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

Anna M. Biller [a, b, c, **], Karin Meissner [a, d], Eva C. Winnebeck [a, e], Giulia Zerbini [f, *]

[a] Institute of Medical Psychology, Ludwig Maximilian University Munich, Munich, Germany
[b] Graduate School of Systemic Neuroscience, LMU Munich, Germany
[c] Institute of Psychology, Bundeswehr University Munich, Munich, Germany
[d] Division of Health Promotion, Coburg University of Applied Sciences and Arts, Coburg, Germany
[e] Neurogenetics, Technical University of Munich, and Institute of Neurogenomics, Computational Health Center, Helmholtz Center Munich, Munich, Germany
[f] Department of Medical Psychology and Sociology, University of Augsburg, Augsburg, Germany

Introduction

Early school start times (SSTs) have been recognized as one of the leading causes of inadequate sleep in adolescents worldwide. They clash with the longer and later sleep needs of adolescents [e.g., 1–4], leading to wide-spread, chronic sleep restrictions in the student population [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 positive associations between delaying SSTs and sleep duration and/or 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 reasonable 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.

Academic achievement, however, is notoriously difficult to quantify, and the measures commonly used in school settings, grades and test scores, suffer from inherent limitations with regards to objectivity (e.g., teacher bias), reliability and validity [24]. This results from complex influences at many levels (student, family, teacher, school etc) as well as the generally broad scope of

* Corresponding author. Department of Medical Psychology and Sociology, Stenglinstraße 2, 86156, Augsburg, Germany.
** Corresponding author. Institute of Medical Psychology, Goethestrasse 31, 81373, Munich, Germany.
E-mail addresses: anna.biller@med.uni-muenchen.de (A.M. Biller), giulia.zerbini@med.uni-augsburg.de (G. Zerbini).

grades including important soft factors such as participation in class, effort, or behaviour [24]. Thus, grades and scores are strongly influenced by many factors that can and need to be considered in statistical analyses for meaningful interpretation of potential SST effects.

Maybe because of this difficulty, previous reviews have mostly summarized the relationship between SSTs and variables other than achievement (e.g., sleep, tardiness rates, absences, motor vehicle accidents and health [14,16,25]). We identified a total of 12 peer-reviewed reviews [15,16,25–34] – only three of them systematic reviews [15,16,31] – that discuss SSTs also in relation to academic achievement albeit not as their main focus. Most of the reviews concluded that the evidence was mixed, that the few positive effects reported in the studies were weak, and that many of the studies analysed suffered from methodological limitations. Minges and Redeker [16] and Marx et al. [15] reported the results of only two and three studies respectively, which limits their conclusions. Morgenthaler et al. [31] reviewed eight studies and none of these found significant improvements in academic achievement associated with delayed SSTs. Nonetheless, newspaper articles often purport it as established scientific fact that later SSTs improve academic achievement [35–37], while some public outreach programs also convey this message [38], mostly referring to single studies that found positive associations.

Since academic achievement shapes future career trajectories [39–41], answering the question whether delaying SSTs improves achievement goes beyond simple and genuine scientific curiosity, and 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 [42] and the ROBINS-I tool [43], we assessed the existing evidence of the relationship between SST and academic achievement and 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 vs. college students, who differ considerably in their sleep characteristics and class schedules. Our main question was whether changes in school start times in middle or high schools (or international equivalents) have any effect on academic achievement as measured by (standardised) test scores or course grades (both from subjective self-reports or objective records). Given the heterogeneity in study types and data treatment, intervention strength, exposure duration, data analysis and outcomes as well as the high risk of bias in many studies, we decided against an overall meta-analysis or meta-analyses on subgroups of studies [44]. Instead, 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

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. No restrictions were made with respect to languages, article types or year of publications. The following search string was 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 screened for further studies. We included two unpublished articles that are currently under review in peer-reviewed journals [45,46].

Study selection criteria

The recommended PRISMA guidelines for study selection, data synthesis and systematic reviews were followed [47]. Fig. 1 summaries the study selection process. After removal of duplicate records, the titles and abstracts of the retrieved records were screened for relevance regarding the study question and clearly irrelevant records were excluded. The full texts of the remaining articles were retrieved, screened and included for qualitative analysis if the following study selection criteria were fulfilled: 1) academic achievement was assessed as course grades or (standardised) test scores; 2) participants were middle school or high school students; 3) studies reported both a change/variation in SSTs and linked it to course grades or (standardised) test scores.

Data abstraction and treatment

AMB and GZ independently and systematically extracted predefined study characteristics as summarised in Table 1. Authors were contacted when information was missing, not clearly defined or if further analyses were available upon request. If authors did not answer or failed to provide the requested information, this is marked as “not available” (“NA”) in Table 1 and flagged orange or red (depending on the severity of risk of bias) in the reporting bias category of the risk of bias assessment (Fig. 3).

Both in Table 1 and Fig. 3, studies were grouped based on the type of data analysis performed with respect to grades/test scores (which sometimes differed to other assessed outcomes, such as sleep duration). We identified longitudinal analyses (i.e., within-subject) that investigated academic achievement in relation to a change in SSTs for a specific cohort of students over time 1) including a control group that did not change SSTs or 2) without a control group. Furthermore, we identified cross-sectional analyses that compared academic achievement of different, independent groups of students (i.e., between-subject) with varying start times either at one specific time point or over several years.

Risk of bias assessment

A pre-defined risk of bias assessment was conducted independently by two of the authors (AMB and GZ; Fig. 3). Given the lack of 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-RCTs. To this end, items from the GRADE scheme [42] and ROBINS-I tool [43] used for non-RCTs 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-RCTs are high risk by definition.

Allocation concealment bias: Researchers knew the sequence or method of randomisation and hence could predict the next allocation. Non-RCTs 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.

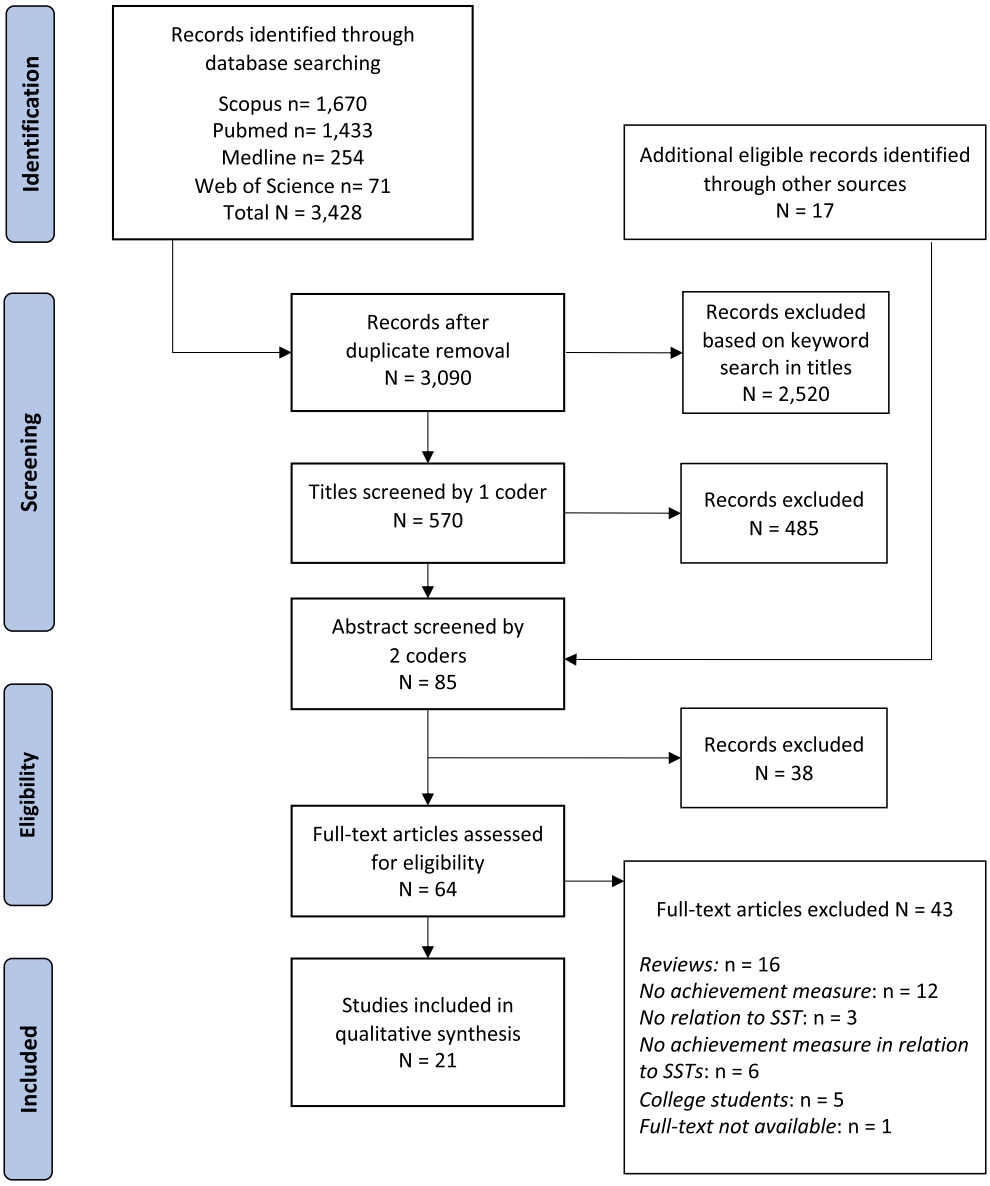

Fig. 1. PRISMA flowchart. The PRISMA flow diagram for our systematic review process detailing the database searches, the number of identified records, titles and abstracts screened, the final studies included in qualitative synthesis and reasons for exclusion of studies.

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 considered blinded. This criterion also covers a potential self-selection bias towards taking part in a study.

Dissimilarity of baseline characteristics: Authors did not check and/or report the dissimilarity of baseline characteristics between cross-sectional groups or between control and treatment groups.

Inappropriate statistical models: Statistical analyses did not account for confounders and/or were inappropriate for the given study type.

Cohort bias (no control group present): Longitudinal changes might be due to specific cohort characteristics (e.g., gender, ethnic background) 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. When assessment differed between AMB and GZ, mutual agreement was sought after discussion of critical points. If no agreement could be reached, two independent scorers (ECW and KM) evaluated the respective studies, and a consensus was found across all scorers. A total quality-of-evidence score was calculated as follows: scores for each bias category were added up (green contributed 1 point, orange .5 points and red 0 points) and then divided by the maximal possible score (eight for the longitudinal studies with control group, seven for the longitudinal studies without control group, and seven points for cross-sectional studies). The quality-of-evidence score was the proportion (%) of the maximum score (e.g., six out of max eight points = 75%). The

Table 1
Detailed descriptions of included studies. Studies are ordered by grade/test score analyses (longitudinal with control group, without control group and cross-sectional).
Abbreviations: ACT, American College Test; b, unstandardised beta coefficient; β, standardised beta coefficient; CG, control group; CI, confidence inverval; CSAT, College Scholastic Ability Test; DID, difference-in-difference estimation approach; ES, elementary school(s); FE, fixed effects; FRPL, free or reduced price lunch; GPA, grade point average; HS, high school (s); IG, intervention group; mo, month; MS, middle school(s); μ, average; NA, not available; n.s., not statistically significant on min. 5%-level; OLS, ordinary least square; OR, odds ratio; p, significance level; SES, socio-economic status; SD, standard deviation; SST, school start time; y, year. Terms describing statistical descriptions were taken from the original articles.

Study/ Risk of bias score	Study Design	Sample characteristics	Outcome	Type of analysis	Results	Key findings
LONGITUDINAL ANALYSES (within-subject) with control group						
Jung [61] Score: 6/8	Schools: sample drawn from Gyeonggi Education Panel Study: 85 ES (only age 11), 63 MS; some delayed (IG), others not (CG) SST: pre: 8:00–8:20, post: 9:00 Assessment: 2012–2017; change: 2014; pre/post: 3/2 y Exposure: 2 y Resolution: 1/y	**Group 1: longitudinal cohort** $N_{students}$(IG) = 2,562 $N_{students}$(CG) = 220 Grade: 4th-9th, Age: 11-16 Gender: 51% (IG)-58% (CG) male Ethnicity/race: NA SES: NA but see covariates **Group 2: cross-sectional cohorts** $N_{students}$(IG) = 2,562 (2015) $N_{students}$(CG) = 4,026 (2012) Grade: 7th, Age: 14 Gender: ~51% male Ethnicity/race: NA SES: NA but see covariates Location: Gyeonggi, South Korea	Korean, English & math test score, end of spring semester Scale: NA Provided by: governmental agency	DID & OLS regression Predictor: 9 o'clock policy implemented (binary: yes, no) Covariates student-level: 8 items Extended personal covariates: 18 items Extended other covariates: 5 items FE: y, individual	(1–3) Specifications with y FE & personal covariates only/and other covariates: increase in math (.27/.23/.22 SD, p < .05), English (.13/.24/.26 SD, p < .01), not Korean (.13/.11/.12 SD, n.s.) (4) Specifications with y FE & extended personal covariates: increase in English (.18 SD, p < .01), n.s. in math (.16 SD) & Korean (.05 SD) (5) Specification with y & individual FE: n.s. and negative in math (−.17) & English (−.14), Korean becomes p < .01 (−.29 SD) Cross-sectional robustness check confirms longitudinal results	- Increase in English scores, but no change in Korean/math test score when controlled for (extended) per-sonal covariates; - Korean scores decrease when individual FE is applied Caveat: sleep duration did not differ between CG & IG
Lenard [55] Score: 6/8	Schools: 19 HS; 5 advanced SSTs (IG) & 14 not (CG) SST: in IG: pre: 8:05, post: 7:25; in CG: 7:25 Assessment: 2008–2019; change: 2012/13; pre/post: 4/7 y Exposure: 7 y Resolution: 1/y	$N_{students}$~10,000 per each 8 cohorts $N_{observations}$ ≤52,854 (ACT scores) Grades: 8th-12th (inferred), Age: NA Gender: ~50% males Ethnicity/race: ~52% White, 26% (CG) vs. 23% (IG) African American, ~12% Hispanic, 11% other SES: 35% FRLP Location: Wake County, NC, USA	ACT scores in 11th grade (composite & individual scores for English, reading, math & science) Scale: 1–36 (= best) Provided by: Wake County administration	DID & comparative interrupted time series Predictor: SST change (binary: yes, no) Personal covariates: 5 items School-level covariates: 5 items FE: student, school, grade(sensitivity tests)	- n.s. results for ACT composite and individual subject ACT scores - independent of length of exposure ACT composite scores (all p > .05): Partial exposure: b = .023, early start all y: b = −.167, treated schools all: b = .273, p > .05 Scores were trending in all schools: math scores dropped over subsequent cohorts; English, reading & science rose	No changes in individual or composite ACT scores after start times were **advanced** (independent of length of exposure)
Edwards [54] Score: 5.5/8	Schools: 20 MS, 9 MS with total of 14 SST changes (9 delays, 5 advances), 11 MS no change SST: pre: 7:30–8:45, post: 7:30 −8:25, no change: 7:30–8:05 Assessment: 1999–2006; change within these years Exposure: NA Resolution: 1/y	$N_{students}$~15,000 per each 7 cohorts, $N_{observations}$ ≤102,506 Grades: 6th-8th, Age: 11–14.5 Gender: ~51% males Ethnicity/race: 21–24% African American, 2–10% Hispanic SES: 14–28% FRPL Location: Wake County, NC, USA	Reading & math end-of-year standardised test scores Scale: 0%–100% (inferred); converted to percentile scores/ student within grade & current year Provided by: Wake County administration	Pooled OLS models, quantile regression model Predictor: absolute SSTs/change in SSTs Covariates student-level: 8 items Covariates school-level: 5 items FE: student & school	Per 1h delay in SST: 1.8–2.9% points (.06-.07 SD) increase in maths, 1.0–3.4% points (.04–.05 SD) increase in reading when using within school variation or both within & between school variation (both p < .01) Selected covariate results for maths & school fixed effect (all p < .01): Black: b = −15.8, Hispanic: b = −5.4, female: b = −1.7, FRPL: b = −5.4, parent education (y):b = 2.6 Age effect: Conditional quantile effect of 1h later start on percentile rank on grades: Math: students in bottom half:b = ~2 −3; upper half: b = .75-2 Reading: students in bottom half: b~1.5 −2, upper half: b~0-1	- Better end-of-year standardised scores in maths & reading with later starts - Students who previously achieved lower grades & older students benefitted more
Shin [45]	Schools: all MS in 2 districts (599 schools in Gyeonggi (IG), 383 schools in Seoul (CG))	$N_{observations}$ ≤33,282 Grades: 7th-9th, Age: NA Gender: ~50% male	Semester grades (standardised) for math & reading	DID Predictor: 9 o'clock policy implemented	Increase in math (.03 SD) & reading grades (.02 SD; both p < .001)	Improvements in math & reading semester grades

Score: 5.5/8	SST: in IG: pre: ~8:20, post: 9:00; in CG: <8:00−9:00 Assessment: 2013−2015; change: 2014; pre/post: 1/1 y Exposure: 1 y Resolution: 1/semester	Ethnicity/race: mostly Asian (gender & ethnicity pers. comm. author) SES: NA Location: Gyeonggi & Seoul, South Korea	Scale: numeric; 0−100; normalized by distribution Provided by: Korean Education & Research Information Service	(binary: yes, no) Covariates student-level/other: 6 items Covariates school-level: 5 items FE: y, mo		
Kim [63] Score: 5.5/8	Schools: HS from 2 districts; Gyeonggi delayed (IG), Seoul not (CG) SST: in IG: pre: <7:40−9:00, post: 9:00; in CG: < 8:00−9:00 Assessment: 2009−2016; change: 2014; pre/post: 5/2 y Exposure: 2 y Resolution: 1/y	$N_{observations}$ = 1,479,131 Grades: 9th-12th, Age: 15-18 Gender: 52% males Ethnicity/race: NA SES: NA Location: Gyeonggi & Seoul, South Korea	(1) Ann. Nat. Assessment of Edu. Achiev. (Korean, math, English) for 9th & 11th grade (2) College Scholastic Ability Test (CSAT) for 12th grade Scale: NA Provided by: EduDataService System	DID Predictor: 9 o'clock policy implemented (binary: yes, no) Covariates student-level: 5 items FE: individual, year, region, school type Several robustness checks	Results 11th graders ($p < .01$): Math overall: .07−.1 SD Math male: .08-.14 SD Math female: .06-.07 SD Robust when adding covariates Korean & English scores n.s. when control variables are added (except male Korean: .07 ($p < .05$) Results 12th graders: For CSAT no statistically significant benefit after delay	- Better standardised scores in math not in Korean or English - Boys benefitted more - No change in CSAT scores (note: the CSAT is scheduled before 9:00!)
Rhie [62] Score: 2.5/8	Schools: several MS & HS; Gyeonggi district delayed (IG), 3 other districts not (CG) SST: in IG: pre: 7:30−8:10, post: 9:00 (MS delayed 30−60 min; HS delayed 60−90 min); CG: 7:30 −8:00 Assessment: 2012−2016; change: 2014; pre/post: 2/2 y; 2014 excluded Exposure: 2 y Resolution: 1/y	$N_{students}$(IG) = 42,517 $N_{students}$(CG) = 28,287 Grades: 7th-11th or 12th, Age: NA Gender: ~52% male Ethnicity/race: NA SES: NA Location IG: Gyeonggi district & Daegu/Gyeongbuk/Ulsan district, South Korea	Self-reported GPAs Scale: % of students having "high & moderate GPAs" Provided by: participants	Logistic regression analysis using complex samples to compare IG & CG by year Predictor: Group Covariates: NA (analyses across years by group not explained)	Percentage of students reporting "high & mid high GPAs" across years and separated by group. (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 $p < .05$	No change in self-reported GPAs
LONGITUDINAL ANALYSES (within-subjects) without control group						
Biller [46] Score: 3/7	School: 1 HS-equivalent SST: pre: mostly 8:00, post: 8:00 or 8:50 (daily choice) Assessment: 2013−2017; change: 2016; pre/post: 2.5/1.5 y Exposure: .5−1.5 y Resolution: 4/y	$N_{students}$ = 63-157 $N_{observations} \leq$16,724 Grades: 7th-12th, Age: 14-21 Gender: 30−40% males Ethnicity/race: NA SES: NA Location: Alsdorf, Germany	Quarterly grades of 12 subjects in 3 disciplines (sciences, social sciences, languages) Scale: numeric, 0% −100% (= best) Provided by: school registry	Linear mixed models 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), frequency of 8:50AM-use Covariates: gender, grade, discipline, academ. quarter	- Flexible system did not predict grades: .00 SD, $p > .05$ - Sleep duration did not predict grades: −.05 SD, $p > .05$ - Changes in sleep duration or chronotype (from baseline to the flexible system) did not predict grades - Except social jetlag (post: .03 SD, $p = .027$) Covariates (from models 3a-d): Male: .07 SD ($p > .05$) Grade level 12: .06 SD ($p < .001$) Quarter 4: .05 SD ($p < .001$) Social Sciences: .17 SD ($p < .001$)	No changes in quarterly school-reported grades
Thacher [56] Score: 2.5/7	School: 1 public HS SST: pre: 7:45, post: 8:30 Assessment: 2010−2014; change: 2012, pre/post: 2/2 y Exposure: 1−2 y Resolution: 1/y	$N_{students}$ = ~650−800 across 4 y (but t-test for cross-sectional comparisons shows ~250−330 students) Grades: 9th-12th, Age: μ~16.5 Gender: NA Ethnicity/race: NA SES: ~18% eligible for free lunch Location: Glen Falls, NY, USA	(1) Weighted average GPAs & subject-specific GPAs (2) Standardised test scores from Regents Exams for cross-sectional comparison Scale: numeric; 100 point scale for GPAs Scale for Regents Exam:	**Longitudinal comparisons:** mixed effect analyses including within-subject effect control **Cross-sectional comparisons:**	**Longitudinal comparison:** - No statistically significant evidence for change in GPA (overall & subject) - Higher grade levels (12th) better grade (p-value not reported), males worse ($p = .011$-$.059$), FRPL worse ($p < .001$) but independent of SST - No exact numbers are reported **Cross-sectional by grade level:** (1) **GPA**:	No (systematic) change in overall GPAs in longitudinal comparison - GPA test scores of 11th graders were higher in students

(*continued on next page*)

Table 1 (*continued*)

Study/ Risk of bias score	Study Design	Sample characteristics	Outcome	Type of analysis	Results	Key findings
			NA Provided by: school	independent-samples t-tests for grades & standardised test scores	11th graders' GPAs higher by 2.55% points with later SSTs: Pre mean: 78.79% (SD 11.11), post mean: 81.34 (SD 8.79), $t_{295} = 2.20$, $p = .028$ (2) **Regents exams**: 2 of 20 subject test scores (10th grade Earth Sciences & 11th grade Algebra) better before the change ($p < .007$)	who started school later - No systematic difference in test scores cross-sectionally
Wahl-strom [52] Score: 1.5/7	Schools: 7 HS SST: pre: 7:15, post: 8:40 Assessment: 6 y; change: 1997/98; pre/post: 3/3 y Exposure: 3 y Resolution: NA	$N_{students}$ = NA $N_{observations} \geq 1$ million Grades: 9th − 12th?, Age: NA Gender: NA Ethnicity/race: NA SES: NA Location: Minneapolis, MN, USA	All letter grades (semester & trimester grades) Scale: letter grades (categorical) Provided by: school district	Statistical analysis: NA Covariates: NA	"A small improvement in grades earned overall but not statistically significant" → No actual numbers are reported	No change in letter grades
Owens [49] Score: 1/7	School: 1 HS (boarding & day school) SST: pre: 8:00, post: 8:30 Assessment: Dec 2008, Mar 2009 Exposure: 2 mo (Jan−Mar) Resolution: 1/time point	$N_{students}$ = 201 Grades: 9th − 12th, Age: μ~16.5 Gender: ~43% males Ethnicity/race: NA SES: NA Location: Rhode Island, USA	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 = .70; 95% CI = .41−1.20, $X^2 = 1.71$, $p = .22$	No change in self-reported grades
Boergers [48] Score: 0/7	School: 1 HS (boarding school) SST: pre (t_1): 8:00, post (t_2): 8:25, back (t_3): 8:00 Assessment: Nov 2010 (t_1), Mar 2011 (t_2), May 2011 (t_3) Exposure: ~5 mo Resolution: 1/time point	$N_{students}$ = 197 Grades: 9th − 12th, Age: μ = 15.6 Gender: 41% males Ethnicity/race: 52% White, 7% African American, 6% Hispanic, 24% Asian, and 10% multiracial/other SES: NA Location: Rhode Island, USA	Self-reported grades Scale: categorical; "mostly Bs or better" Provided by: students	Statistical analysis: NA 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 significant change
CROSS-SECTIONAL ANALYSES (between-subject)						
Groen [57] Score: 5/7	Schools: 790 HS from Panel Study of Income Dynamics (nationally-representative sample) SST: 7:00−9:15 (average 7:53) Assessment: 2002/03, 2007/08 Exposure: NA Resolution: 1/y	$N_{students} \leq 1200$ Grades: 9th − 12th, Age: 13-18 Gender: 50% males Ethnicity/race: 61−65% White or other race, non-Hispanic, 15−22% Black, 9−16% Hispanic SES: 6−32% free or reduced-price lunch recipient Location: USA	Broad-reading test score & applied-problems (math) test score of the Woodcock-Johnson Revised Tests of Basic Achievement (age-adjusted) Scale: NA; normalised by survey year Provided by: NA (probably by research assistant)	Linear OLS model & Oster model (bounded effects); instrumental-variable estimates Predictor: SSTs Covariates student/ family/school/district/ county/state-level and sunlight: 8/5/5/3/1/1/2 items	OLS regression: per 1 h later SSTs: - higher scores in females' reading by .16 SD ($p < .01$) - math (males & females) & reading (male) n.s. - .27/.32 SD higher scores in reading for FRPL students (males & females) Oster model (bounded effects): Per 1 h later SSTs, .16-.28 SD higher scores in reading for females; .05-.12 SD higher scores in applied-problems for males →Likely mediated by longer (36 min) sleep duration/1 h of later SSTs for females but not for males	- Reading scores higher for females in schools with later SSTs, no differences for males - No difference in math scores for either gender
Hinrichs [53] Score: 5/7	Schools: 48 districts (73 schools); Minneapolis (incl. some suburbs) delayed (IG); St. Paul (incl. some suburbs) not (CG); school types NA SST	$N_{observations}$ = 196,617 $N_{students}$ = NA; slightly less than number of observations (pers. comm. author) Grades: 10th-12th, Age: NA	Individual composite ACT test scores Scale: numeric; 0−36 (=best)	OLS regression, quantile regression Predictor: SST (h post 7:00) Covariates student-	No association between SST and ACT scores (from full specification #8): Per 1 h later: b = −.02, p > .05 CI(SD) = -.23, .18 (−.05, .04) Subroup analyses by race, income,	No difference in ACT scores when comparing schools with various SSTs No differential results

	in IG: pre: 7:15, post: 8:40 SST in CG: 7:30 Assessment: 1993–2002; change: 1997/98; pre/post: 4/5 y Exposure: ~5 y Resolution: 1/y	Gender: 44% males Ethnicity/race: 79% White, 3% Black, 7% Asian, 1% Hispanic, 7% missing SES: 12% family income <$30,000 Location: Twin Cities metropolitan area, MN, USA	Provided by: ACT test company	level: 5 items Covariates school/ district-level: School length, set of school-specific linear time trends FE: grade, year	gender, quantile regression or large schedule change were n.s. (actual numbers not reported) Covariates: Males: b = .25, p < .01 Black: b = −2.47, p < .01 Low income: b = −.92, p < .01	depending on race, SES, or previous performance
	Schools: all HS in Kansas state SST: NA Assessment: 2000–2006 (reading/ maths 2001–2006; social science/ science 2000–2006) Exposure: NA Resolution: 1/y	$N_{schools}$ = 1,666 Grades: 10th-11th, Age: NA Gender: 40% White females, 9% non-White females, 9% non-White males Ethnicity/race: 72% White, 18% non-White SES: ~19% free lunch status Location: KS, USA	School-level test score data on state-wide Kansas Assessments in math, reading, science & social studies Scale: 0–100% (inferred) Provided by: Kansas Department of Education	OLS regression, quantile regression Predictors: SSTs (h post 7:00) Covariates: Length of School Day, racial distribution, FRS, school time trends FE: school, year	No association of SST and test scores (maths, reading, science, or social studies) For reading: 1 h later SSTs (from full specification #8): b = .95, p > .05	No difference in Kansas Assessment scores in math, reading, science or social studies
	Schools: 75 schools in 19 districts SST Change: NA; some delays Assessment: 2000–2007; change: 2001/02; pre/post: 1/6 y Exposure: ~6 y Resolution: 1/y	$N_{observations}$ = 171 (number of district-by-year pairs) Grade/Age/Gender/Ethnicity/race/ SES: NA Location: Virginia suburbs of Washington, DC, USA	End-of-course exams or standardised tests? Scale: 0–100% Provided by: Virginia Department of Education	OLS regression Predictors: SSTs Covariates: several on school-level (not detailed)	"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 & confirmed	No significant difference
Bastian [58] Score: 5/7	Schools: 410 HS (all public school students in NC) SST: 7:00–9:30; Of 410 schools 23 schools changed SSTs, 9 by ≥ 30 min; 44 districts (278 schools) had across-school variation in SSTs; average time difference earliest-latest: 33 min (range: 5–120 min); 69 districts (132 schools) had no variation Assessment: 2011–2015; change within these years Exposure: NA Resolution: NA(1/y?)	$N_{students}$ = 770,623 Grades: 9th-12th (inferred) Age: 14–18 (inferred) Gender: NA Ethnicity/race: White with 46% racial/ethnic minority (i.e., Black, Hispanic, American Indian, Asian, multiracial) SES: 49% eligible for free/reduced school meals Location: NC, USA	(1) Average course grades & course grades in 1st period classes in math, English, science & social studies Scale: 4-point scale, converted from numeric into unweighted grade points (2) Test scores from statewide standardised end-of-course exams (EOC) in algebra, biology, English Scale: standardised (3) ACT composite scores Scale: 0-36 Provided by: NC Department of Public Instruction	Linear regressions Predictors: SST Covariates student-level: 6 items Covariates school-level: 6 items Additional covariates (for EOC): >7 items FE: school & district (used for robustness checks)	(1) **Course grades**: - Overall: no association Per 1 h delay: .012 SD, p > .05 - Course grades in 1st period: ≥8:30 h vs. < 7:30 h start: .05 SD, p < .05 - Disadvantaged students (economic, lower achievement or minority) higher grades overall & in 1st period with later SSTs Per 1 h delay: .05-.07 SD, p < .05 or .01 (2) **EOC scores**: Mixed findings overall (incl. dis-advantaged) algebra: higher but p > .05; biology: lower, p < .05; English: lower but p > .05 (3) **ACT scores**: overall: .11 SD, p > .05; students with lower achievement: .28 SD, p < .05	- No difference in average course grades when comparing schools with various SSTs - 1st period grades higher with start ≥8:30 h - No significant results for EOC or ACT overall - Disadvantaged students benefitted more (grades & ACT)
Lewin [66] Score: 3/7	Schools: 26 MS with variable SSTs (country-wide surveillance data) SST: "Earliest": 7:20–7:30 "Early": 7:40 –7:55 "Late": 8:00–8:10 Assessment: 2008, 2010, 2012 Exposure: NA; since admission to school? (3 y?) Resolution: 1/time point	$N_{students}$ ~ 32,000 (pooled from all sample years) Sample 2008: n = 6,936; 2010: n = 11,991; 2012: n = 10,768 Sample "Earliest" SSTs: n = 7,206; Sample "Early" SSTs: n = 13,161; Sample "Late" SSTs: n = 12,613 Grade: 8th, Age: 13-14 Gender: 50% males Ethnicity/race: 42% White, 58% non-White SES: 22% low, 58% medium, 20%	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 Predictor: SSTs by group Mediator: sleep duration (Sobel test) Covariates student-level: survey y, gender, race Covariates school-level: FRPL Hierarchical structure:	Grades in "earliest schools" were lower (b/β = −.29 SD, p = .01), n.s. for "earlier schools" (b/β = −.11 SD, p = .13) compared to later starting schools (Note: unclear if coefficient was standardised) Association of SST overall mediated by sleep duration: b/β = .12, p < .001 Covariates: Female: b/β = .31, p < .001 Non-White: b/β = −.32, p < .001	- SSTs later than 30 min associated with better self-reported grades - Longer sleep duration associated with better grades

(*continued on next page*)

Table 1 (*continued*)

Study/ Risk of bias score	Study Design	Sample characteristics	Outcome	Type of analysis	Results	Key findings
		high FRS Location: NA but likely USA		students nested within schools	Free lunch status: up to b/β = −.67, p < .001	
Kelley [65] Score: 3/8	School: 1 English state-funded HS SST: pre = Year 0 (A): 8:50 post = Year 1−2 (B): 10:00, pre = Year 3 (A): 8:50 Assessment: 4 y Exposure: 1−2 y Resolution: 1/y	Year 0: $n_{students}$ = 169 Year 1: $n_{students}$ = 166 Year 2: $n_{students}$ = 164 Year 3: $n_{students}$ = 179 Grades: NA, Age: 14-16 Gender: NA Ethnicity/race: NA SES: NA Location: urban-area with achievement below national average, England	Standard National Examination (GCSE) Scale: G-A* (= best) Provided by: UK Office of National Statistics	T-test; Cohen's d & h; Value-added analysis; % students achieving "good academic progress" (i.e., ≥5 GCSE grades of C or better in English, math & min. 3 other subjects) Covariates: NA	Change in value-added as % of national (compared to national average, all p < .0005): Year 1 vs. 0: +15%; Year 2 vs. 0: +20%; Year 3 vs. 2: −7% % students making good academic progress compared to national average: Year 0: −40%, p < .0005 Year 1: −9%, p = .18 Year 2: −11%, p = .08; Year 3: −15%, p = .01	Later SSTs associated with higher % of students making good academic progress & higher value-added number compared to national average
Wolfson [60] Score: 3/7	Schools: 2 MS SST: School E: 7:15, School L: 8:37 Assessment: fall 2003, spring 2004 Exposure: NA; since admission to school? (2−3 y?) Resolution: 1	$N_{students}$ = 205 School E: 79, School L: 126 Grades: 7th (n = 99), 8th (n = 106) Age: NA Gender: 40% males Ethnicity/race: 46 vs. 60% White (School E vs. School L), 8−9% African American, 16−19% Hispanic, 6−10% Asian, 10−16% other SES: 18% FRPL Location: New England, USA	Fall quarter grade based on mean of English, science, math & social studies Scale: numeric; 0 −100% (= best) Provided by: schools	MANOVA, incl. Bonferroni correction for group comparisons Independent variables: school (=SST/cohort), grade, gender Covariates: no	Significant School × Grade interaction: F(1,208) = 17.06, p < .001; i.e., there were no school differences for 7th graders but 8th graders; School L students had higher grades (and more White students) than School E students: F(1,104) = 10.60, p < .01 7th grade: pre mean grade: 83.16% (SD 7.16), post mean grade: 80.46% (SD 10.11) 8th grade: pre mean grade: 76.85% (SD 9.45), post mean grade: 83.79% (SD 8.80)	- Higher average grades for 8th graders (not 7th graders) in school with later SST - No gender differences
Dunster [59] Score: 3/7	Schools: 2 public HS (RHS & FHS) SST: pre: 7:50, post: 8:45 Assessment: spring 2016 (pre) & spring 2017 (post) Exposure: NA; ~7 mo? Resolution: 1/y	$N_{students}$ = 178 total, from 2 independent samples: Sample 2016: 51 students (RHS) + 41(FHS), Sample 2017: 41(RHS) + 41(FHS) Grade: 10th, Age: μ~16 Gender: ~47% male Ethnicity/race: 76/75% & 2/19% White (2016/2017 RHS & FHS), 10/5% & 54/46% Asian, 6/5% & 7/7% Hispanic, 8/5% & 32/22% African American, 0/10% & 10/10% unknown/otherSES: 31% vs. 88% economically disadvantaged students (RHS vs. FHS) Location: Seattle, WA, USA	One 2nd semester grade from a biology lab class Scale: NA (probably 0% −100%) Provided by: teacher	Generalized linear models (binomial) predicting year (=SST/cohort) Predictors: 2nd semester biology grades Covariates: school, sleep offset, mood, chronotype, sleepiness	Grade was predictive of year (=SST/cohort) after adjusting for other variables e.g., sleep offset on schooldays No model coefficients provided Median grade 2016: 77.5% (mean: 74.6%) Median grade 2017: 82% (mean: 76.6%)	Higher biology grades were predictive of students stemming from year with later SST
Milić [64] Score: 2/7	Schools: 2 grammar, 2 vocational schools SST: weekly alternating morning or afternoon schedules; early schedule: 7:00/13:00 (2 schools); late schedule: 8:00/14:00 (2 schools) Assessment: May−Jun 2011 Exposure: NA; since admission to	$N_{students}$ = 821 Sample Early schedule: n = 452 Sample Late schedule: n = 369 Grade: NA, Age: 15−19 y Gender: across entire sample: 54% males; early schedule- sample: 73% males; late schedule-sample: 30% males Ethnicity/race: NA	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 < .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

	school? Resolution: once	SES: NA Location: Osijek, Croatia			
Wahlstrom [51] Score: 1/7	Schools: 6 HS, 3 districts SST: pre: 7:35−7:50, post: 8:00 −8:55 Assessment: MN: 2010−2011; CO: 2011−2012; WY: 2011/12 (pre) vs. 2012/13 (post) Exposure: 1 y? Resolution: 1/time point	$N_{students}$: NA (grade analyses) Grades: 9th-12th, Age: 13-19 Gender: 51% males Ethnicity/race: 70% White, 9% African American, 7% Hispanic, 6% Asian, 8% Other (indicative only!) SES: NA Location: MN/CO/WY, USA	(1) English, maths, social studies, science grades in 1st & 3rd period-classes or GPAs Scale: categorical; "mostly A's = 9" to "mostly F's" = 1 (2) State-wide achievement tests or PLAN Scale: NA Provided by: GPAs by districts; categorical grades by students	Independent t-tests, correlations Covariates: NA	(1) **GPA** All grade levels: 3 schools reported higher GPAs with later starts, 2 mixed results, 1 n.s. (2) **ACT/PLAN** Math: 1 school reported increases in test scores, 1 decreases, 3 n.s., 1 NA; Reading: 5 n.s., 1 NA; Writing: 1 decrease, 4 n.s., 1 NA; Science: 4 n.s., 1 NA, 1 not tested Composite ACT or PLAN scores: higher scores in 2 schools with later starts, n.s. results in 4 other schools	- In 3 out of 6 schools, higher GPAs with later starts independent of grade level - 11th graders GPA always higher when data was available - Mixed & often non-significant associations with standardised test scores
Wahlstrom [50] Score: .5/7	Schools: MS & HS in 3 districts; 1 district delayed, 2 not SST: MS: district A (IG): 7:35, B: 8:00, C: 8:00; HS: district A (IG): 8:30, B: 7:25, C: 7:15 Assessment: NA Exposure: NA Resolution: NA	$N_{students}$ = a not further defined sample was drawn from 7,168 students of 17 districts Grades: 10th − 12th & 7th − 8th Age: NA Gender: NA Ethnicity/race: NA SES: NA Location: MN, USA	Self-reported grades Scale: NA Provided by: participants	Statistical analysis: NA Covariates: NA	Mean self-reported grades in district A were highest ($p < .05$) compared to district B & C for 10-12th graders (District A: 7.08, B: 6.50, C: 6.37) but not for 7-8th graders (District A: 6.66, B: 6.91, C: 6.60)	- HS-students who started later reported higher grades - MS-students who started later reported better or similar grades

different bias categories were not weighted. We defined scores <25% as low, ≥25% and <75% as moderate and ≥75% as good.

Results

Literature search

Overall, 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 many clearly irrelevant titles, a second automated search with the same search string was carried out on titles only, resulting in 570 articles. One coder (AMB) then screened titles manually and excluded 485 studies. The abstracts of the remaining 85 studies were *independently* screened by two coders (AMB and GZ), who selected a total of 50 studies of which 39 studies were overlapping (80% inter-rater agreement; step not shown in Fig. 1). The inclusion of the 11 studies for which there was no initial agreement was further discussed and eight of these 11 were included for full text reading resulting in 47 chosen articles (i.e., 38 exclusions) plus 17 articles identified through other sources leading to a total of 64 studies. After full text reading, AMB and GZ selected 21 studies that fulfilled all inclusion criteria (Fig. 1).

Study characteristics and quality

In the following paragraphs, summary information concerning the 21 included studies are reported (see also Fig. 2 and Table 1). For written summaries of individual studies please refer to the SI.

School type and cohort characteristics

Most studies were conducted in the US (13) [48−60], followed by South Korea (4) [45,61−63], Germany [46], Croatia [64], England [65], and one unknown location [66] (Fig. 2a and Table 1). The majority of studies collected data in high schools (>900 schools), of which two were also boarding-schools [48,49], two grammar schools and two vocational schools [64]. Other school types were middle schools (>140 schools) [45,50,54,60−62,66] and elementary schools (85, not considered here except for students aged 11 in 4th grade in [61]). In one study, school type was not specified [53]. The sample sizes varied drastically between 157 and >770,000 individual students and up to >1 Million 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 and 19 y. Gender ratios ranged from 30 to 60% males with a median of 50% males (five studies did not report gender ratios) [50,52,56,58,65]. Most included participants were White (2−79%), followed by Hispanic/Latino (1−19%), Black/African American (3−32%), Asian/Pacific Islanders (5−54%; except for the Korean studies that included "mostly Asian participants"), and other races/ethnicity (1−16%). Ten studies did not report ethnicity/race [46,49,50,52,56,61−65], while of the remaining studies only four considered ethnicity/race in their statistical analysis [53,54,58,66]. SES was mostly measured as free or reduced lunch status and ranged from 6 to 49%.

Study types based on the data analysis performed on grades or test scores

We identified longitudinal analyses (within-subject) and cross-sectional analyses (between-subject) based on the type of grades/test score data analysis performed. The 11 studies with longitudinal analyses all included a change in SSTs [45,46,48,49,52,54−56,61−63]. However, only six studies included an additional control group with no change in SSTs [45,55,61−63] or with both no change and advance of SSTs [54](Fig. 2b). Of the 10

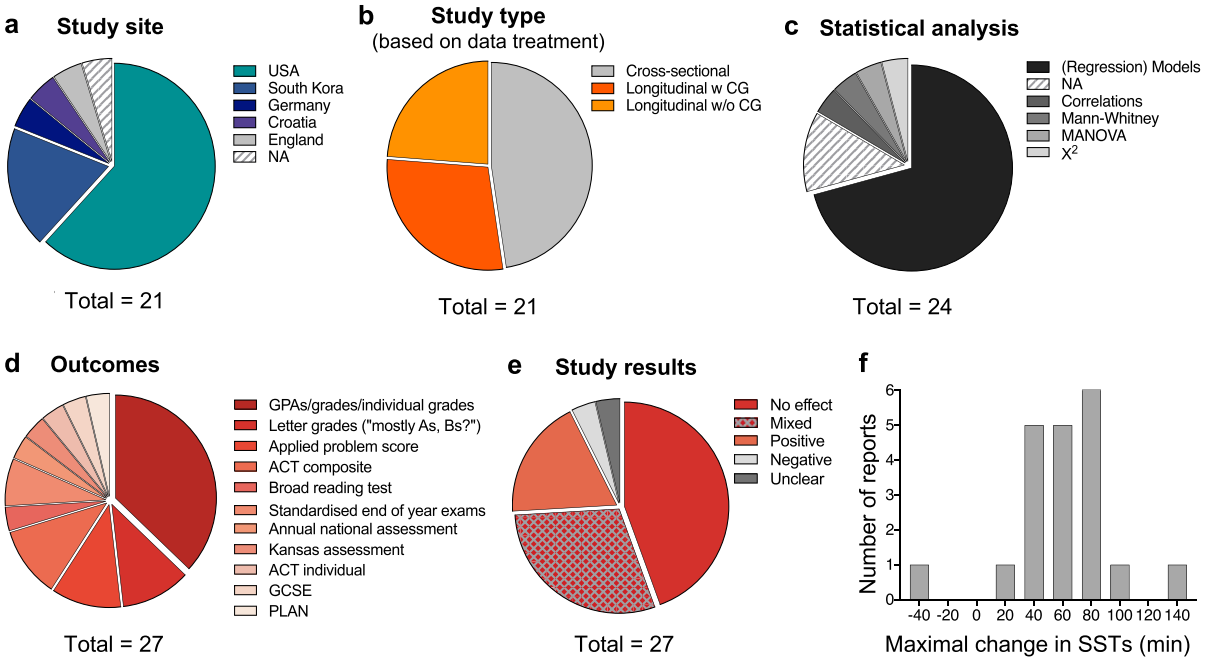

Fig. 2. Characteristics of included studies. a-e, Pie charts depicting key characteristics and main findings of the 21 studies included in the final review. Since several studies used multiple types of analyses or assessed multiple outcomes, the total number in c,d,e is > 21. **f,** Histogram displaying 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.

	Longitudinal analyses (within-subject) with control group						Longitudinal analyses (within-subject) without control group					Cross-sectional analyses (between-subject)									
	Jung [61]	Lenard [55]	Edwards [54]	Shin [45]	Kim [63]	Rhie [62]	Biller [46]	Thacher [56]	Wahlstrom [52]	Owens [49]	Boergers [48]	Groen [57]	Hinrichs [53]	Bastian [58]	Lewin [66]	Kelley [65]	Wolfson [60]	Dunster [59]	Milić [64]	Wahlstrom [51]	Wahlstrom [50]
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	X	X										X	X	X							
At least 50% good evidence score	X	X	X	X	X							X	X	X							
Results	→	→	↑	↑	↑→							↑→	→	↑→							

Fig. 3. Risk of bias assessment. Included studies are ordered based on their 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. NA, not applicable. [1]These studies also included cross-sectional grades or scores analyses (between-subject); either as a robustness check or as secondary analyses. For results on these see the result section and Table 1. [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 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 = .5 points; red = 0 points. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

studies with cross-sectional analyses, four studies compared schools in various districts without an intervention but based on their different SSTs [57,60,64,66]. The remaining studies with cross-sectional analyses included a change in SST, providing repeated between-subject analyses of schools or districts over approximately one [59] or several years [53,58], or at one time point after the change [50,51]. One study with a cross-sectional analysis 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 y [65].

Statistical analyses

A vast range of different statistical analyses was reported (Table 1 and Fig. 2c). Notably, regressions were the dominant analysis method, ranging from general OLS regressions [53,54,57,58,61], quantile regression [53,54], difference-in-difference methods [45,55,61,63] and binomial regressions [59,62] to linear mixed models [46,56] and path analysis with probit regression [66]. One study reported Oster models with bounded effects and instrumental estimates [57]. Another study used MANOVA [60], while several simpler analysis methods not controlling for covariates were also used. These were t-tests [51,56,65], X^2-tests [49], Mann-Whitney Test [64] and correlations [51]. Notably, several studies did not report the nature of their statistical analyses [48,50,52].

Study outcome measures

About half of the studies provided grades as outcome measures, while the other half provided (standardised) test scores (Table 1, Fig. 2d). However, since several 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) [53,55,58], national achievement scores or PLAN scores [51], standardised test scores from Regents Exams [56], standardised end-of-course exams [58], annual national assessment of achievement in South Korea [63], GCSE in the UK (General Certificate of Secondary Education) [65], and the broad-reading test score and applied-problems (math) test score from the Woodcock-Johnson Revised Test of Basic Achievement [57]. These test scores were objectively reported (except for Groen et al., for which the source was unclear [57]) (Fig. 2d). The remaining studies analysed other types of objective scores or grades [45,46,52,54,58,60,61], subjective grades [48–51,59,62,66], while in one study the exact outcome was unclear (the authors only stated “final grade in last semester”) [64]. Sampling resolution was mostly once per year, the highest reported resolution was once per academic quarter (Table 1) [46]. An overview of the study results (positive, negative, mixed or no associations between later SSTs and grades/test scores) is depicted in Fig. 2e.

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

The SST delay reported was on average 64 min (median = 60, SD = 26) with a range of 25–135 min (Fig. 2f). This average is based on the maximal delay reported by each study and thus an approximation. Since some studies only provided SST ranges or a minimal start delay, the numbers are not precise. One study investigated exclusively SSTs advances by 40 min [55]. One study changed to a flexible SST in which students could choose daily whether to attend school at 8:00 h or 8:50 h [46]. Exposure duration to the (new) start time ranged from 2 mo to 7 y (Table 1). 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 study results

Overall, five studies found clear positive associations of later/delayed school starts with academic achievement [45,50,54,65,66], five reported mixed results [51,57,58,60,63], eight did not detect significant associations [46,48,49,52,53,56,61,62], one reported a negative association [64], and one study's finding was unclear [59] (Fig. 2e, Table 1). One study investigated advancing SSTs by 40 min and found no changes in ACT scores after the change [55]. Notably, of the 21 studies, four studies investigated the same 9 o'clock policy (i.e., delay of SST to 9:00 h) in South Korea [45,61–63]. 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 two studies by Wahlstrom et al. conducted in the same district: the report in 2002 [52] might be a longitudinal follow-up of the report from 1997 [50], but we were unable to confirm this. In the following, we grouped the studies based on the type of analyses performed, school type (middle vs. high school) and outcome measure (grades vs. test scores) to identify potentially hidden associations.

Longitudinal vs. cross-sectional analyses

Of the studies with longitudinal analyses and control group, two reported positive effects [45,54], three studies found no effect [55,61,62] and one showed mixed results [63], while the studies without a control group found overall no associations between delaying SSTs and academic performance [46,48,49,52,56]. Of the studies with cross-sectional analyses, three found positive associations [50,65,66], one negative [64], one no associations [53], four mixed results [51,57,58,60], and one result was unclear [59].

Middle vs. high school students

Two studies investigated both middle and high school students but did not report the results separately [61,62]. Five studies investigated exclusively middle-school students [45,50,54,60,66], all of which measured course grades (not test scores, presumably since middle-school students usually do not take standardised tests yet). Of these studies, all reported positive associations between later SSTs and grades. In contrast, positive associations were found only in two [56,58] out of 10 studies that investigated grades as outcome measures in high school students and these were only found in specific subgroups of these students (e.g., older students, females, disadvantaged students).

Grades vs. test scores

Course grades and standardised scores most likely reflect different aspects of students' learning and knowledge and could thus be differentially sensitive to SST changes. However, when grouping studies based on outcome measure, there was no tendency or differential results on either 1) objective test scores (two positive [54,65], three null findings [53,55,56], and four mixed results [51,57,58,63]), 2) objective grades (two positive [45,59], four null findings [46,52,56,61] and two mixed results [58,60]), 3) or self-reported grades (two positive [50,66], three null findings [48,49,62], and one each for mixed [51] and negative [64]).

Risk of bias assessment

To judge the evidence quality of the included studies, we performed a risk of bias assessment (Fig. 3). 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 seven out of 21 studies. This meant that over half of the studies stayed below 75% of the good-evidence-score within their respective category. Therefore, the quality of the evidence can be deemed only moderate which also precluded conducting a meta-analysis [44].

On the positive side, especially the longitudinal studies with a control group showed a high evidence quality with two [55,61] out of six studies reaching at least a 75%-score and three more studies [45,54,63] >50%. Two studies [45,63] 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).

Discussion

Chronic sleep restriction in adolescents has become a serious health concern worldwide [e.g., 8,67]. 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., 3,68,69]. Delaying school start times has the great potential of improving cognitive functioning, physical health and well-being of students mediated by improved sleep (as reviewed elsewhere [16,25,28]) with possibly relatively little costs [70,71]. But does a delay in SSTs also translate into improved academic achievement? Our systematic literature search identified 21 studies that investigated whether SSTs have any systematic effect on/are associated with course grades or standardised test scores in middle and high school students. The analyses revealed that about half of the studies did not find any positive effect or association, while the other half found mixed, positive and unclear results. Given the high risk of bias 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 and to allow for conducting a meta-analysis (see the Research Agenda for suggested improvements).

Methodological considerations

Our systematic risk of bias assessment showed that the evidence level of included studies was mostly moderate, with eight out of 21 studies achieving a score of ≥50% within their respective 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 [72]. 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 only used in 14 out of 21 studies.

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. We decided against a subgroup meta-analysis on the five studies with longitudinal analyses and control group that reached >50% good evidence score [45,54,55,61,63] given that we could not exclude the possibility that four of them included data collected in the same school districts [45,61–63]. This would seriously mislead the interpretation of the results and was thus deemed inappropriate.

Studies with low risk of bias also reveal no clear picture

What do studies with low risk of bias (i.e., a ≥75% good-evidence-score) conclude about the influence of SSTs on academic achievement? Among the longitudinal studies, Lenard et al. (2020) [55] found that advancing SSTs by 40 min was not linked with changes in ACT scores, while Jung (2018) [61] showed that delaying start times by 40–60 min was also not associated with changes in 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 .03–.07 SD gains in math and .03–.05 SD gains in reading [45,54], with a larger effect for students previously achieving lower scores [54], and one reports small effects on math but not on Korean nor English [63]. Three cross-sectional studies also achieved a good evidence score of >75% [53,57,58]. 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 reading and problem-solving items from the Woodcock-Test but only for females and reading [57], 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 [53]. Bastian and Fuller (2018) reported that an 8:30 h or later start was necessary for positive associations with 1st period grades [58]. Furthermore, the authors showed that especially students with low average achievement in the past, students from ethnic minorities and students with a low SES tend to benefit from later starts. In summary, good evidence studies report either no, relatively small, or non-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 [73] or racial segregation [74]. However, general test scores might be less sensitive to acute changes in SSTs because they measure the accumulated knowledge over several schooling years [55]. Different standardised tests might also measure different underlying concepts and knowledge. Moreover, standardised tests are often scheduled in the morning [53] and therefore confounded by time-of-day effects on attention and fluid intelligence (e.g., logic, reasoning, problem solving) [55,75–77]. In the case of ACT or PLAN scores, tests are usually only taken by students with a record of good grades and who also apply for admission to college – a specific student population, who is prone to ceiling effects, making these students less likely to benefit from later SSTs compared to students with lower academic achievement, as two other studies also confirmed [54,58].

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 [78]. A "teacher bias" can also particularly influence the results of interventional studies if not controlled for. Moreover, differences in average grades between school districts could be especially problematic when comparing SSTs in cross-sectional multi-sites studies. Schools and teachers could also globally adjust their grading system if overall students' grades improve, precluding the detection of positive effects. However, large-scale studies, especially when many covariates are collected, also allow to tease apart between-school, between-class and between-teacher effects when accounted for by appropriate statistics. 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.

Differences between middle and high school students

Given that chronotype markedly delays during puberty [e.g., 3,79] and that students with later chronotypes have been shown to achieve on average lower grades [80], one could expect that high school students benefit more from later SSTs than middle school students. Unfortunately, the number of studies investigating exclusively middle school students (mostly grades 6th-8th) were only five out of 21 [45,50,54,60,66], and no study directly compared the effect of delaying SSTs on course grades between middle and high school students, thus limiting our conclusions. Still, the proportion of studies which showed positive associations with course grades (no study investigated test scores, presumably because middle school students do not take standardised tests yet) was higher among the middle school studies, an unexpected result that should be further addressed in future studies.

Considerations of power, magnitude of associations and dose

An alternative explanation for the mixed results could be a lack of statistical power in small-scaled studies, which could lead to more null-findings. However, only six studies had a sample size of about $N \leq 200$ [46,48,49,59,60,65]. Among these studies, the results were also mixed as in the studies with greater sample sizes. Overall, most of the 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 magnitude of associations between SSTs and achievement was generally relatively low: only four studies reported standardised beta coefficients >.15, which equals to effect sizes of $r = .2$ [81]. Traditionally, these are considered small effects according to Cohen (1988) [82] but considering Hattie's interpretation for education settings, $r = .2$ is at the bottom of "the zone of desired effects" [81].

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 [53] tried to model this hypothesis using spline regressions but found no clear answer. Furthermore, the latest start time in the studies reviewed here was 10:00 h and the largest delay was 135 min (Fig. 2f). 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 Paediatrics recommends starting schools not earlier than 8:30 h [83], which is supported by Bastian and Fuller [58] who found that only when school started at 8:30 h or later, significant positive effects were detected on 1st period grades only, 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 multitude of 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 [80], ethnic or racial background [58], conscientiousness [78] or prior knowledge [84]) to family-level factors (e.g., parental involvement [85], parental education [61], or SES [86,87]), and to classroom- and school-level factors (e.g., classroom size [73] and atmosphere [84], teacher quality and assessment style [88]).

Regarding student or family-level factors, structural disadvantages (e.g., being from an ethnic minority or low SES background) may require particular attention in the study of SSTs and achievement, since these disadvantages are linked with both lower achievement and suboptimal sleep [e.g., 87]. For example, students with difficult social backgrounds are prone to reduced, poorer and more variable sleep than their more advantaged peers [87,89,90]. This may arise from longer commuting times, less parental monitoring of bedtimes, media use, more caffeine/drugs intake, or more family conflicts/economic problems provoking anxiety and stress – all of which can contribute to shorter and poorer sleep [87,91] and, in turn, to lower achievement [92]. Therefore, students with a low SES might particularly benefit from later SSTs in terms of sleep AND achievement, so that disparities can be reduced. This is supported by stratified analyses performed in three of the included studies, showing that especially students from economically disadvantaged backgrounds, from ethnic/racial minority groups and who perform at the lower end of the achievement distribution benefitted from later starts [54,57,58], although Hinrichs did not find any differential effects for these groups [53]. Future studies should continue this important work to identify target groups that particularly benefit from later SSTs.

Next to stratified analyses, there is also a lack of mediation and moderation analyses that could shed some light on the mechanisms behind possible improvements in academic achievement. Only a few studies tested the mediating role of sleep duration, reporting that later SSTs were associated with longer sleep, which, in turn, was linked to better academic achievement [e.g., 57,66]. Such analyses could test whether mediating variables that might need time to improve, such as sleep and learning, could have positive effects on academic achievement long-term. In general, reflecting on confounders, mediating and moderating variables, their influence on academic achievement and on how they might also be affected by changes in SSTs is important for future study designs and analyses.

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 non-peer reviewed studies to reduce a possible publication bias in favour of positive results. Previous reviews [e.g., 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 [34]). However, lifestyle 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 [79,93].

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 type, quality and chosen outcome measure and consequently a lack of generalisability of individual study results that also prevented conducting a meta-analysis in line with recommendations issued by the Cochrane collaboration (see the Research Agenda 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 an SST-delay. This suggests that SSTs could be delayed while academic achievement is very likely 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 since students are likely better rested and therefore cognitively more capable and efficient (but this needs to be assessed in future studies). In combination with other reported positive outcomes on sleep, daytime sleepiness, mood and motivation, computer gaming, attendance rates, or tardies and suspensions [e.g., 14,15,25,27,28], this likely remains a valid argument in favour of delaying SSTs.

Practice points

1. Grades and test scores influence future academic and working opportunities, however clear and systematic evidence on whether school start times improve academic achievement is currently lacking largely due to methodological shortcomings and heterogeneity of studies.
2. Later school start times might have stronger effects in specific sub-populations (e.g., later chronotypes or students achieving lower grades) and might depend on the amount of delay and exposure length, which future studies should clarify.
3. There are likely no adverse consequences of later start times since achievement levels were either maintained (with *potentially* less cognitive effort or time spent studying) or increased in all studies reviewed here (except for one).

Research agenda

To clarify the evidence and provide policy makers and educators with evidence-based guidelines, future studies should focus on:

1. Planning

- **Design**: multi-site, longitudinal (intra-individual) studies with pre-post analyses including a control group are recommended; randomisation at class or school level is feasible and could improve the evidence quality
- **Sample size**: large sample sizes are necessary to control for the numerous covariates to be considered and small effect sizes and effects potentially only occurring in sub-groups
- **Placebo/nocebo effects**: assessment and control of expectations of students, teachers and parents should be implemented
- **Achievement measures**: recommended are high-resolution, objective grades from a range of different academic subjects and across the year or standardised test scores. Note: grades and scores measure different concepts/capacities; avoid self-reports, composite scores ("mostly A"), or low resolution (subject, teacher, time)

2. Statistical analyses

- **Appropriate statistics** for the (nested and/or extensive longitudinal) study design, which consider influence of covariates and time trends
- **Stratified analyses** to detect sub-group effects (e.g., students showing high vs. low achievement, disadvantaged students)
- **Mediation analysis** to identify pathways
- **Analysis of dose response effects** of amount of delay/advance of start time and duration of exposure to the new start time

3. Reporting in detail about:

- **Study designs and data treatment**
- **Outcome variables** (grading scales, standardised tests, etc.)
- **Basic demographics of the studied population** (incl. $N_{students}$, $N_{observations}$)
- **Effect sizes** (also relative to outcome scales)
- **Educational system** (brief overview for international readers)

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.

Conflicts of interest

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

Acknowledgements

We thank all contacted authors who replied and helped us to report their findings as accurately as possible.

Abbreviations

ACT	American College Test
b	unstandardised beta coefficient
CG	control group
CSAT	College Scholastic Ability Test
GCSE	General Certificate of Secondary Education
GPA	grade point average
IG	intervention group
μ	average
NA	not available/applicable
OLS	ordinary least squares
OR	odds ratio
RCT	randomised controlled trial
SD	standard deviation
SES	socio-economic status
SST	School start times
y	years

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.smrv.2021.101582.

References

[1] Crowley SJ, Van Reen E, LeBourgeois MK, Acebo C, Tarokh L, Seifer R, et al. A longitudinal assessment of sleep timing, circadian phase, and phase angle of entrainment across human adolescence. PLoS One 2014;9(11).

[2] Crowley SJ, Acebo C, Carskadon MA. Sleep, circadian rhythms, and delayed phase in adolescence. Sleep Med 2007;8(6):602–12.

[3] Crowley SJ, Wolfson AR, Tarokh L, Carskadon MA. An update on adolescent sleep: new evidence informing the perfect storm model. J Adolesc 2018:55–65.

[4] Carskadon MA, Acebo C, Jenni OG. Regulation of adolescent sleep: implications for behavior. Ann N Y Acad Sci 2004;1021(1):276–91.

[5] Gibson ES, Powles ACP, Thabane L, O'Brien S, Molnar DS, Trajanovic N, et al. "Sleepiness" is serious in adolescence: two surveys of 3235 Canadian students. BMC Publ Health 2006;6:1–9.

[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 Med Rev 2012;16(3):203–11.

[7] Keyes KM, Maslowsky J, Hamilton A, Schulenberg J. The great sleep recession: changes in sleep duration among US adolescents, 1991-2012. Pediatrics 2015;135(3):460–8.

[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 Med 2011;12(2):110–8.

[9] Raniti MB, Allen NB, Schwartz O, Waloszek JM, Byrne ML, Woods MJ, et al. Sleep duration and sleep quality: associations with depressive symptoms across adolescence. Behav Sleep Med 2017;15(3):198–215.

[10] Baum KT, Desai A, Field J, Miller LE, Rausch J, Beebe DW. Sleep restriction worsens mood and emotion regulation in adolescents. J Child Psychol Psychiatry Allied Discip 2014;55(2).

[11] Short MA, Gradisar M, Lack LC, Wright HR. The impact of sleep on adolescent depressed mood, alertness and academic performance. J Adolesc 2013;36(6):1025–33.

[12] Garaulet M, Ortega FB, Ruiz JR, Rey-López JP, Béghin L, Manios Y, 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 2011;35(10):1308–17.

[13] Mullington JM, Haack M, Toth M, Serrador JM, Meier-Ewert HK. Cardiovascular, inflammatory, and metabolic consequences of sleep deprivation. Prog Cardiovasc Dis 2009;51(4):294–302.

[14] Bowers JM, Moyer A. Effects of school start time on students' sleep duration, daytime sleepiness, and attendance: a meta-analysis. Sleep Health 2017;3(6):423–31.

*[15] Marx R, Tanner-Smith EE, Davison CM, Ufholz LA, Freeman J, Shankar R, et al. Later school start times for supporting the education, health, and well-being of high school students. Campbell Syst Rev 2017;13(1):1–99.

*[16] Minges KE, Redeker NS. Delayed school start times and adolescent sleep: a systematic review of the experimental evidence. Sleep Med Rev 2016;28:82–91.

[17] Beebe DW, Rose D, Amin R. Attention, learning, and arousal of experimentally sleep-restricted adolescents in a simulated classroom. J Adolesc Health 2010;57(5):523–5.

[18] Killgore WDS, Kahn-Greene ET, Lipizzi EL, Newman RA, Kamimori GH, Balkin TJ. Sleep deprivation reduces perceived emotional intelligence and constructive thinking skills. Sleep Med 2008;9(5):517–26.

[19] Hysing M, Haugland S, Bøe T, Stormark KM, Sivertsen B. Sleep and school attendance in adolescence: results from a large population-based study. Scand J Publ Health 2015;43(1):2–9.

[20] Walker MP, Stickgold R. Sleep, memory, and plasticity. Annu Rev Psychol 2006;57:139–66.

[21] Stickgold R. Sleep-dependent memory consolidation. Nature 2005;437(7063):1272–8.

[22] Maquet P. The role of sleep in learning and memory. Science 2001;294(5544):1048–52 (80-).

[23] Alhola P, Polo-Kantola P. Sleep deprivation: impact on cognitive performance. Neuropsychiatric Dis Treat 2007;3(5):553–67.

[24] Allen JD. Grades as valid measures of academic achievement of classroom learning. Clear House A J Educ Strategies, Issues Ideas 2005;78(5):218–23.

[25] Wheaton AG, Chapman DP, Croft JB, Chief B, Branch S. School start times, sleep, behavioral, health and academic outcomes: a review of literature. J Sch Health 2017;86(5):363–81.

[26] Gomez Fonseca A, Genzel L. Sleep and academic performance: considering amount, quality and timing. Curr Opin Behav Sci 2020;33:65–71.

[27] Wolfson AR, Carskadon MA. Understanding adolescents' sleep patterns and school performance: a critical appraisal. Sleep Med Rev 2003;7(6):491–506. 2004/03/17.

[28] Alfonsi V, Scarpelli S, D'Atri A, Stella G, De Gennaro L, D'Atri A, et al. Later school start time: the impact of sleep on academic performance and health in the adolescent population. Int J Environ Res Publ Health 2020;17(7).

[29] Berger AT, Widome R, Troxel WM. Delayed school start times and adolescent health. In: Sleep and health. Santa Monica, CA, United States: Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN, United States RAND Corporation; 2019. p. 447–54.

[30] Hershner S. Sleep and academic performance: measuring the impact of sleep. Curr Opin Behav Sci 2020;33:51–6.

[31] Morgenthaler TI, Hashmi S, Croft JB, Dort L, Heald JL, Mullington J. High school start times and the impact on high school students: what we know, and what we hope to learn. J Clin Sleep Med 2016;12(12):1681–9.

[32] Wahlstrom KL, Owens JA. School start time effects on adolescent learning and academic performance, emotional health and behaviour. Curr Opin Psychiatr 2017;30(6):485–90.

*[33] Wolfson AR, Ziporyn T. Adolescent sleep and later school start times. In: Sleep, health, and society: from aetiology to public health; 2018. p. 215–23.

[34] Fuller SC, Bastian KC. The relationship between school start times and educational outcomes. Curr Sleep Med Reports 2020:18–9.

[35] Schmidt S. Later school starts linked to better teen grades [Internet]. 2019 [cited 2020 Dec 21]. Available from: https://www.sciencenewsforstudents.org/article/later-school-starts-linked-better-teen-grades.

[36] Urton J. Teens get more sleep, show improved grades and attendance with later school start time, researchers find [Internet]. 2018 [cited 2020 Dec 21]. Available from: https://www.washington.edu/news/2018/12/12/high-school-start-times-study/#:~:text=12 in the journal Science,minutes of sleep each night.

[37] Lee K. More evidence finds that delaying school start times improves students' performance, attendance, and sleep [Internet]. 2018 [cited 2020 Dec 21]. Available from: https://www.everydayhealth.com/kids-health/delaying-school-start-times-improves-students-performance-health/.

[38] Ackerman X, Phan T, Gee A, Kim A, Imani S, Welkie D, et al. School start times [Internet]. 2019 [cited 2020 Dec 21]. Available from: https://ccb.ucsd.edu/_files/bioclock/Infographic PDF. School Start Times 2019, Ackerman, Phan, Gee, Kim, Imani, Welkie, Golden.pdf.

* The most important references are denoted by an asterisk.

[39] French MT, Homer JF, Popovici I, Robins PK. What you do in high school matters: high School GPA, educational attainment, and labor market earnings as a young adult. E Econ J 2015;41(3):370–86.
[40] Geiser S, Santelices MV. 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 2007;35.
[41] Ma J, Pender M, Welch M. Education pays 2016. Coll Board Trends High Educ Ser; 2016. p. 1–44.
[42] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence - study limitations (risk of bias). J Clin Epidemiol 2011;64(4):407–15.
[43] Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct;355:i4919.
[44] Deeks J, Higgins J, Altman D. Chapter 10: analysing data and undertaking meta-analyses [Internet]. In: Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al., editors. Cochrane handbook for systematic reviews of interventions; 6.2. 2021. Available from: www.training.cochrane.org/handbook.
[45] Shin J. Sleep more, study less? The impact of delayed school start time on sleep and academic performance [Internet]. 2018 [cited 2021 Jul 29]. Available from: https://repositories.lib.utexas.edu/bitstream/handle/2152/65930/SHIN-DISSERTATION-2018.pdf?sequence=1.
[46] Biller AM, Molenda C, Obster F, Zerbini G, Förtsch C, Roenneberg T, et al. Are flexible school start times associated with higher academic grades? A 4-year longitudinal study. bioRxiv; 2021. https://doi.org/10.1101/2021.07.29.452310.
[47] Moher D, Liberati A, Tetzlaff J, Altman D. Preferred reporting items for systematic reviews and MetaAnalyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097.
[48] Boergers J, Gable CJ, Owens JA. Later school start time is associated with improved sleep and daytime functioning in adolescents. J Dev Behav Pediatr 2014;35(1):11–7.
[49] Owens JA, Belon K, Moss P. Impact of delaying school start time on adolescent sleep, mood, and behavior. Arch Pediatr Adolesc Med 2010;164(7):608–14. 2010/07/07.
[50] Wahlstrom KL, Frederickson J, Wrobel G. School start time study : technical report, volume II analysis of student survey data. 1997.
[51] Wahlstrom KL, Dretzke BJ, Gordon MF, Peterson K, Edwards K, Gdula J. Examining the impact of later high school start times on the health and academic performance of high school students. A Multi-Site Study; 2014.
[52] Wahlstrom K. Changing times: findings from the first longitudinal study of later high school start times. NASSP Bull 2002;86(633):3–21.
*[53] Hinrichs P. When the bell Tolls: the effects of school starting times on academic achievement. Educ Finance Policy 2011;6(4):486–507.
*[54] Edwards F. Early to rise? The effect of daily start times on academic performance. Econ Educ Rev 2012;31(6):970–83.
*[55] Lenard M, Morrill MS, Westall J. High school start times and student achievement: looking beyond test scores. Econ Educ Rev 2020;76.
[56] Thacher PV, Onyper SV. Longitudinal outcomes of start time delay on sleep, behavior, and achievement in high school. Sleep 2016;39(2):271–81.
*[57] Groen JA, Pabilonia SW. Snooze or lose: high school start times and academic achievement. Econ Educ Rev 2019;72(5):204–18.
*[58] Bastian KC, Fuller SC. Answering the bell: high school start times and student academic outcomes. AERA Open 2018;4(4). 2332858418812424.
[59] Dunster GP, de la Iglesia L, Ben-Hamo M, Nave C, Fleischer JG, Panda S, et al. Sleepmore in Seattle: later school start times are associated with more sleep and better performance in high school students. Sci Adv 2018;4(12):eaau6200.
[60] Wolfson AR, Spaulding NL, Dandrow C, Baroni EM. Middle school start times: the importance of a good night's sleep for young adolescents. Behav Sleep Med 2007;5(3):194–209.
*[61] Jung H. A late bird or a good bird? The effect of 9 o'clock attendance policy on student's achievement. Asia Pac Educ Rev 2018;19(4):511–29.
[62] Rhie S, Chae KY. Effects of school time on sleep duration and sleepiness in adolescents. PLoS One 2018;13(9):e0203318.
*[63] 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 2018;1–26.
[64] Milić J, Kvolik A, Ivković M, Čikeš AB, Labak I, Benšić M, et al. Are there differences in students' school success, biorhythm, and daytime sleepiness depending on their school starting times? Coll Antropol 2014;38(3):889–94.
[65] Kelley P, Lockley SW, Kelley J, Evans MDRR. 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 2017;11(12).
[66] Lewin DS, Wang G, Chen YI, Skora E, Hoehn J, Baylor A, et al. Variable school start times and middle school student's sleep health and academic performance. J Adolesc Health 2017;61(2):205–11.
[67] Chattu V, Manzar M, Kumary S, Burman D, Spence D, Pandi-Perumal S. The global problem of insufficient sleep and its serious public health implications. Healthcare 2018;7(1):1.
[68] Carskadon MA. Factors influencing sleep patterns of adolescents. In: Adolescent sleep patterns: biological, social, and psychological influences. New York: Cambridge University Press; 2002.
[69] Wittmann M, Dinich J, Merrow M, Roenneberg T. Social jetlag: misalignment of biological and social time. Chronobiol Int 2006 Jan 7;23(1–2):497–509.
[70] Hafner M, Stepanek M, Troxel WM. The economic implications of later school start times in the United States. Sleep Health 2017;3(6):451–7.
[71] Jacob BA, Rockoff JE. Organizing schools to improve student achievement: start times, grade configurations, and teacher assignments. Hamilt Proj 2011;9:24.
[72] Illingworth G, Sharman R, Jowett A, Harvey C-JJC-J, Foster RG, Espie CA. Challenges in implementing and assessing outcomes of school start time change in the UK: experience of the Oxford Teensleep study. Sleep Med 2019;60:89–95.
[73] Krueger AB, Whitmore DM. 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 2001;111(468):1–28.
[74] Card D, Rothstein J. Racial segregation and the black–white test score gap. J Publ Econ 2007;91(11–12):2158–84.
[75] Hansen M, Janssen I, Schiff A, Zee PC, Dubocovich ML. The impact of school daily schedule on adolescent sleep. Pediatrics 2005;115(6):1555–61.
[76] Fimm B, Brand T, Spijkers W. Time-of-day variation of visuo-spatial attention. Br J Psychol 2016;107(2):299–321.
[77] Zerbini G, van der Vinne V, Otto LKM, Kantermann T, Krijnen WP, Roenneberg T, et al. Lower school performance in late chronotypes: underlying factors and mechanisms. Sci Rep 2017;7(1):4385.
[78] Rimfeld K, Kovas Y, Dale PS, Plomin R. True grit and genetics: predicting academic achievement from personality. J Pers Soc Psychol 2016;111(5):780–9.
[79] Roenneberg T, Kuehnle T, Pramstaller PP, Ricken J, Havel M, Guth A, et al. A marker for the end of adolescence. Curr Biol 2004 Dec 29;14(24):1038–9.
[80] Zerbini G, Merrow M. Time to learn: how chronotype impacts education. PsyCh J 2017;6(4):263–76.
[81] Lenhard W, Lenhard A. Berechnung von Effektstärken [Internet]. 2016 [cited 2021 Jul 8]. Available from: https://www.psychometrica.de/effektstaerke.html.
[82] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale NJ: Erlbaum Associates; 1988.
[83] American Academy of Pediatrics. School start times for adolescents. Pediatrics 2014;134(3):642–9.
[84] Neumann K, Kauertz A, Fischer HE. Quality of instruction in science education. In: Fraser BJ, Tobin KG, McRobbie CJ, editors. Second international handbook of science education. Berlin: Springer; 2012. p. 247–58.
[85] Juang LP, Silbereisen RK. The relationship between adolescent academic capability beliefs, parenting and school grades. J Adolesc 2002;25(1):3–18.
[86] Pokropek A, Borgonovi F, Jakubowski M. Socio-economic disparities in academic achievement: a comparative analysis of mechanisms and pathways. Learn Indiv Differ 2015;42:10–8.
[87] Buckhalt JA, El-Sheikh M, Keller P. Children's sleep and cognitive functioning: race and socioeconomic status as moderators of effects. Child Dev 2007;78(1):213–31.
[88] Rockoff JE. The impact of individual teachers on student achievement: evidence from panel data. Am Econ Rev 2004;94(2):247–52.
[89] Jarrin DC, McGrath JJ, Quon EC. Objective and subjective socioeconomic gradients exist for sleep in children and adolescents. Health Psychol 2014;33(3):301–5.
[90] El-Sheikh M, Kelly RJ, Buckhalt JA, Benjamin Hinnant J. Children's sleep and adjustment over time: the role of socioeconomic context. Child Dev 2010;81(3):870–83.
[91] Pereira ÉF, Moreno C, Louzada FM. Increased commuting to school time reduces sleep duration in adolescents. Chronobiol Int 2014;31(1):87–94.
[92] Dewald JF, Meijer AM, Oort FJ, Kerkhof GA, Bögels SM. The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: a meta-analytic review. Sleep Med Rev 2010;14:179–89.
[93] Dahl RE, Allen NB, Wilbrecht L, Suleiman AB. Importance of investing in adolescence from a developmental science perspective. Nature 2018;554(7693):441–50.