 (2006).

73. Hafner, M., Stepanek, M. & Troxel, W. M. The economic implications of later school start times in the United States. *Sleep Heal.* **3**, 451–457 (2017).

74. Jacob, B. A. & Rockoff, J. E. Organizing schools to improve student achievement: Start times,

grade configurations, and teacher assignments. *Hamilt. Proj.* 24 (2011).

75. Illingworth, G. *et al.* Challenges in implementing and assessing outcomes of school start time change in the UK: experience of the Oxford Teensleep study. *Sleep Med.* **60**, 89–95 (2019).

76. Krueger, A. B. & Whitmore, D. M. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *Econ. J.* **111**, 1–28 (2001).

77. Card, D. & Rothstein, J. Racial segregation and the black–white test score gap. *J. Public Econ.* **91**, 2158–2184 (2007).

78. Hansen, M., Janssen, I., Schiff, A., Zee, P. C. & Dubocovich, M. L. The impact of school daily schedule on adolescent sleep. *Pediatrics* **115**, 1555–1561 (2005).

79. Fimm, B., Brand, T. & Spijkers, W. Time-of-day variation of visuo-spatial attention. *Br. J. Psychol.* **107**, 299–321 (2016).

80. Zerbini, G. *et al.* Lower school performance in late chronotypes: underlying factors and mechanisms. *Sci Rep* **7**, 4385 (2017).

81. Rimfeld, K., Kovas, Y., Dale, P. S. & Plomin, R. True grit and genetics: Predicting academic achievement from personality. *J. Pers. Soc. Psychol.* **111**, 780–789 (2016).

82. American Academy of Pediatrics. School Start Times for Adolescents. *Pediatrics* **134**, 642–649 (2014).

83. Zerbini, G. & Merrow, M. Time to learn: How chronotype impacts education. *Psych J* **6**, 263–276 (2017).

84. Neumann, K., Kauertz, A. & Fischer, H. E. Quality of instruction in science education. in *Second international handbook of science education* (eds. Fraser, B. J., Tobin, K. G. & McRobbie, C. J.) 247–258 (Berlin: Springer, 2012). doi:10.1007/978-1-4020-9041-7

85. Juang, L. P. & Silbereisen, R. K. The relationship between adolescent academic capability beliefs, parenting and school grades. *J. Adolesc.* **25**, 3–18 (2002).

86. Pokropek, A., Borgonovi, F. & Jakubowski, M. Socio-economic disparities in academic achievement: A comparative analysis of mechanisms and pathways. *Learn. Individ. Differ.* **42**, 10–18 (2015).

87. Rockoff, J. E. The impact of individual teachers on student achievement: Evidence from panel data. *Am. Econ. Rev.* **94**, 247–252 (2004).

88. Jarrin, D. C., McGrath, J. J. & Quon, E. C. Objective and subjective socioeconomic gradients exist for sleep in children and adolescents. *Health Psychol.* **33**, 301–305 (2014).

89. El-Sheikh, M., Kelly, R. J., Buckhalt, J. A. & Benjamin Hinnant, J. Children's sleep and adjustment over time: The role of socioeconomic context. *Child Dev.* **81**, 870–883 (2010).

90. Dahl, R. E., Allen, N. B., Wilbrecht, L. & Suleiman, A. B. Importance of investing in adolescence from a developmental science perspective. *Nature* **554**, 441–450 (2018).

91. Roenneberg, T. *et al.* A marker for the end of adolescence. *Curr. Biol.* **14**, 1038–1039 (2004).

Tab. 2 | Detailed descriptions of included studies. Studies are ordered by their type of study design (longitudinal with control group, without control group and cross-sectional). Study designs of two studies could not clearly be clearly defined. Abbreviations: SST, school start time; OLS, ordinary least square; SD, standard deviation; b, unstandardised beta coefficient; β, standardised beta coefficient; μ, average; CG, control group; IG, intervention group; CSAT, College Scholastic Ability Test; GPA, grade point average; ACT, American College Test; OR, odds ratio; NA, not available.

Author(s) (Year)	Study Design	Sample characteristics	Measure of school achievement	Type of analysis	Change in SSTs results in	Key findings
LONGITUDINAL (within-subjects comparison) with control group						
Edwards *et al.* (2012) [55]	Schools: 22-28 or 15-17 middle schools? 9 schools delayed, 4 advanced, 11 did not change SST Change: pre-change: 07:30 - 08:45 post-change: 7:30-8:25 Assessment: 1999-2006 Exposure time: NA Sampling resolution: once per year	$N_{students}$=20,530 or 10,544 + 6,082? (1999-2000) $N_{students}$=27,686 or 7,191 + 7,675? (2005-2006) $N_{observations}$: up to 102,506 Grade levels: 6^{th} – 8^{th} Age: 11-14.5 Gender ratio: ~ 51% males Ethnicity/race: Caucasian, Black, Hispanic Location: Wake County, North Carolina, USA	Outcome: End of year standardised test scores in reading and math Scale: 0%-100% (inferred); converted to percentile scores for each student within their grade and current year Provided by: Wake Country administration	Pooled OLS models, quantile regression model predicting scores Covariates: several on student-level and school-level Fixed-effects: student and school	Per 1h delay in SST: 1.8-2.9 percent points (0.06-0.07 SD) increase in maths and 1.0-3.4 percent points (0.04-0.05 SD) increase in reading when using within schools variation or both within and between school variation (both p<0.01) Some covariate results for maths and school fixed effect: Black colour: β= -0.50 (p<0.01) Hispanic: β= -0.17 (p<0.01) Female: β= -0.054 (p<0.01) Free lunch status: β= -0.17 (p<0.01) Parent education (years): β =0.08 (p<0.01)	Up to 0.07 SD gains in maths and 0.05 SD increase in reading end-of-year standardised scores even when adjusted for covariates The effect was stronger for lower achieving students
Jung (2018) [65]	Schools: drawn sample from 85 elementary and 63 middle schools; Cohorts from the Gyeonggi Education Panel Study SST Change: pre-change: 8:00-8:20 post-change: 9:00 Assessment: 2012-2017; SST delay in 2014; i.e. data cover 3 years prior and 2 years after the change Exposure time: 2 years Sampling resolution: once per year	**Group 1: longitudinal cohort** with change in SST (IG) compared to 220 students (CG) who did not delay: N_{total}=2,562 Grade levels: 4^{th} – 9^{th} Age: 11 – 16 Gender ratio: 50.5% (IG) -57.7% (CG) male Ethnicity/race: NA **Group 2: cross-sectional cohorts** N_{IG}=4,026 (2015) N_{CG}=2,562 (2012) Grade levels: 7^{th} Age: 14 Gender ratio: ~51 % male Ethnicity/race: NA Location: Gyeonggi, South Korea	Outcome: Korean, English and math grades at the end of the spring semester Scale: NA Provided by: governmental agency	Difference-in-difference estimation / OLS estimation Covariates: various student-level and school-level characteristics, Fixed effects: year, individual Cross-sectional comparison as robustness check	At first sight, math and English test scores increased (also Korean but p>0.05); when controlling for personal covariates the effect becomes smaller and n.s. for math; when applying individual-fixed effect estimation the result becomes negative (significantly for Korean) Model specification 4 with Year fixed effects and all personal covariates: Korean: β= 0.048 (p>0.05) Math: β= 0.16 (p>0.05) English: β= 0.18 (p<0.01) Cross-sectional robustness check confirms longitudinal results	No effect on Korean, English or math test scores when controlling for personal covariates or unobserved individual heterogeneities Caveat: Sleep duration did not differ between CG and IG!

Kim (2018) [67]	Schools: high schools from 2 districts; Gyeonggi delayed (IG) and Seoul did not delay (CG) SST Change: pre-change in IG: varying between earlier than 7:40 and 9:00 post-change in IG: 9:00 SST in CG: varying between earlier than 8:00 and 9:00 Assessment: 2009 to 2016; policy change to 9:00 starts in Sept,2014 in Gyeonggi; i.e. data cover 5 years prior and 2 years after the change Exposure time: 2 years Sampling resolution: once per year	Group A: Schools in IG Group B: Schools in CG $N_{observations}$= up to 1,479,131 Grade levels: 9th-12th Age: 15-18 Gender ratio: 52% males Ethnicity/race: NA Location: Gyeonggi and Seoul, South Korea	Outcomes: **(1**) Annual National Assessment of Educational Achievement (Korean, math, English) for 9th and 11th graders **(2)** College Scholastic Ability Test (CSAT) for 12th graders Scale: NA Provided by: EduDataService System	Difference-in-differences method Covariates: regional time trends Fixed effects: individual and school Several robustness checks	Results 11th graders: Math scores especially in male students increased by 0.06-0.1 SD ($p<0.01$). Results are robust when adding covariates to the model. Korean and English scores become non-significant when control variables are added to the model. For 12th graders: For CSAT no statistically significant benefit from the 9 o'clock-policy	Small effects on math, but not on Korean or English standardised scores No effect on CSAT scores (possible time-of-day interference: the CSAT is scheduled before 9 am)
Rhie & Chae (2018) [66]	Schools: several middle (MS) and high schools (HS); Gyeonggi district delayed SST (IG) 3 other districts did not delay (CG) SST Change: pre-change (IG): 7:30 – 8:10 post-change (IG): 9:00 (MS delayed by 30-60min; HS delayed by 60-90min) SST in CG: 7:30 – 8:00 Assessment: 2 years of data before and after the change (2012−2016); 2014 as the year of change was excluded Sampling resolution: once per year Exposure time: 2 years	N_{IG}=42,517 N_{CG}=28,287 Grade levels: 7th – 11th or 12th Age: NA Gender ratio: ~52% male Ethnicity/race: NA Location IG: Gyeonggi district and Daegu/Gyeongbuk/Ulsan district, South Korea	Outcome: Self-reported GPAs Scale: Percentage of students having" high and moderate GPAs" Provided by: participants	Logistic regression analysis using complex samples Covariates: NA	Percentage of students reporting "high and mid high GPAs": Years 2012, 2013, 2015, 2016: IG= 34.3%, 33.9%, 38.4%*, 37.8%* CG= 39.8%*, 36.7%, 40.6%*, 39.4%* *: different from 2013 on $p<0.05$	No 9AM policy effect on self-reported GPAs

Shin (2018) [48]	Schools: all middle schools in 2 districts (599 schools in Gyeonggi and 383 schools in Seoul) SST Change: pre-change in Gyeonggi (IG): ~ 8:20 post-change in Gyeonggi (IG): 9:00 SST in Seoul (CG): varying between before 8:00 and 9:00 Assessment: 2013-2015 with the policy change to 9:00 starts in Sept 1,2014 in Gyeonggi Exposure time: 1 year Sampling resolution: once/semester	$N_{observations}$= up to 33,282 Grade levels: 7th -9th Age: NA Gender ratio: ~50% male (direct contact with author) Ethnicity/race: mostly Asian (from private correspondence with author) Location: Gyeonggi and Seoul, South Korea	Outcome: Semester grades (standardised) for math and reading Scale: numeric; 0-100; normalized by population distribution Provided by: Korean Education & Research Information Service	Difference-in-difference methods Covariates: various individual and school-level variables Fixed effects: year, month	Increase in math (0.03 SD) and reading grades (0.02 SD; (both p<0.001)	Up to 0.03 SD increase in math and 0.02 SD increase in reading semester grades when adjusted for time-trends and other covariates
Lenard *et al.* (2020) [56]	Schools: 19 high schools, of which 5 schools **advanced** SST (IG) and 14 did not (CG) SST Change: pre-change (IG): 8:05 post-change (IG): 7:25 SST in CG: 7:25 Assessment: data span 2008-2019 with SST advance in 2012-2013 Exposure time: 7 years Sampling resolution: once per year	$N_{students}$~10,000 per each 8 cohorts $N_{observations}$= up to 52,854 (ACT scores) Grade levels: 8th – 12th (inferred) Age: NA Gender ratio: ~ 50% males Ethnicity/race: White, African American, Hispanic, other Location: Wake County, North Carolina, USA	Outcome: ACT scores in 11th grade (composite and individual scores for English, reading, math and science) Scale: 1-36 (= best) Provided by: Wake County administration	Difference-in-difference estimation approach Comparative interrupted time series Covariates: various individual and school-level variables	No effect of earlier start times on ACT composite or individual subject ACT scores (independent of length of exposure) ACT composite scores: Partial exposure: β= 0.023, p>0.05 Early start all years: β= -0.167, p>0.05 Treated schools all: β= 0.273, p>0.05 Scores were trending in all schools with math scores dropping over subsequent cohort groups, while English, reading and science were rising	No effect on individual or composite ACT scores when start times are **advanced**
Author(s) (Year)	**Study Design**	**Sample characteristics**	**Measure of school achievement**	**Type of analysis**	**Change in SSTs results in**	**Key findings**
LONGITUDINAL (within-subjects comparison) without control group						
Biller *et al.*[49]	School: 1 high school SST Change: pre-change: mostly 8:00 post-change: 8:00 or 8:50 (flexible choice on a daily basis) Assessment: data span 4 years (2.5	$N_{students}$= 63-157 $N_{observations}$= up to 16,724 Grade levels: 7th – 12th Age: 14-21 Gender ratio: 30-40% males Ethnicity/race: NA Location: Alsdorf, Aachen region, Germany	Outcome: quarterly grades of 12 academic subjects of 3 disciplines (sciences, social sciences, languages) Scale: numeric, 0%-100% (= best) Provided by: school registry	Linear mixed models predicting quarterly grades Covariates (in all models): gender, grade level, academic discipline, academic quarter	No effect of flexible system on grades: β = 0.00 (p>0.05) No absolute sleep effects on grades No effects of changes in sleep duration or chronotype from baseline to the flexible system on grades except for social jetlag (post β=0.03, p=0.027)	No effect of the flexible system nor sleep duration or chronotype on quarterly objective grades when controlled for covariates

	before and 1.5 years after the change) Exposure time: 0.5-1.5 years Sampling resolution: 4 times/year			Predictors (in some models): chronotype (+change from t_0-t_1), social jetlag (+change from t_0-t_1), sleep duration (+change from t_0-t_1), amount of 8:50AM-use	Covariates (numbers from models 3a-d): Male: β=0.07(p>0.05) Grade level 12: β=0.06 (p<0.001) Quarter 4: β=0.05 (p<0.001) Social Sciences: β=0.17 (p<0.001)	
Boergers *et al.* (2014) [51]	School: 1 independent high school (boarding school) SST Change: pre-change (t_1): 08:00 post-change (t_2): 08:25 back-change (t_3): 08:00 Assessment: Nov 2010 (t_1), Mar 2011 (t_2), and May 2011 (t_3) Exposure time: ~5 months (during winter term) Sampling resolution: once/ time point	$N_{students}$=197 Grade levels: 9th – 12th Age: μ = 15.6 Gender ratio: 41% males Ethnicity/race: White, Black, Hispanic Asian, multiracial or other Location: Rhode Island, USA	Outcome: Self-reported grades Scale: categorical; "mostly Bs or better" Provided by: students	No type of analyses stated Covariates: NA	After the delay in SST, the percentage of self-reported "mostly Bs or better" changed from 93% (t_1) to 91 % (t_2)	Unclear as statistics are not reported; authors report no effect
Owens *et al.* (2010) [52]	School: 1 independent high school (boarding and day school) SST Change: pre-change: 8:00 post-change: 8:30 Assessment: Dec 2008, Mar 2009 Exposure time: 2 months (January-March 2009) Sampling resolution: once/time point	$N_{students}$=201 Grade levels: 9th – 12th Age: μ~16.5 Gender ratio: ~ 43% males Ethnicity/race: NA Location: Rhode Island, USA	Outcome: Self-reported grades Scale: categorical; "mostly B's or better" Provided by: participants	χ^2 analysis Covariates: NA	After the delay in SST, the percentage of self-reported "mostly B's or better" changed from 82.2% to 87.1% OR=0.70; 95% CI=0.41-1.20, X^2=1.71, p=0.22	No effect on self-reported grades
Thacher & Onyper (2016) [57]	School: 1 public high school SST Change: pre-change: 7:45 post-change: 8:30 Assessment: data span 2010-2014; 2 years of data before and after the change in 2012 Exposure time: 1-2 years	$N_{students}$ ~ 650 – 800 across 4 years (but t-test for cross-sectional comparisons seems to be ~250-330) Grade levels: 9th – 12th Age: μ~16.5 Gender ratio: NA Ethnicity/race: NA Location: Glen Falls, NY, USA	Outcomes: **(1)** Weighted average GPAs and subject-specific GPAs (English, science, math, social studies, art, music, foreign language, and health studies) **(2)** Standardised test scores from Regents Exams for cross-sectional comparison	**Longitudinal comparisons:** mixed effect analyses including within-subject effect control	**Longitudinal comparison:** no statistically significant evidence for improvement/decline in GPA (overall and by subject) Effect of grade level (higher grade levels better grade), gender (males worse; p=0.011-0.059), free lunch status (p<0.001) independent of SST No exact numbers are reported	No (systematic) effect on overall GPAs in longitudinal comparison

	Sampling resolution: once/year		Scale: numeric; 100 point scale for GPAs Scale for Regents Exam: NA Provided by: school	**Cross-sectional comparisons:** independent-samples t-tests for grades and standardised test scores	**Cross-sectional by grade level:** (1) GPA: Only 11th graders' GPAs increased by 2.55 percent points after the change: Mean before: 78.79% (SD 11.11) Mean after: 81.34& (SD 8.79) t_{295}=2.20, p=0.028 (2) Regents exams: 2 of 20 subject test scores (10th grade Earth Sciences and 11th grade Algebra) were significant better **before** the change (p<0.007)	GPA test scores of 11th graders improved cross-sectionally with later SSTs No systematic effects for individual subjects of standardised test scores cross-sectionally
Wahl-strom (2002) [64]	Schools: 7 comprehensive high schools SST Change: pre-change: 7:15 post-change: 8:40 Assessment: 6 years; data cover 3 years before and after the change in the 1997-1998 school year Exposure time: 3 years Sampling resolution: NA	$N_{students}$= NA $N_{observations}$= >1 million Grades levels: 9th – 12th? Age: NA Gender ratio: NA Ethnicity/race: NA Location: Minneapolis, Minnesota, USA	Outcome: all letter grades (semester and trimester grades) Scale: categorical letter grading Provided by: school district administration	Statistical analysis not reported Covariates: NA	"A small improvement in grades earned overall but not statistically significant" No actual numbers are reported	No effect on letter grades
Author(s) (Year)	**Study Design**	**Sample characteristics**	**Measure of school achievement**	**Type of analysis**	**Change in SSTs results in**	**Key findings**
CROSS-SECTIONAL (between-subjects comparison)						
Groen & Pabilonia (2019) [58]	Schools: 790 high schools Data from the Child Development Supplement to the Panel Study of Income Dynamics (PSID-CDS) and the Common Core of Data SST Ranges: from 07:00 to 09:15 (average start time of 7:53) Assessment: data from years 2002/03 and 2007/08 Exposure time: NA Sampling resolution: once/year	$N_{students}$= up to 1200 Grade levels: 9th – 12th Age: 13-18 years Gender ratio: 50% males Ethnicity/race: White, Black, Asian, Hispanic Location: USA	Outcome: broad-reading test score and applied-problems (math) test score; both age-adjusted and from the Woodcock-Johnson Revised Tests of Basic Achievement Scale: NA; normalised by survey year Provided by: NA (probably by research assistant)	Linear OLS model predicting test scores; Oster model (bounded effects); instrumental-variable estimates Predictors: SST Covariates: several on student-level and school-level, e.g. school day length	Per 1h delay: increase in females' reading scores by 0.16 SD (p<0.1); no sign. effect for females' math scores and for both males' applied problems and reading scores From Oster model (bounded effects): 0.16-0.28 increase in reading for females; 0.05-0.12 increase in applied-problems for males → probably mediated by an increase of 36 min in sleep duration for every 1 h of later SSTs for females but not for males	Woodcock-Johnsons Test scores increased by up to 0.28 SD in reading for females, no significant effect for males' scores No significant effect on applied-problems scores for either females or males

Hinrichs 2011[54]	Schools: 48 districts (73 schools) Minneapolis and some suburbs delayed (IG); St. Paul and other suburbs maintained schedules (CG) SST Change: pre-change (IG): 7:15 post-change (CG): 8:40 SST in CG: 7:30 Assessment: data span 1993-2002 with change in 1997/1998 Exposure time: ~ 5 years Sampling resolution: once/year	$N_{observations}$=196,617 Number of students not exactly known, but slightly less than the number of observations according to author (private correspondence) Grade levels: 10th-12th Age: NA Gender ratio: 44% males Ethnicity/race: White, Black, Asian, Hispanic, Other Location: Twin Cities metropolitan area, Minneapolis, USA	Outcome: individual composite ACT test scores Scale: numeric; 0-36 (= best) Provided by: ACT test company with permission from schools	OLS regression predicting ACT scores Predictors: SST, school day length Covariates: several on student-level and school/district-level	No effect of SSTs on ACT scores (from full specification): 1h later SSTs: 0.02 SD, p>0.05 Covariates: Males: b=0.25 SD, p<0.01 Black: b=-2.47 SD, p<0.01 Low income: b=-0.92 SD, p<0.01	No effect on ACT scores
	Schools: every public high school in Kansas state SST Range: NA Assessment: 2000-2006 (11th grade reading and 10th grade maths between 2001-2006; 11th grade social science and 10th grade science between 2000-2006) Exposure time: NA Sampling resolution: once/ year	$N_{schools}$=1,666 Grades: 10th-11th Age: NA Gender ratio: 40% white females, 9% non-white females, 9% non-white males Ethnicity/race: white and non-white Location: Kansas, USA	Outcome: school-level test score data on state-wide Kansas Assessments in math, reading, science and social studies Scale: 0-100% (inferred, not stated) Provided by: Kansas Department of Education	OLS regression predicting Kansas assessment test scores Covariates: several on school-level variables	No effect on any of the test scores (maths, reading, science, or social studies) For reading: 1h later SSTs (from full specification): b=0.95, p>0.05	No effect on Kansas Assessment scores in math, reading, science or social studies
	Schools: 75 schools in 19 districts in Virginia SST Change: NA Assessment: data span 2000-2007; some delays in 2001/2002 Exposure time: ~ 6 years Sampling resolution: once/year	$N_{observations}$=171 (number of district-by-year pairs) Grade level/Age/Gender ratio/Ethnicity/race: NA Location: Virginia suburbs of Washington, DC, USA	Outcome: End of course exams Scale: 0-100% Provided by: Virginia Department of Education	Analysis: OLS regression predicting end of course exams Covariates: several on school-level	"The results, which are not reported here but are available upon request, are somewhat imprecise, but they do not give evidence for an effect of the timing of the school day on test scores." → requested and confirmed	No effect on end of course exams
Bastian & Fuller (2018)[59]	Schools: 410 high schools; 1,591 schools by year (includes all public school students in North Carolina) SST Range: 7:00 to 9:30 23 schools changed start times, 9 changed start times by ≥30 min	$N_{students}$=770,623 Grade level: 9th-12th (inferred) Age: 14-18 years (inferred) Gender ratio: NA	Outcomes: **(1)** Average course grades and course grades in 1st period classes in math, English, science and social studies Scale: 4-point scale	Linear regression models Covariates: several on student-level and school-level; incl. year-fixed effects	(1) Course grades: Per 1h delay: No effect of SSTs on overall course grades β =0.012, p>0.05 SSTs effect on course grades in 1st period: For ≥8:30h starts (compared to <7:30h start): β =0.050, p<0.05	No significant relationship between SSTs and average course grades Grades in 1st period class were improved

	44 districts (278 schools) had across-school variation in SSTs; average time difference between earliest and latest was 33min (5min to max 2h); remaining 69 districts (132 schools) had no variation Assessment: data span school years 2011-2012 through 2014-2015 Exposure time: NA Sampling resolution: NA; probably once/year	Ethnicity/race: White, Black, Hispanic, American Indian, Asian, multiracial Location: North Carolina, USA	Conversation of numeric course grades into unweighted grade points **(2)** Test scores from state wise standardised end-of-course exams (EOC) in algebra, biology, English Scale: normalised **(3)** ACT composite scores Scale: 0-36 Provided by: North Carolina Department of Public Instruction	Specification checks with school fixed effect and with school district fixed effects (=robustness check of main results)	Economically disadvantaged students, minority and low-performing students benefited more in course grades overall and in 1st period (per 1h later): β= from 0.049 to 0.074, p>0.05 or 0.01 (2) EOC scores**:** Mixed results for EOC Algebra (improvements but p>0.05.), EOC Biology (reductions, p<0.05), and EOC English (reductions but p>0.05.) (3) ACT scores: Per 1h delay: overall SSTs effect on ACT composite: β=0.107, p>0.05; low-performing students: β=0.277, p<0.05	by 0.05 quality points associated with a later start (≥8:30h) Later SSTs did not systematically predict EOC or ACT but low-performing students did better on the ACT Later SSTs were associated with better course grades (overall and 1st period) for disadvantaged students
Dunster *et al.* (2018)[60]	Schools: 2 public high schools (RHS and FHS) SST Change: pre-change: 07:50 post-change: 08:45 Assessment: spring 2016 (pre) and spring 2017 (post) Exposure time: NA; ~ 7 months? Sampling resolution: once/year (second semester)	$N_{students}$=178 (pooled from both schools during each year) 2 independent samples: Sample 2016: n=51(RHS) + n=41(FHS) Sample 2017: n=41(RHS) + n=41(FHS) Grade level: 10th Age: μ~16 Gender ratio: ~47% male Ethnicity/race: White, Black, Asian, African American, unknown/other Location: Seattle, USA	Outcome: One 2nd semester grade from a science lab class Scale: NA (probably 0%-100%) Provided by: teacher	According to method section: Generalized linear models (binomial) "with year as the dependent variable, testing the hypothesis that years differed based on the basis of the other variables" Other variables: school, sleep offset, grade, mood, chronotype, sleepiness	*Years* differed after controlling for differences, including grade and sleep variables Median grade 2016: 77.5% Mean grade 2016: 74.6% Median grade 2017: 82% Mean grade 2017: 76.6%	Higher biology grades were predictive of years
Milić *et al.* (2014) [53]	Schools: 4 schools (2 grammar schools and 2 vocational schools), all with weekly alternating morning and afternoon schedules SST Schedules: 2 schools with early schedule: 07:00 (morning schedule) and 13:00 (afternoon schedule) 2 schools with late schedule: 08:00 (morning schedule) and 14:00 (afternoon schedule) Assessment: May and June 2011	$N_{students}$= 821 Sample Early schedule: n=452 Sample Late schedule: n=369 Grade levels: NA Age: 15-19 years Gender ratio: across entire sample: 54% males sample early schedule: 73% males sample late schedule: 30 % males Ethnicity/race: NA Location: Osijek, Croatia	Outcome: Final grade in last semester Scale: numeric; 1-5 (= best) Provided by: NA	Mann-Whitney Test Covariates: no	Students attending the **early schedule** obtained better grades (p<0.001) SST at 07:00: Mean grade: 3.60 (SD 1.08)= 72.0% SST at 8:00: Mean grade: 3.28 (SD 1.19)= 65.6%	Final semester grades were better in **earlier** starting schools

	Exposure time: NA; probably since admission to school Sampling resolution: once					
Wolfson *et al.* (2007) [61]	Schools: 2 middle schools SST Differences: School E: 7:15 School L: 8:37 Assessment: fall 2003, spring 2004 Exposure time: NA; probably since admission to school (2-3 years?) Sampling resolution: once	$N_{students}$= 205 School E: n=79 School L: n=126 Grade levels: 7th (n=99) – 8th (n=106) Age: NA Gender ratio: 40% males Ethnicity/race: White, African American, Hispanic, other Location: New England, USA	Outcome: average fall quarter grade based on mean of 4 subjects (English, science, math and social studies) Scale: numeric; 0-100% (= best) Provided by: schools	MANOVA, Bonferroni correction for group comparisons Variables: school, grade, gender Other covariates: no, but "schools were similar in SES, size, and ethnic distribution of students" (except for a higher percentage of Whites in School L (60% vs. 46%)	Significant School x Grade interaction: $F(1,208)=17.06$, $p<0.001$; i.e. there were no school differences for 7th graders but 8th graders Students at School L had higher average grades than students at School E: $F(1,104)=10.60$, $p<0.01$; SST at 7:15: Mean grade: 83.16% (SD 7.16) 7th graders Mean grade: 76.85% (SD 9.45) 8th graders SST at 8:37: Mean grade: 80.46% (SD 10.11) 7th graders Mean grade: 83.79% (SD 8.80) 8th graders No gender differences were found	Increased averaged fall quarter grades for 8th graders in later school No significant difference for 7th graders
Lewin *et al.* (2017) [69]	Schools: 26 middle schools with variable SSTs (country wide surveillance data) SST Ranges: "Earliest" SSTs: 7:20 – 7:30 "Early" SSTs: 7:40 – 7:55 "Late" SSTs: 8:00 – 8:10 Assessment: surveys in 2008, 2010, and 2012 Exposure time: NA; probably since admission to school (3 years?) Sampling resolution: once/time point	$N_{students}$ ~ 32,000 Pooled from all sample years Sample 2008: n=6,936 Sample 2010: n=11,991 Sample 2012: n=10,768 Sample "Earliest" SSTs: n=7,206 Sample "Early" SSTs: n=13,161 Sample "Late" SSTs: n=12,613 Grade levels: 8th Age: 13-14 years Gender ratio: 49.8% males Ethnicity/race: White, non-white Location: NA but most likely USA	Outcome: Self-reported grades Scale: 4-point categorical; "Do you mainly get A's, B's, C's, or D's/F's?" Provided by: participant	Path analysis with probit regression predicting grades: Predictor: SSTs Mediator: sleep duration (Sobel test) Covariates student-level: survey year, gender, race Covariates school-level: free lunch status Hierarchical structure: students nested within schools	Self-reported grades of students attending the "earliest schools" were significantly lower ($\beta=-0.286$, $p=0.012$), no sign. effect for "earlier schools" ($\beta=-0.114$, $p=0.126$) The negative effect of SST on grades was overall mediated by sleep duration: $\beta=0.115$, $p<0.001$ Covariates: Female: $\beta=0.312$, $p<0.001$ Non-white: $\beta=-0.321$, $p<0.001$ Free lunch status: up to $\beta=-0.668$, $p<0.001$	An advance of at least 30 min was associated with worse self-reported grades Longer sleep duration was also associated with increased grades
Wahl-strom 1997 [62]	Schools: 3 districts; 1 delayed SSTs, 2 did not delay; high schools and middle schools SST Ranges: High schools (10-12th grades): SST at district A (IG): 8:30 SST at district B: 7:25 SST at district C: 7:15	$N_{students}$= a not further defined sample was drawn from 7,168 students of 17 districts Grade levels: 10th – 12th and 7th – 8th Age: NA Gender ratio: NA Ethnicity/race: NA Location: Minnesota, USA	Outcome: Self-reported grades Scale: NA Provided by: participants	Statistical analysis: NA Covariates: NA	Mean self-reported grades in district A were highest ($p<0.05$) compared to district B and C for 10-12th graders: District A: 7.08 District B: 6.50 District C: 6.37 But not for 7-8th graders: District A: 6.66	Students in a district with later start time reported getting higher grades than in two districts with earlier SSTs (high schools)

	Middle schools (7-8th grade): SST at district A (IG): 7:35 SST at district B: 8:00 SST at district C: 8:00 Assessment: NA Exposure time: NA Sampling resolution: NA				District B: 6.91 District C: 6.60	For middle schools, students who started later were either better or similar to students from a school which started earlier
Kelley *et al.* (2017) [68]	School: 1 English state-funded high school SST Change: Year 0 = pre-change (A): 08:50 Year 1-2 = post-change (B): 10:00 Year 3 = change back (A): 08:50 Assessment: 4 years Exposure time: 1-2 years Sampling resolution: once/year	Year 0: $n_{students}$=169 Year 1: $n_{tudents}$=166 Year 2: $n_{tudents}$=164 Year 3: $n_{tudents}$=179 Grade levels: NA Age: 14-16 Gender ratio: NA Ethnicity/race: NA Location: urban-area of 0.7 million in a region where achievement was lower than national average, England	Outcome: Standard National Examination (GCSE) Scale: G-A* (= best) Provided by: UK Office of National Statistics	T-test; Cohen's d and h for effect size Value-added analysis (predictions) Percentage of students achieving "good academic progress" (= achieving 5 or more GCSE grades of C or better in English, math and at least 3 other subjects) Covariates: NA	Change in value-added as % of national (compared to national average): Year 1 vs 0: +15%, p<0.0005 Year 2 vs 0: +20%, p<0.0005, Year 3 vs 2: -7%, p<0.0005 Percentage of students making good academic progress compared to national average: Year 0: -40%, p<0.005 Year 1: -9%, p=0.182 Year 2: -11%, p=0.081 Year 3: -15%, p=0.014	Delay is associated with higher % of students making good academic progress and higher value-added number compared to national average
Wahl-strom (2014) [63]	Schools: 8 public high schools in 5 school districts in 3 states changed SSTs and participated in a survey on sleep habits Grades were retrieved from 6 schools in 3 districts SST Change: pre-change: 7:35-7:50 post-change: 8:00-8:55 Assessment: 2010-2011 (Minnesota), 2011-2012 (Colorado); 2011-2012 (pre-change in Wyoming) vs 2012-2013 (post change in Wyoming) Exposure time: 1 year? Sampling resolution: once/time point	$N_{students}$= 9,089 (sleep habits survey) $N_{students}$: NA (grade analyses) Grade levels: 9th-12th Age: 13-19 Gender ratio: 50.6% Ethnicity/race: White, Black/African American, Hispanic/Latino, Asian/Asian American, Other Location: Minnesota/Colorado/Wyoming, USA	Outcomes: **(1)** Grades in English, maths, social studies, science in 1st and 3rd period-classes or GPAs Scale: categorical; "mostly A's=9" to "mostly F's"=1 **(2)** Standardised test scores (state-wide achievement tests or PLAN) Scale: NA Provided by: GPAs from districts; categorical grades by students	t-tests, correlations Covariates: NA	Longitudinal standardised test scores: mainly non-significant results and some mixed results for both composite scores (PLAN) and individual subjects	Mixed and often non-significant effects on GPAs and standardised test scores

Supplementary information for

"School start times and academic achievement - a systematic review on grades and test scores"

Authors:

Anna M. Biller, Karin Meissner, Eva C. Winnebeck & Giulia Zerbini

Corresponding authors:

anna.biller@med.uni-muenchen.de

giulia.zerbini@med.uni-augsburg.de

Tab. S1 | Protocol detailing reasons for risk of bias assessment decisions of Tab. 1.

LONGITUDINAL STUDIES (within-subjects comparison) WITH CONTROL GROUP			
Study	**Bias**	**Decision**	**Reasons**
Edwards 2012[55]	Reporting bias on author level	Green	Sample characteristics are well described; actual sample size for the different analyses is not always clear. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Orange	State-wide data are used to construct percentile scores for each student within their grade and current year; no direct comparison with the national average
Jung 2018[65]	Reporting bias on author level	Green	Sample characteristics are well described. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Green	Yes
Kim 2018[67]	Reporting bias on author level	Orange	Sample characteristics are not well described. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported.
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs. The EDSS provided a 70% randomly extracted sample from the population (includes 3 exam scores, gender and school ID)
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Green	Yes
Rhie 2018[66]	Reporting bias on author level	Orange	Sample characteristics are well described but some information is contradictory (cfr Table 1 and Table S1 grade 7th to 11th or 12th). Statistical analyses are not fully reported. P values are reported but statistical tests are not reported. Overall there is no effect of the 9AM policy, but the third sentence of the discussion says: "Self-reported school performance of the intervention group was more improved than the control". This analysis is not reported and it is contradictory with the rest of the results
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	(Dis)similarity of baseline characteristics	Green	Considered (in terms of sleep onset, sleep offset and sleep duration; not for gender)
	Appropriate statistical models which control for confounders	Orange	No. The authors used a logistic regression with complex samples comparing each year to a baseline year, and for intervention and control group but no covariates were included
	Control group present	Green	Yes
Lenard 2020[56]	Reporting bias on author level	Green	Sample characteristics are well described. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported

	Responder bias on student level	Green	Data objectively reported (ACT scores)
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Green	Yes
Shin 2018[48]	Reporting bias on author level	Orange	Sample characteristics are not well described. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported.
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Green	Yes
LONGITUDINAL STUDIES WITHOUT CONTROL GROUP (within-subjects comparison)			
Study	**Bias**	**Decision**	**Reasons**
Biller 2021[49]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Green	Data objectively reported
	Blinding	Red	Even though grades were obtained objectively from the school registry, students were aware they were participating in the experiment (they had to fill in questionnaires)
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Red	No
Boergers 2014[51]	Reporting bias on author level	Red	Statistical analyses regarding grades are not reported
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	Appropriate statistical models which control for confounders	Red	Unable to judge due to missing information
	Control group present	Red	No
Owens 2010[52]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses). Ethnicity is missing
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	Appropriate statistical models which control for confounders	Red	Simple Chi-Square Test not controlling for confounders
	Control group present	Red	No
Thacher 2016[57]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Green	Data objectively reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	Appropriate statistical models which control for confounders	Orange	The longitudinal analysis is appropriate (mixed linear model with some moderators/covariates); cross-sectional analyses are only simple t-tests; it is not clear why the authors do not run a mixed within-between model and combine longitudinal and cross-sectional analyses. Nevertheless, several analyses are reported which supports the notion that the data were extensively explored to reach the conclusions of the paper
	Control group present	Red	No

Wahl-strom 2002[64]	Reporting bias on author level	Red	Sample size (n students) for the grade analyses not reported, statistical analyses not reported, demographic characteristics of the sample not fully reported; author(s) were contacted but did not respond
	Responder bias on student level	Green	Data objectively reported
	Blinding	Orange	Students were aware they were participating in the experiment (they had to fill in questionnaires) but grades were collected also in students who did not fill in questionnaires
	Appropriate statistical models which control for confounders	Red	Incomplete information to judge
	Control group present	Red	No
CROSS-SECTIONAL STUDIES (between subject comparison)			
Study	**Bias**	**Decision**	**Reasons**
Groen 2019[58]	Reporting bias on author level	Green	Statistical analyses overall are well described. Only analyses regarding results in Table 1 are not reported; not always specified that the statistically significant results were at the 0.1 level
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
Hinrichs 2011[54]	Reporting bias on author level	Green	Originally orange but author provided all details after contacting
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
Bastian 2018[59]	Reporting bias on author level	Green	Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported.
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
Dunster 2018[60]	Reporting bias on author level	Green	Information about sample and study design are well described. Statistical analyses are well described but are contradictory reported in the methods and in the results. Only p-values are reported and not the statistical tests performed.
	Responder bias on student level	Green	Data objectively reported
	Blinding	Red	Students were aware they were participating in the experiment, and teachers who provided the grades as well (the authors also recognize a potential teacher-level bias)
	(Dis)similarity of baseline characteristics	Orange	Variables such as ethnicity and other possible confounders were assessed but not considered in the analyses; higher percentage of whites were found in 2017 sample
	Appropriate statistical models which control for confounders	Orange	Better grades were predictive of school year but this was after accounting for other predictive variables. Authors thus tested year and not grade as dependent variable.
Milic 2014[53]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Orange	Grades are possibly subjectively reported from questionnaires but this is unclear; author(s) were contacted but did not respond

	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	(Dis)similarity of baseline characteristics	Orange	The different gender composition of the early schedule and late schedule samples is reported and discussed but not controlled for in the analyses
	Appropriate statistical models which control for confounders	Red	Simple Mann-Whitney Test without controlling for confounders
Wolfson 2007[61]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Green	Data objectively reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	(Dis)similarity of baseline characteristics	Orange	Schools are very similar, except for a higher percentage of white students in the school with later SSTs
	Appropriate statistical models which control for confounders	Orange	MANOVA with Bonferroni correction for multiple comparisons. One problem is that ethnicity was not controlled for although there were more white students in the school with later SSTs.
Wahl-strom 1997[62]	Reporting bias on author level	Red	All information necessary to critically read the results are missing (sample information, statistical analyses etc.)
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	(Dis)similarity of baseline characteristics	Orange	The author reports that the schools are similar but data are not reported. One caveat is that students from district B and C also spent more time on homework.
	Appropriate statistical models which control for confounders	Red	Insufficient information to be judged
Lewin 2017[69]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	(Dis)similarity of baseline characteristics	Green	Reported and controlled for
	Appropriate statistical models which control for confounders	Green	Yes
Wahl-strom 2014[63]	Reporting bias on author level	Red	Sample size (n students) for the grade analyses are not reported; statistical analyses are not reported (t-test, p-values); the average GPAs before and after the change are only reported in the appendix but without specifying the SD and range and therefore it is difficult to judge the effect size
	Responder bias on student level	Green	Grades are probably objectively reported (GPAs and test scores); author(s) were contacted but did not respond
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires). The authors do not report whether grades were collected also in other students that did not fill in questionnaires.
	Appropriate statistical models which control for confounders	Red	One can only infer that independent t-tests were used as the authors do not clearly report it. T-tests without controlling for co-variates would be problematic
Kelley 2017[68]	Reporting bias on author level	Orange	Sample characteristics are not well described. Statistical analyses and models are described, p-values are reported but t-tests and also Cohens-coefficients are not always reported
	Responder bias on student level	Green	Grades are objective
	Blinding	Green	Data retrieved from archive

	(Dis)similarity of baseline characteristics	Red	Not reported
	Appropriate statistical models which control for confounders	Red	No. Simple t-tests without covariates
	Control group present	Orange	Comparison with a national sample which is probably not a (gender) matched control group