OF TEENAGE OWLS AND EARLY SCHOOL

Longitudinal Effects of Delayed and Flexible School Start Times on Sleep, Psychological Benefits, and Academic Performance

Anna Magdalena Biller

Graduate School of Systemic Neurosciences
LMU Munich

Dissertation at the
Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

December 2020

Supervisor
Prof. Dr. Till Roenneberg
Institute of Medical Psychology
Ludwig-Maximilians-Universität München

First Reviewer:	Prof. Dr. Till Roenneberg
Second Reviewer:	Prof. Dr. Benedikt Grothe
External Reviewer:	Prof. Dr. Henrik Oster
Date of Submission:	31 December 2020
Date of Defense:	26 April 2021

There is a time for many words,
and there is also a time for sleep.

Homer

TABLE OF CONTENTS

0
Abstract

The circadian clock, an internal timing system that synchronises to the 24h light-dark cycle, governs sleep-wake timing in humans. Sleep is beneficial for numerous physiological processes, e.g. neuronal network organisation, immune function, memory and learning, or (mental) health. Acute and chronic sleep restriction in turn increases the risks to develop a variety of diseases or to adopt unhealthy behaviours, such as smoking. Adolescents are at particular risk of suffering from sleep restriction since their circadian clock delays during puberty – they tend to become night owls – which severely clashes with early school start times. Several studies have investigated the impact of delaying school start times on various outcomes. While positive effects have been described for sleep, long-term effects on sleep and clear effects on academic achievement are missing.

We thus conducted two studies investigating the long-term effects of a *flexible* school start system (daily choice of 8:00 or 8:50-start) on sleep, psychological outcomes and academic achievement in a secondary school in Germany. On average, all students – independent of chronotype, gender or age – extended their sleep by ~1h immediately and after 1 year in the flexible system but only on days when school started later. Sleep duration did not increase in the flexible system overall compared to the old start system (fixed start mainly at 8:00) which was connected to a relatively low uptake of the late start option. Importantly though, on average students did not phase-delay and additionally reported numerous psychological benefits both in the flexible system and on later start days, e.g. better well-being and sleep quality, increased concentration, higher motivation to attend school or better learning quality. Students also reduced their alarm-driven waking. Additionally, girls were more successful in keeping stable sleep onset times longitudinally on later start days compared to boys. Based on linear mixed model analyses including 4 years of ~17,000 individual, quarterly grades, we did not identify any increase in academic performance in the flexible system per se, or with sleep or sleep changes. On the other hand, covariates such as quarter, grade level or discipline robustly and significantly predicted grades in our sample.

To put our findings into perspective and clarify the evidence on this controversial topic, we also conducted a systematic literature review. Following the PRISMA guidelines, we synthesised the evidence of 21 studies that investigated the impact of school start times on grades and (standardised) test scores. We included a systematic risk of bias assessment that identified relatively poor reporting and medium evidence level of included studies. We concluded, that grades/scores are suboptimal outcome measures and that the current evidence (quality) does not allow to make any sound conclusions on whether school start times substantially impact grades or scores.

The flexible system is a widely neglected school start system that offers unique benefits for students. Vey importantly, it allowed teenagers to maintain their grades while, at the same time, they clearly benefitted in terms of sleep and psychological outcomes. Future studies need to study effects on cognition and learning in the real-life context and extend studies periods >1 year of follow-up to identify effects that might emerge after longer time periods or for specific (groups) of individuals.

1

General introduction

1.1. Rhythms of life[a]

1.1.1. A short history of the world in four rhythms[b]

> *"Above the Earth is the region of the ether; and still higher is the vault of the heaven. Beneath this vault the Sun, Moon and stars perform their motions, rising out of Okeanos in the morning and returning thither at night."* [1]

At the beginning, the cosmological ideas of the Greek philosophers were more a mixture of Greek and Babylonian mythology before advancing to more serious philosophical yet still speculative concepts[1]. Above, Homer refers to the flat, disc-like Earth, which was partly covered by the sea and surrounded by the great river Okeanos from which all things originate. While this sounds very poetic, it is not quite what we think today. Indeed, it was not until Thales, the founder of the Ionian School[2] that cosmological theories were separated from ancient mythology, which brought him the fame of being the first scientific philosopher we know of today[1]. He proposed water to be the first of everything, the essence of life. A circadian biologist might be more drawn to Heraklit of Ephesus' concept of the world - the universal flux. For him, "all things flow", which brought him to believe that fire was the first principle[1]. But circadian researchers could also sympathise with Pythagoras: "Number not merely represents the relations of the phenomena to each other but is the substance of things, the cause of every phenomena of nature"[1]. Pythagoras and his followers came to this conclusion by observing the regular movements of the celestial bodies and how the harmony of musical sounds depends on regular intervals[1].

It is fascinating that in one way or another, they all had a point, reminiscent of the great parable of the blind men describing parts of an elephant[3]. While these *physiologoi* (φυσιολόγοι), as Aristotle called them[4], disagreed on many minor or major parts, one thing was undoubted: the predictable and constant movement of the stars, the sun, the moon, and the ocean.

> "*The earth turned on its axis and split time into day and night. Its tilt gave us seasons*."
>
> (Kreitzman & Foster)[5]

With the celestial constellations in place, the history of the world might have started off with a light-dark rhythm of about 22 hours (or less) since Earth was several hours quicker than today[c] to rotate around its axis[6]. According to the *Giant impact hypothesis*, a proto-Earth consisting only of rock and

[a] Title from a wonderful and inspiring book by Prof. Russell Foster and Leon Kreitzman[5].

[b] By analogy with BBC Radio 4's „A History of the World in 100 Objects".

[c] Since then, days on earth have steadily become longer leading to the 24h-day we know today. This means that Earth's angular velocity is 4 min per longitudinal degree which translates to 15° every hour[42].

lava was hit by a protoplanet called Theia 4.5 billion years ago that ejected material into space, some of which eventually consolidated and formed the moon[7]. Earth's tilt of 23.5° was probably the result of this crash, so if we assume this theory to hold true, seasonal rhythms as we know them today (365.24 days) probably evolved only afterwards. The orbiting moon gave rise to lunar rhythms of 29.53 days and, with water on Earth, also to intertidal rhythms of about 12.8 hours, which might have timed marine and coastal life. These four rhythms – circadian, seasonal, lunar and tidal – govern much of our life on earth and thus "provide the temporal cues for the coordination of most behaviours ranging across daily feeding patterns, daily and seasonal migration, growth, reproduction, hibernation and much else".[8]

1.1.2. Biological clock (works)

All organisms, from bacteria, through plants, to insects and humans, orientate their behaviour and physiology towards predictable changes in their surroundings presumably because this allows anticipation of rhythmic changes in the environment: food sources come and go, predators only hunt at specific times, and daylength and thus temperature rapidly and profoundly change across the day and season. Adapting to these variations enables organisms to occupy spatial and social niches but also temporal ones across the 24h-day. The first biological clocks might have evolved in ancestors of cyanobacteria about 3 billion years ago[9] but other organisms with internal timing subsequently appeared, such as nocturnal, diurnal or crepuscular animals[10]. Another advantage of a regular organisation is that it prevents everything within the body from taking place at the same time, giving the body a structure[10]. The circadian clock tells the time of a 24h-day; thus, it follows rhythms that are about (lat: *circa*) a day (*dies*) long, a term coined by Franz Hallberg[11] to stress that the period of the internal clock of humans is approximately 24h, but not precisely - as will be described below. A short review of the history of circadian clock research and its clockworks will be given in the next sections. Other rhythms, such as circannual (*i.e.* seasonal), circatidal or circalunar rhythms are also extremely interesting but not the topic of this thesis. They have been reviewed elsewhere and I highly recommend referring to these resources[*e.g.* 8].

A short history of clock research

The history of circadian research is inextricably intertwined with Jean Jacques Otrous de Mairan, a French astronomer who is famous for a wonderful yet simple experiment with a *mimosa* plant in 1729[12,13]. By enclosing the mimosa in his cupboard, he deprived it from sunlight and subjected it to constant darkness[13]. Nevertheless, the plant kept unfolding and closing its leaves in a regular fashion. This was very intriguing since it remained mysterious how the plant knew the time of day. One hypothesis was that the plant could have sensed the variation in temperature across the day[14]. Consequently, Henri Louis Duhamel du Monceau placed some plants in salt mines in which constant temperature conditions prevailed - but he also observed the rhythmic leaf movement[15]. In 1832, Alphonse de Candolle, a Swiss botanist, then noticed that leaf movements varied between individual plants under constant conditions – they slightly deviated from 24h[12,16]. This was an exciting piece of evidence that the rhythm he observed was endogenous (*i.e.* within the plant): if they merely received input from the outside, all individual plants should have exhibited the same period. It took another 100 years until subsequent studies by Erwin Bünning[17] and Colin Pittendrigh[18], among others, confirmed this so-called *free-running rhythm* under constant conditions, thus it became clear that plants and animals have an endogenous clock of a period near 24h which free-runs when they are deprived of external cues which would otherwise synchronise their activity. But what are these external cues, also called *Zeitgeber* since they "give time" to the organism?

To investigate this more closely in humans, Aschoff and Wever built a bunker in the South of Germany in the 1960s to control for all potential influences from the outside world, ranging from daylight, to temperature, electromagnetic fields or even cosmic ray showers[5,19]. Deprived of any time cues, participants (often students who were writing up their thesis) lived in the bunker for several weeks while their core body temperature, hormonal secretion, loco-motor activity and sleep was studied[19]. Aschoff and colleagues found evidence that individuals extended their rhythm's period by up to 4h[20], which sometimes resulted in awkward moments when students were still happily writing up their thesis but the experiment was already finished – some had "lost" up to several days by constantly shifting later (a shortening of up to 5h was also observed). Although these were rare exceptions, the findings from the famous "Bunker experiments" and other observations, such as from Colin Pittendrigh[18] or Nathaniel Kleitman[21] (one of the fathers of sleep research), again supported the notion of an i) endogenous, ii) self-sustained oscillator that free-runs under constant conditions with a period *tau* (τ) of almost, but not exactly 24h, and that also showed iii) temperature-compensation. The last feature is worth another note: according to the Q_{10} rule, biochemical reactions double in speed with a rise of 10°C (as reviewed in[22]). Thus, if an endogenous clock does not compensate for exogenous temperature variations it would speed up or slow down as a function of temperature. A hamster with a rhythm of 24h at 20°C would suddenly exhibit a rhythm of only 12h at 30°C – a possibly fatal behaviour if waking hours now coincided with those of a predator.

Circadian clock features

As soon as the notion of an endogenous clock was accepted the question begged where such an oscillator would be in the body, and how it would work? One of the most important findings in circadian research was the location of the central pacemaker - the master clock of the body - in the suprachiasmatic nucleus (SCN) in mammals. The SCN is a small cell assembly of about 20.000 neurons in the anterior hypothalamus above the optic chiasm, the crossing of the optic nerves[23]. Transplantation studies revealed in the 1980s that gold hamsters without the SCN got completely arrhythmic but when the SCN of another hamster was implanted they exhibited the donor's period[24,25]. How does the SCN sustain its rhythmicity without any external input? Research into the fine-tuned and complicated molecular clockwork, which consists of several interlocking feedback loops, is still ongoing[*e.g.* 26]. However, the main concept of a transcriptional-translation feedback loop in the SCN could already be described in 1990: put simply, a gene is transcribed in the cell's nucleus and subsequently translated into a protein in the cytoplasm of the cell; the protein then enters the nucleus and suppresses its own transcription[27]. Eventually, the protein will be degraded, triggering a new start of the cycle (Fig. 1). In mammals, the key genes involved in these loops are *clock* and *bmal1* whose proteins form a heterodimer and initiate transcription of *period* (per1, 2, 3) and *chryptochrome* (cry1, 2). Once *per* and *cry* are translated into proteins, they also from a heterodimer and are transported to the cell's nucleus via phosphorylation processes (casein kinases 1 δ/ε) where they inhibit *bmal1* and *clock* and consequently their own transcription (Fig. 1).

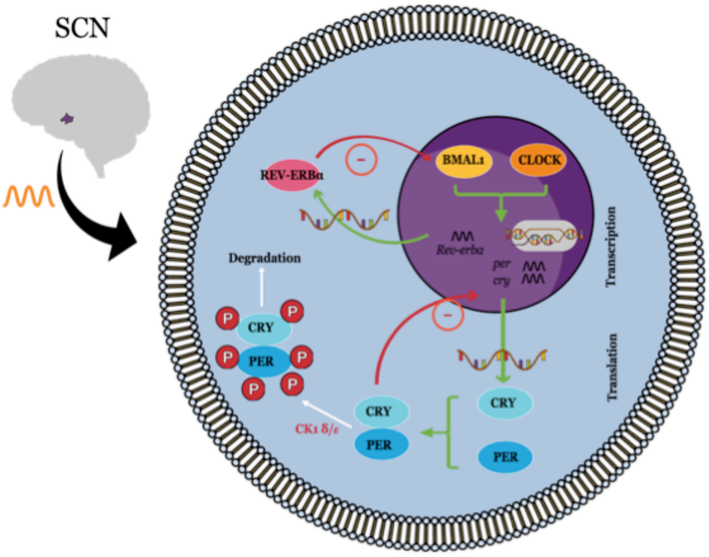

Fig. 1 | The transcriptional-translation feedback loop in mammals. The SCN in the hypothalamus produces rhythmic expression of *CLOCK* and *BMAL1* genes. CLOCK and BMAL1 proteins then form a heterodimic transcriptional activator (forward loop, green arrows) in the nucleus producing rhythmic gene expression of *per*, *cry*, and *REV-erbα*. After transcription in the cell, CRY and PER proteins are translated in the cytoplasm and also form a heterodimer. Post-translational modifications are made by CK1 that phosphorylates the PER-CRY (marked by the P), degrades it and also localizes it back to the nucleus where CRY-PER eventually stops its own transcription (red arrows). A secondary loop is REV-erbα, which is also rhythmically expressed and transcribed in the nucleus, then translated in the cytoplasm and eventually inhibits transcription of BMAL1. The RNA string indicates translation in the cytoplasm. Abbreviations: SCN, suprachiasmatic nucleus; per, period; cry, cryptochrome; CK1, casein kinase 1. Figure created in the Mind the Graph platform (*www.mindthegraph.com*).

We know today that probably every cell in our body contains the molecular make up for such a negative feedback loop, meaning that each cell contains an oscillator[28]. However, these peripheral clocks need input from the SCN to prevent dampening over time, thus they are not self-sustained[29], (although this has also been questioned[28]). For this reason, the SCN has been traditionally viewed as the master clock, a sort of conductor, who orchestrates the downstream clocks to keep the entire body playing in harmony (Fig. 2).

But how does the SCN actually manage to align with the outside world? It has been shown that the SCN receives light input from the retina of the eyes directly through the retino-hypothalamic tract[30], which entrains the self-sustained rhythm of the SCN to the outside world[31]. So far, no other *non-ocular* photoreceptors have consistently been found[32,33]. However, an unknown type of retinal cells was identified when mice were observed to still entrain to light stimuli even though they lacked rods and cones (the traditional photoreceptors;[34,35]). Amazingly, 1% of retinal ganglion cells were shown to contain a photopigment called melanopsin and are therefore also able to sense light (*i.e.* they are photosensitive[36,37]. These cells were eventually termed intrinsically photosensitive retinal ganglion cells (ipRGCs)[35,38,39] and are sufficient to entrain an animal through light[40]. It has subsequently been found that rods and cones, however, are also important in photic entrainment but are not strictly necessary[38].

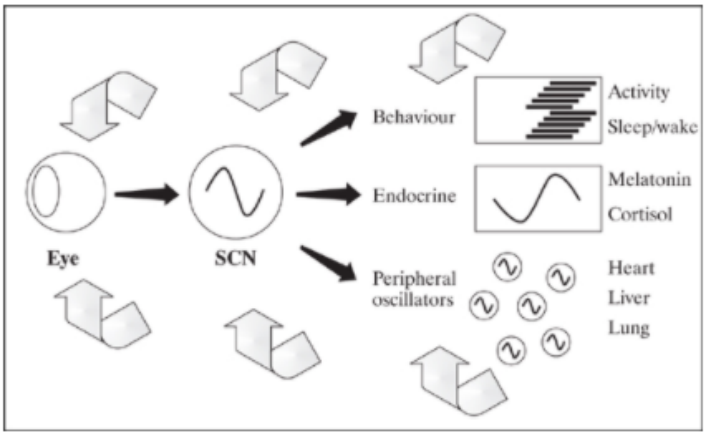

Fig. 2 | Schematic of the relationship between the eye, SCN and downstream rhythms. *Source*: Kreitzman & Foster[5] (reproduced with permission).

1.1.3. (Phase of) entrainment

A clock is only really useful if it accurately tells the time. A good clock says it is 12:00h when it is actually 12:00h (according to the sun). Let us imagine your watch tells you it is 11:00h even though it is 15:00h. If your watch has the same period (tau=24h) and is temperature compensated (see chapter 1.1.2) you could still tell the time if you managed to figure out the phase-angle difference of 4h between the two, since your watch and the sun time show a stable phase relationship. In this case, the two are locked, or in other words: your watch is synchronised to the sun but shows a phase angle difference of 4h (the phase of entrainment). But if you do not have a high-quality Swiss model or the batteries need recharging, your watch might slow down, hence it extends its period and the stable relationship is lost. In this case, the two have different periods and are out of phase.

It is useful to keep this picture in mind to understand how our endogenous clock keeps ticking in synchrony with the world around us. Biological clocks need environmental signals, *Zeitgebers*, to keep them entrained. Entrainment is an active synchronisation process, in which the circadian clock of an organism assumes a phase relationship with an external rhythm[41]. For most organisms, the main zeitgeber is the natural light-dark (LD) cycle of the sun that provides appropriate i) long light durations, ii) strong light intensity differences between day (photoperiod) and night (scotoperiod), and iii) has a period of 24h[42]. Under these conditions, an organism can entrain to the sun. Such an internal time keeping mechanism is very useful for any organism, since it allows for anticipation and thus prediction of cyclic events beyond mere reaction to a stimulus such as the light. The notion of entrainment is so critical that it constitutes the third characteristic of an endogenous clock, besides self-sustainability and temperature compensation[18,43,44]. Historically, there have been several ideas on how to conceptualise entrainment of organisms, ranging from non-parametric to parametric and integrated approaches of entrainment. They all assume that entrainment is the result of a matching internal period (τ) with the external (light-dark) period of 24h (T), *i.e*, entrainment is reached when τ = T. In analogy, entrainment happens when your watch is in accordance with sun time. These concepts, however, slightly differ and are briefly reviewed in the next section.

Non-parametric and parametric entrainment

One could test the responsiveness of such a system to light by exposing an animal to transient light pulses across the 24h-day when it otherwise lives in constant darkness (DD). Colin Pittendrigh did exactly this and plotted the magnitude of the phase shift of an animal (Δϕ) against external time (T) which results in a *phase response curve* (PRC)[45]. Such PRCs usually show a phase advancing and phase delaying portion (on the y-axis), separated by a dead zone in which exposure to a brief light stimulus has no phase-shifting effect – the animal keeps its old phase. The term non-parametric refers to the underlying idea that a light pulse has the power to instantaneously make the old phase "jump" to a new phase but keeping the oscillator's velocity stable. Jürgen Aschoff understood light as a continuous stimulus and assumed that the endogenous oscillator itself alters its velocity to adjust to perturbating light signals and thus ensures entrainment (τ = T)[46]. He constructed *velocity response curves* (VRC) that are derived from PRCs and display how the system speeds up or slows down in response to light at a given time point (phase).

Integrated approach to entrainment - CIRCs

These concepts use transient or continuous light as the main zeitgeber (also termed *photic-entrainment*) but light itself is a complicated signal. It has various characteristics, such as intensity, duration, or wavelength, and one can experience light at different times of day and be exposed to different light histories. Given this complexity in real life, an integrated approach to entrainment*, the circadian response characteristics* (CIRC) has been suggested by Roenneberg and colleagues[47]. The shape and asymmetry of the CIRC predicts the effect of light on the intrinsic period of an organism (Fig. 3A) and is dimensionless. In this view, the internal cycle is either compressed or expanded as depicted by the curve asymmetry, while the dead zones are described by the shape of the curve. In contrast to PRCs or VRCs, the CIRC is assessed from an entrained organism and does not make any a priori assumptions with regards to how internal and external periods are synchronised[47].

Non-photic entrainment

Human rhythms can be studied on the biological, psychological and sociological level since "time cuts across all three domains".[5] Even though light remains the strongest zeitgeber for our clock, other signals, such as food intake[48], physical activity[49], sleep-wake cycles or social cues can potentially also entrain an organism (reviewed in [50,51]). If these signals are out-of-phase with the light-dark cycle, the information can be misleading: which clocks should the organism follow? Historically, without electricity and artificial lighting, the *Social clock* was synchronised both with the *Sun clock* and with our *Biological clock*. Life in a 24/7 society (*e.g.* shift-work) and constant but relatively dim illuminations at the wrong time has challenged this synchrony. Countries that chose an unfortunate time zone (*e.g.* Spain) or the implementation of day light saving time (DST) in the summer months adds to the desynchronization of the social clock with the sun clock. This results in biological clock shifts which can have severe health consequences, especially when experienced long-term (see also chapter 1.3.3. where this will be followed up).

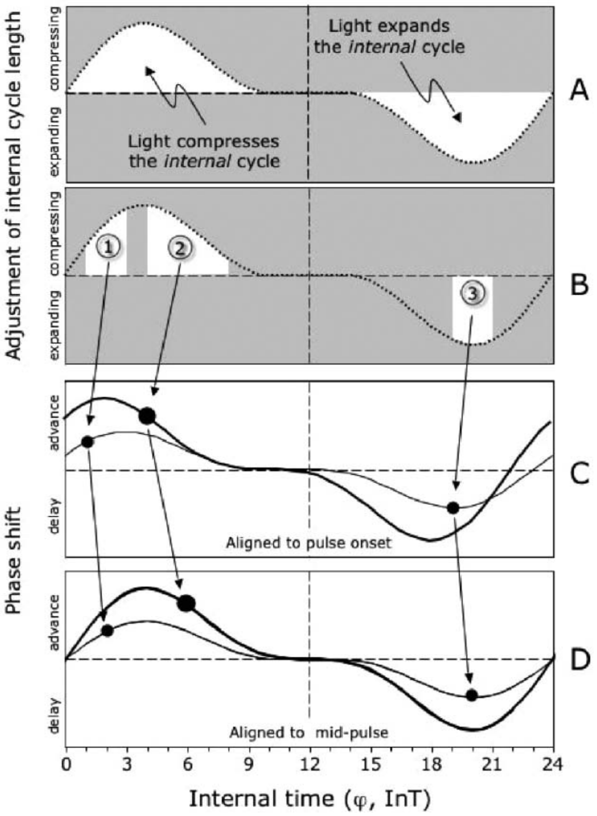

Fig. 3 | The circadian response characteristics. Panel A shows the compression and expansion characteristics of light depending on the internal time. Panel B depicts light pulses given at various points which results in different PRCs in Panel C and D. *Source*: Roenneberg *et al.,* 2010 (reproduced with permission)[47].

1.1.4. Measuring phase of entrainment in humans

Within a certain range of entrainment individuals can entrain differently to their environment, which explains individual variations in physiological or behavioural peaks and troughs. These timing differences are called chronotypes (the famous larks and owls) and inconsistencies between the clocks can affect chronotypes differentially. Since there are many peripheral clocks in the body with no single phase of entrainment, estimating the phase of the complete organism remains difficult.[42] Thus, some proxies of chronotype have proven to be more easily accessible, reliable and thus useful for studying human entrainment than others. Biological phase markers are traditionally core body temperature[52], clock gene expression or hormonal secretion (*e.g.* cortisol or melatonin levels)[53,54]. Especially dim-light melatonin onset is considered a stable and direct phase marker of the SCN and is widely used[55,56]. One limitation all of the physiological phase markers share is their impractical and costly use in very large-scale studies. In these cases, questionnaires can be used to estimate chronotype from subjectively reported sleep timings. Numerous such questionnaires exist, *e.g.* the Composite Scale of Morningness (CSM)[57] or the Morningness-Eveningness Questionnaire (MEQ)[58], with the latter being widely used in the field. These two questionnaires however do not assess sleep separately on school/work days and free days even though there is considerable variance between these two (see also Fig. 7b and 7c). The Munich ChronoType Questionnaire (MCTQ) overcomes this problem since it asks about sleep onset, offset and duration separately for these days[59]. MSF_{sc}, the midsleep on free days (corrected for potential oversleep), which can be computed from these variables[d] is used as a proxy for phase of entrainment of the sleep-wake cycle and was shown to better predict DLMO compared to onset or offset[60].

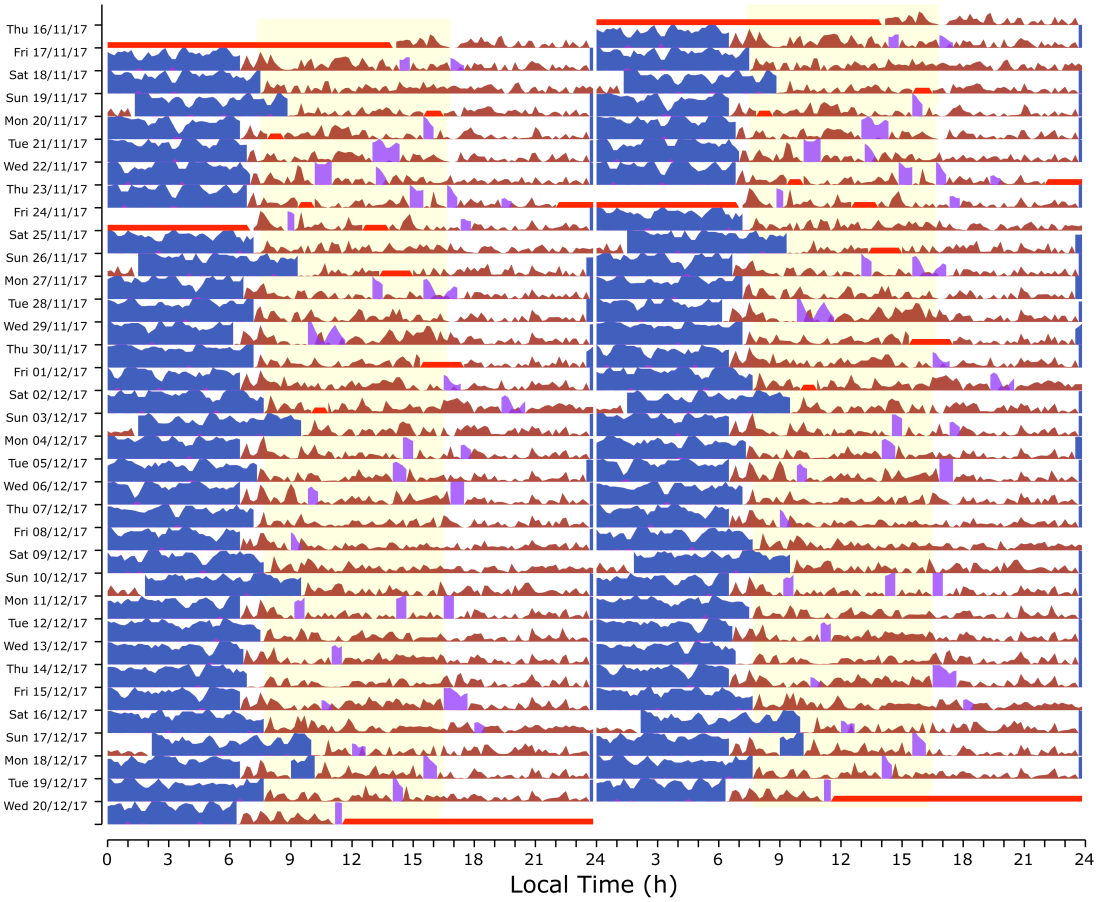

Fig. 4 | Double plot of a participant's sleep-wake rhythm. The graph shows local time on the x-axis, plotted twice in one row (*i.e.* day 1: 0-24h followed by day 2: 0-24h). The next row repeats the previous day and adds the next day afterwards. In this way, sleep episodes are not separated across midnight, instead a more cyclic picture of sleep-wake activity emerges. Blue shows estimated sleep periods, purple indicates day sleep (naps), red curves visualise activity patterns and bright red columns are missing data (device was not worn). The yellow background signals the photoperiod. The activity scale is unitless and simply indicates more or less activity. Note that day and night activity are scaled differently. The algorithm to estimate sleep from activity used here is our in-house Munich Actimetry Sleep Detection Algorithm (MASDA)[61,65]. Clearly visible are delayed sleep episodes from Saturday to Sunday - a typical weekend lie-in pattern. *Source*: own data, unpublished.

Another powerful tool to assess sleep-wake timings and thus chronotype is locomotor activity detected via actimetry. A recording device is worn on the wrist or leg over several days up to several months, and measures changes in acceleration and orientation, which is then converted to a unitless

[d] MSF_{sc} = sleep onset$_{\text{free days}}$ + ½ sleep duration$_{\text{work days}}$

outcome of activity[61]. Several algorithms exist that estimate sleep timing from inactivity, *i.e.* relative rest periods (*e.g.*[62–64]). The in-house algorithm we use is called MASDA (Munich Actimetry Sleep Detection Algorithm) and shows good validity in estimating sleep-wake in field studies[61,65]. When a cosine curve is fitted to the raw activity data (least square approach), numerous variables can be computed such as amplitude, range of oscillation, period, or frequency of phase. Adding a one-harmonic fit to the raw data gives rise to the so-called centre of gravity, the acrophase[61] or peak of the activity rhythm, which can be used as a phase marker for rest-activity cycles of any animal, not just humans. Fig. 4 visualises a participant's sleep-wake cycle over one month. It clearly shows inactive periods, *i.e.* the estimated sleep periods (blue) in a regular fashion. Strictly speaking, this method only allows for a binary distinction between active and rest periods but does not directly estimate true sleep episodes. But how can we even distinguish these states - rest and activity - from one another? Or put differently, what constitutes sleep?

1.2. The mysterious world of sleep

1.2.1. What is sleep?

"If sleep does not serve an absolute vital function,
then it is the biggest mistake the evolutionary process ever made."

(Allen Rechtschaffen)[66]

We all do it. And we even spend approximately one-third of our life in a mostly unconscious state that we call sleep. You might have pondered about how to optimise this extensive period of presumably wasted time. Imagine how much more one could do if we only napped occasionally on the go - just like the swift that sleeps unihemispherically for a short period of time on the wing[67]. But is it actually wasted time? Surprisingly, many fundamental questions about sleep are yet to be conclusively answered. Sleep is a very costly behaviour for many animals because predators have a simple job when you cannot watch out for them. It seems likely that evolution would thus have dropped such a dangerous behaviour as soon as possible. Still, sleep is highly conserved across animal evolution[68] even though Allison and Cichetti demonstrated that the amount and depth of sleep and its temporal distribution indeed largely depends on ecological niches: larger, carnivorous animals that live on the surface (*e.g.* lions, tigers) tend to sleep longer and deeper when not searching for food or mates, while herbivorous species (*e.g.* rabbits), which mostly live in nests, sleep less[69]. In general, sleep architecture in mammals depends on age, body size, diet, where the animal lives and on the safety of its sleeping place.[70] Given that humans still sleep, there is potentially some biological sense to it – or actually, we can afford the luxury of (relatively long periods of) sleep. What is sleep then, how does it work, and why do we need it?

It is relatively easy for us to identify somebody who is falling asleep by observation: breathing rates slow down, muscle tone and activity decreases, eyes are usually closed and often the person exhibits a sleep-specific posture. It becomes increasingly difficult to wake the person because their sensory threshold starts to increase[71]. However, by only watching a sleeping person, we do not observe their brain's activity. Nathaniel Kleitman and Eugene Aserinsky [72] revolutionised the study of sleep in the

1950s and 60s when they first discovered distinct and rhythmic alterations of rapid-eye movement (REM) and non-REM (NREM) sleep in humans' brain activity throughout the night measured with electroencephalogram recordings (EEG)[e]. Previously, it was widely believed that sleep was a rest period characterised by absent or greatly reduced brain activity.[69] Cerebral blood flow[f] studies, however, soon showed only a 20% decrease during sleep[69,73] and a similar share of increasing and decreasing firing neurons at sleep onset[69,74]. In the following years, it became clear that sleep is a highly active period during which the brain exerts several cyclic "programs". In fact, the amplitude of brain metabolism and neuronal activity changes during sleep are mostly higher than during wake[70,75–77]. The fact that the body tries to catch up with (especially NREM) sleep after sleep deprivation or exposure to stressor (*sleep rebound*) also shows that sleep is not just reduced activity or alertness regulated by circadian or ultradian rhythms[70] but that a homeostatic process also plays a role.

1.2.2. A phenomenological model of sleep-wake: the "Two Process Model"

The interplay between this homeostatic component and the circadian rhythm was first characterised by Alexander Borbély[78] and later refined by Borbély, Serge Daan, and Dormien Beersma[79,80]. It is a phenomenological model without considering the underlying physiology in its original presenting (Fig. 5). Despite - or maybe even because of - its simplicity, the *Two Process Model* is still one of the most influential models about the timing of sleep, and allowed many predictions and hypothesis testing[80]. In this model, Process "S", the homeostatic sleep dept, increases during wake thereby increasing the pressure to fall asleep the longer the wake period, and dissipates again with ongoing sleep duration (Fig. 5). Process "C" is the circadian force that promotes sleep and wakefulness in a sinusoidal manner. It counteracts sleep pressure propensity when the sleep window has not opened yet (a few hours before habitual bedtime) and ensures longer sleep duration towards the end of the subjective night when sleep pressure is almost completely vanished. In an entrained person, the circadian rhythm also enables the timely release of melatonin from the pineal a few hours before habitual sleep onset which promotes sleepiness.[81] Overall, the timing of sleep is thus gated by the circadian pacemaker. It has recently been suggested, though, that the *Two Process Model* should be updated to include a social component since our modern 24/7 world influences sleep-wake behaviour to a large extent (see also chapter 1.1.3).[42]

[e] EEG activity patterns emerge from the summed postsynaptic action potentials of pyramidal neurons.[231]

[f] Cerebral blood flow is a biological marker for neuronal activity. Measured as blood-oxygen-level-dependent (BOLD) signal, it shows local changes in blood flow and oxygenation levels, which are altered by neuronal activity via coupling of neurons with the vasculature of the brain.[96,232]

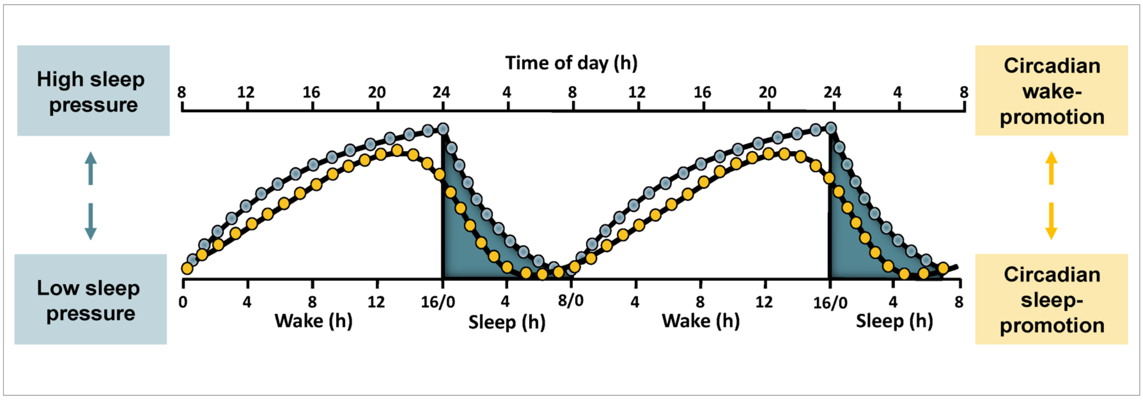

Fig. 5 | The Two Process Model. Schematic of the interplay of process S (the homeostatic component; blue) and process C (the circadian component; yellow) that describes the sleep-wake regulation in humans. The circadian component shows a sinusoidal wave, promotes wake during the subjective day and sleep during the subjective night, and is relatively unaffected by prior sleep-wake history. Sleep need increases with prolonged wakefulness, reaches its peak just before bedtime and dissipates with sleep duration. *Source:* Reichert *et al.,* 2016 (reproduced with permission)[82].

1.2.3. Physiological models of sleep

While the *Two Process Model* served as a very useful concept of wake-sleep regulation, only recently its physiological underpinnings have started to unravel[82]. So, what happens during sleep on a physiological level and how do we fall asleep? To maintain wakefulness and cortical activation, several subcortical structures and neurochemicals are necessary. The *Ascending arousal system* includes excitatory noradrenaline secreted from the locus ceruleus, serotonin from the raphe nuclei, dopamine from the ventral periacqueductal grey matter, acetylcholine from the pedunculopontine tegmentum and the tegmentum of the pons, and lastly orexin that comes from the perifornical area[83]. Since this sounds rather technical, Fig. 6 gives a visual overview of this system and also depicts the sleep-promoting areas that suppress the ascending arousal system mainly by constant inhibition from neurons of the ventrolateral preoptic area (VLPO) via GABA (gamma-aminobutyric acid)[83]. Making the rapid switch from wake to sleep by mutual inhibition is also referred to as the *flip flop switch hypothesis.* It remains subject to further study which molecules and mechanisms initiate the switch in the first place. While accumulating adenosine in the extracellular space of the forebrain during wakefulness seems one important factor (*i.e.* the homeostatic process "S"), other signals must also be involved[83]. The circadian clock (process "C"), which "opens the gate" for sleep, acts directly on the sleep-promoting area since the VLPO also receives input from the SCN, the master clock.

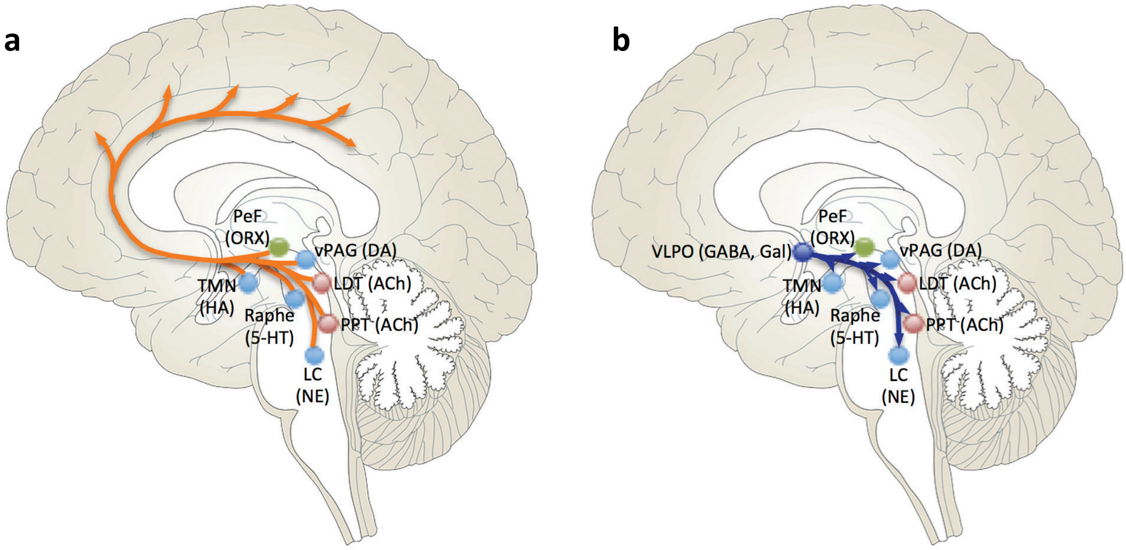

Fig. 6 | Sleep and wake promoting pathways in the human brain. a, key regions and neurochemicals involved in the *ascending arousal system*. **b,** pathways from neurons in the anterior hypothalamus, the VLPO, inhibit the ascending arousal system via GABA thus enabling sleep. Abbreviations: ACh, acetylcholine; DA, dopamine; GABA, gamma amino-butyric acid; Gal, galanin; HA, histamine; LDT, laterodorsal tegmentum; NE, norepinephrine (noradrenaline); ORX, orexin; PeF, perifornical region; PPT, pedunculopontine tegmentum; TMN, tuberomammillary nucleus; vPAG, ventral periaqueductal gray matter; 5-HT, 5-hydroxytryptamine. *Source*: Carley & Farabi, 2016 (reproduced with permission)[83].

When sleep is initiated, REM and NREM sleep take turns in a regular pattern of about 90-min cycles occurring around 4-6 times per night[72]. NREM sleep is further divided into N1, N2, and N3, with the latter also referred to as delta sleep, slow-wave sleep (SWS) or more broadly deep sleep[84]. The cycle runs typically, but not exclusively, in a sequence of N1➝ N2 ➝ N3 ➝ N2 ➝ REM, with greater amounts of N3 sleep earlier at night and more REM sleep towards the end of the sleep episode[*e.g.* 85,86]. N1 sleep is a transitional state characterised by low-amplitude but high-frequency theta activity (4-7 Hz), as seen in EEG recordings, and behaviourally also identifiable through slow rolling eye movements and decreased muscle tone (*e.g.* the typical head dropping when you watch somebody falling asleep). In N2, arousal levels decrease, the sensory threshold thus increases and sleep spindles are observed in the EEG in the 11-16 Hz range. N3 is the deepest sleep stage since the arousal threshold is maximally increased and, in contrast to N1, low-frequency but high-amplitude waves (slow waves) of 0.5-4 Hz dominate the EEG pattern reflecting highly synchronised neuronal activity over widespread cortical areas[85,86].

The switch between NREM and REM sleep is thought to be initiated by reciprocal inhibition of monoaminergic neurons ("REM-off" neurons) and specific cholinergic neurons ("REM-on" neurons) in the brainstem[83,87]. According to the *reciprocal interaction hypothesis* REM-on neurons become highly active when REM sleep starts, thereby inhibiting noradrenergic locus ceruleus neurons and serotonergic raphe neurons[83,88]. REM sleep characteristics differ widely from NREM sleep and seems paradoxical: twitches in the extremities are observed, breathing and heart rates are very variably and the metabolism and neuronal activity are increased compared to NREM - the EEG patterns almost resemble a wake state[70]. Quite intriguingly, REM sleep is accompanied by rapid eye movement (in phasic REM; in tonic REM sleep no rapid eye movements are observed) but a paralysis of the remaining body muscles. Dreams are mostly reported when awaked from this stage and one hypothesis suggests that the paralysis serves as a security mechanism to prevent acting out dreams[89], although it remains questionable why all muscles experience atonia and not just the extremities. Future research also needs to shed light on the ultradian structure of sleep, since its biological and clinical relevance are not clear.[83]

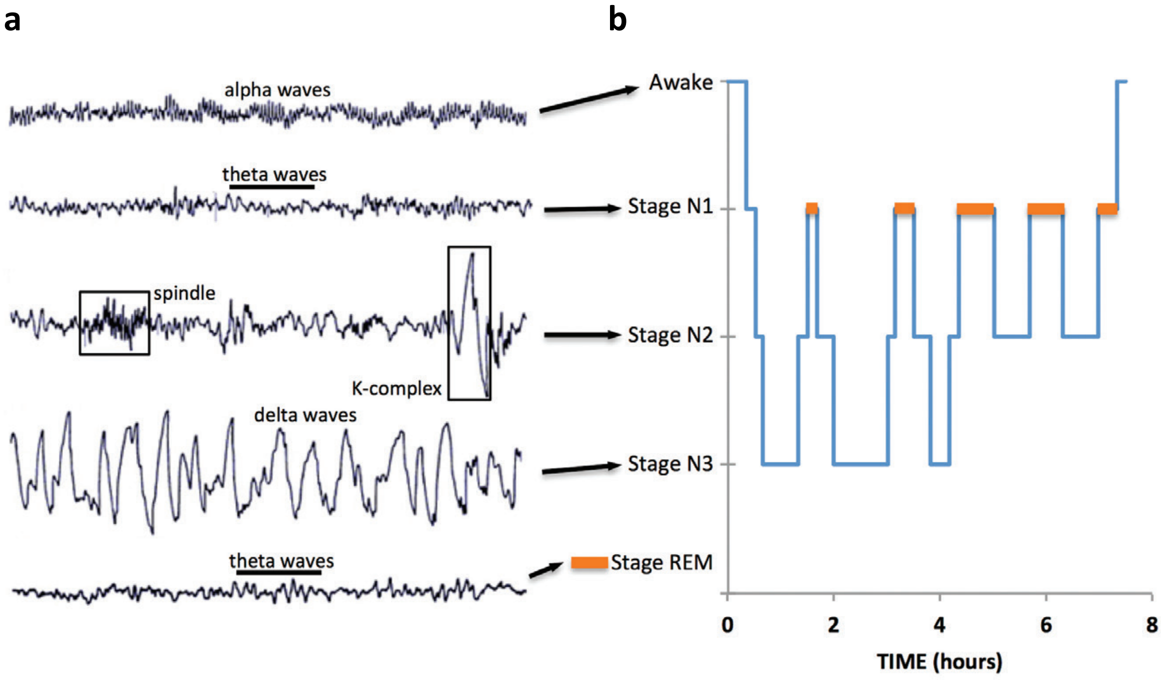

Fig. 6 | Typical EEG characteristics of sleep and wake stages in a healthy human adult. a, EEG activity patterns during wake and during different NREM an REM stages. **b**, A sleep stage graph depicting the temporal organisation of sleep stages across one night. Abbreviations: NREM, non rapid-eye movement sleep; REM, rapid-eye movement sleep. *Source*: Carley & Farabi, 2016 (reproduced with permission)[83].

1.2.4. Do waste your time sleeping: functions of sleep

"It is inevitable that every creature which wakes must also be capable of sleeping,
since it is impossible that it should continue actualizing its powers perpetually.
So, also, it is impossible for any animals to continue always sleeping."
(Aristotle, 350 BCE)[90]

The function of sleep are still not entirely clear and one can choose from a plethora of theories about it, all of which have minor or major drawbacks (as reviewed in [89,91]). While some researchers assume that sleep serves the same function in all animals[*e.g.* 89], this has been widely based on observations from a limited amount of species (mostly terrestrial mammals) and remain to be conclusively shown[70]. Impressively modern, for Aristotle the function of sleep was perfectly obvious: we, and in fact any creature, sleep to save energy. This is indeed one of the most obvious and oldest ideas but has been questioned considerably since the net energy saved seems too marginal to be "sleep's *primary* function"[89,92–94]. Schmidt has recently tried to unify several theories in the *Energy Allocation Model of*

Sleep that states that sleep is based on "the need to optimally allocate limited energy resources to essential biological processes"[89]. In his view, sleep is used for growth, maintenance, repair, immune function and neural network reorganization and allocated to night time to distribute energy uptake across the 24h clock. He does not adequately explain, though, why simple motor quiescence does not do the trick, nor why REM sleep gets longer throughout the night, or changes with age[91].

According to another prominent sleep theory, the *Synaptic Homeostasis Hypothesis*[95], the brain utilises homeostatic downscaling when, at least partly "disconnected", from the environment during sleep to prune up to 20% unnecessary (weak) synaptic connections and strengthen necessary (stronger) connections[96–101]. Wake states, on the contrary, are used for establishing new synaptic connections (*i.e* learning on the synaptic level). But why does the brain need to shut-off to do these downscaling tasks? It is believed that during wake synaptic potentiation eventually saturates, it is more costly to maintain metabolism on a high level, and the brain becomes increasingly less capable of filtering out unnecessary information.[96]

Furthermore, the evidence is strong that sleep improves working memory performance, and fosters learning, and consolidation of new information into long-term memories[102–105]. Specifically, sleep seems to support distinct stages of memory processing, such as encoding, consolidation, retrieval, reconsolidation and integration of information into existing networks[105,106]. Due to the wake-like nature of REM sleep, early research mainly focused on the function of this stage in memory processes (as reviewed in [107]) leading to the *dual process hypothesis*, which essential states that SWS is important for declarative memory (explicit, consciously accessible knowledge) while REM benefits non-declarative memories, such as procedural (implicit learning of motor skills and habits) and emotional memories[86,103,107–110]. More recent evidence puts this simplification into question. While there seems to be sufficient evidence for a crucial role of SWS for declarative memory, procedural memory also depends on sleep spindles occurring during NREM. Emotional memory consolidation has also been demonstrated to not exclusively be linked to REM but also NREM stages, rendering the exact role of REM sleep increasingly mysterious.[86,111]

Taken together, while current theories provide much evidence for specific associations of brain states and functional tasks, they do not, or hardly, provide causal evidence for specific functions of sleep to be absolutely necessary. Some have thus concluded that sleep might serve many functions which vary across animals and are even absent in others[69]. Even though we do not know the single, underlying function that makes sleep necessary (if there is one), it is clear from a wealth of studies that sleep is very import for healthy and optimal performance and psychological wellbeing.

1.3. Changes during adolescence and the challenge of early school

1.3.1. Biological or self-made owls?

"Life is a nightmare that prevents one from sleeping."

(Oscar Wilde)

Adolescence can be a challenging time – not only for the individual but also the people around them. Rebellious times lie ahead until teenagers finally find some sense of self and integrate into society by early adulthood. Puberty is initiated by changes in the brain by around age 10 in girls and 12 in boys[112]. This triggers a variety of astonishing transformations across the entire body, ranging from physical growth, changes in metabolism, physiology and sexual maturation[112]. Inversely, the release of hormones from gonads and adrenals also affects the brain's neuronal structure and functioning, which is associated with cognitive, emotional and motivational changes and re-orientation[112–115]. Especially in early adolescence, functional connectivity between cortical regions strengthens and grey matter in higher-order brain regions experiences thinning and prunes synapses[112,116–118]. Frontal neocortical areas, on the other hand, continue to develop into early adulthood – relatively late considering that these regions are critical for cognition, self-regulation and social behaviours[112]. This late frontal cortical development is also associated with reduced slow-wave sleep (SWS) in these areas in children and adolescents compared to adults, while SWS in general consistently declines across adolescence[106,119–121]. In short, biological and neuronal changes during adolescence provide the basis for a sensitive period of fast-paced learning, which involves self-exploration, discovering autonomy and responsibility, understanding social roles and norms, and increased motivation to gain social status[112].

Teenagers also experience major changes in their sleep-wake timing that are partly, but not exclusively, driven by these underlying biological alterations. Due to progressively later melatonin onsets during puberty, teenagers delay their phase and eventually exhibit later sleep-wake rhythms – they tend to become night owls (Fig. 7a)[122–124]. In addition to these chronotype changes, there is some evidence that sleep pressure accumulates more slowly during adolescence, making teenagers more alert towards the end of their subjective day and therefore less likely to get tired early in the evening[123,125]. Teenagers might also be less responsive to phase-advancing morning light but more sensitive to phase-delaying evening light, which would further exacerbate their late chronotype[126]. Still, more studies are needed to confirm the supposedly altered sleep pressure and light sensitivity in adolescence, whereas the evidence is more conclusive that teenagers really undergo a circadian phase-delay and still have a longer sleep need than the average adult.

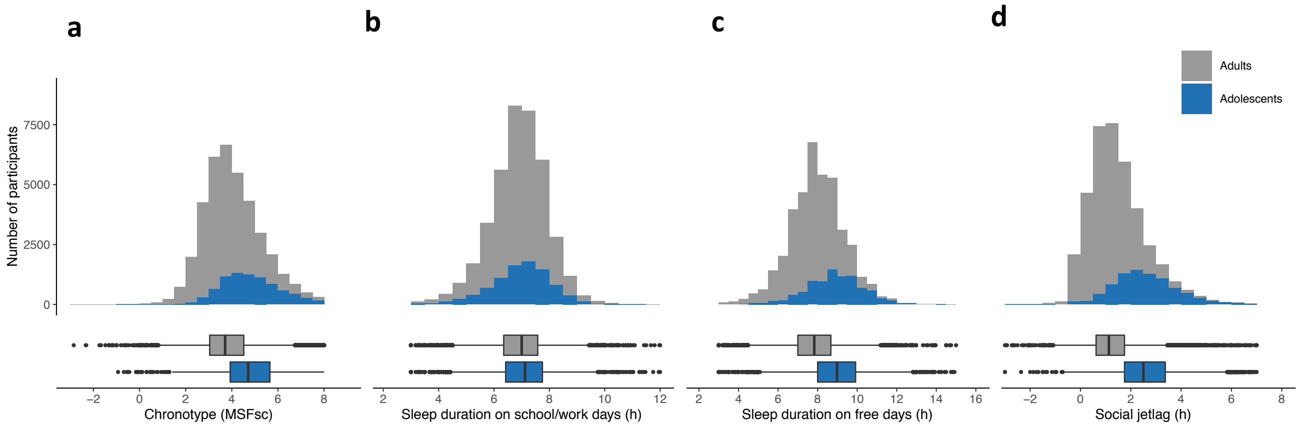

Fig. 7 | Sleep parameters in adolescents compared to adults in Germany. Self-reported sleep parameters from the Munich ChronoType Questionnaire (MCTQ) database from German adolescents aged 14-19 (n=8,388, blue) in comparison to a >3-fold larger German adult sample aged 35-50 (n=29,607, grey). **a,** Chronotype (midsleep time on school-free days corrected for oversleep, MSFsc), **b**, social jetlag, **c**, sleep duration on school/work days, and **d,** sleep duration on free days. Boxplots are Tukey-Boxplots. Sample was drawn in December 2020. *Source:* Roenneberg, private exchange; modified from Winnebeck *et al.*, 2019[127].

This phase-shift, or chronotype delay, has been shown both cross-sectionally over many age groups[*e.g.* 128–132] (Fig. 8) and intra-individually when children were followed over several years.[133,134] It has even been suggested that the tipping point when phase advances again at around age 19 for girls and 21 for boys (Fig. 8) could be used as a biological marker for the end of adolescence[128]. Interestingly, this progressive delay is seen across cultures[129,134,135] and even observed in other mammals at puberty[136]. Cross-cultural comparison, however, indicate that sleep timing is also influenced by social life. A review study by Gradisar and colleagues, for example, showed that Asian populations were especially prone to later bedtimes compared to European and North American study samples[129], possibly due to different social schedules, such as late school times until 20:30h in China or high academic pressure in Korea and Japan, where students often take extra classes or private lessons.[137] Late meal times, which are typical in Uruguay or other cultural similar places like Buenos Aires or Madrid, and attending afternoon school shifts also contribute to late bedtimes and later chronotypes in these populations.[138,139] However, the relative position within a time zone and the difference between the chosen clock time and actual sun time for a specific location also influence chronotype[42,140–142], which might partially confound some of these findings. Indeed, Uruguay, Argentina and Spain adhere to time zones that vary considerable from their sun time.[g] Similarly, Iceland, where later bed and wake up times compared to mainland Europe were also reported[134], adopts UCT (GMT) instead of its designated UCT-1 thus deviating from sun time by >1.5h. The situation is worst in China where all clocks follow Beijing Time that resides at the Eastern border, meaning that cities located at the Western border deviate from the sun clock by up to 5h. This could, at least partly, explain

[g] Spain follows the CET (UCT+1) even though adhering to UCT would align better with sun times by reducing the >1h difference between social and sun times. The same applies to Uruguay and Argentina, which adopted UCT-3 instead of UCT-4.

stereotypical "lateness" in these countries, or between East and Western borders of the same country (as was for example shown in Germany[143]). Type of settlement (urban vs rural) and latitude also influence chronotype[132,141] but all these factors do not explain why Uruguayan teenagers exhibit later chronotypes compared to their Argentinian or Spanish peers[138] since this was comparable in the study by Estevan *et al*[138]. Estevan and colleague showed that school shifts in the afternoon also delayed chronotype by >2.5h compared to morning shift attending students, pointing towards a bi-directional relationship independent of the relative location within the time zone[138]. This further supports that other non-biological or location differences, such as social and cultural factors, also influence chronotype.

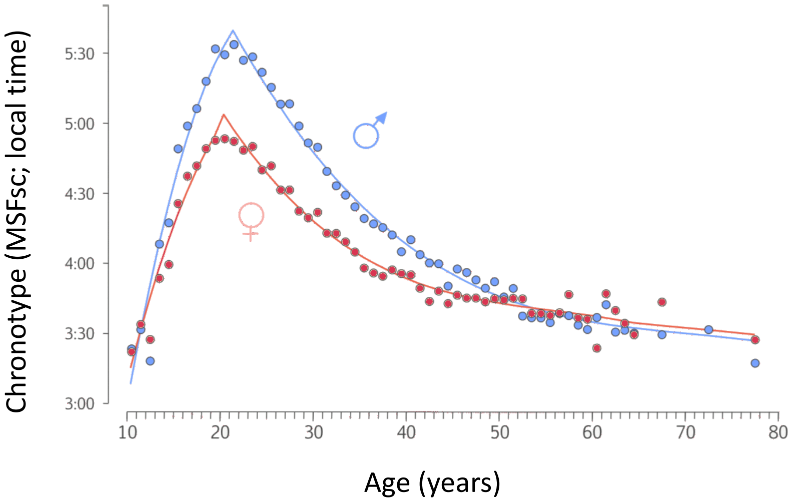

Fig. 8 | Chronotype depends on age and gender. Depicted is the MSFsc distribution (midsleep on free days corrected for oversleep) across age taken from the MCTQ data base. *Source*: Roenneberg *et al.,* 2004 (reproduced with permission).[128]

Indeed, hormonal and neuronal changes during adolescence help to make teenagers more sensitive for cultural and social learning. Teenagers need to take on new social roles and responsibilities. School demands, such as homework and pressure to succeed often increase, part-time jobs become popular, and competitive sports are used for socialising or gaining status[112] – all these factors might extend to late evening hours thus delaying active periods. Very typically, teenagers re-orientate their attention and motivation towards their peers and sexual and romantic interests[112]. This goes hand in hand with more screen time, often at night, which not only increases their exposure to alerting and phase-shifting blue-light at a particular sensitive time of day but might also directly impact on sleep throughout the night (more wake ups from calls, messages, or social media etc)[144,145]. Partying on the weekends or sometimes during the week further delays their circadian rhythms due to late light exposure and food intake, while caffeine intake during the day interferes with their sleepiness. Caffeine acts as an adenosine receptor antagonist that blocks the accumulation of adenosine over time which would normally promote sleepiness.[h] Taken together, teenagers undergo a wealth of physiological and often hormonally but also socially-driven behavioural changes, all of which contribute to their very typical late phase and consequently late sleep timing. Recently, Roenneberg *et al.* thus argued that chronotype might be more state-like instead of a stable personality trait, which enables slight deviations from one's most typical chronotype but not towards the other end of the spectrum (*i.e.* owls do not become larks).[42] While teasing apart the exact biological or social influences on chronotype is relatively irrelevant for the individual – they need to manage and adapt somehow – it might help researchers to develop targeted interventions and guidelines to enable healthy sleep.

1.3.2. The challenge of early school

"*There is no hope for a civilization which starts each day to the sound of an alarm clock*."

(Anonymous)

Due to these alteration during adolescence, one of society's hardest burdens to face as a teenage owl are early school start times[127,129]. These are in direct conflict with teenagers' late internal time[i] and still increased sleep need of at least 8-10 hours compared to adults[146,147]. Let us assume that the typical teenager falls asleep at 23:00h and gets up at 6:30h to reach school in time for 8:00h in Germany. This would result in a net sleep duration of 7.5 hours, which is not just below the recommended minimum of 8h but also relatively optimistic: many students probably fall asleep later and need to get up earlier, especially when buses leave early to reach school in time, which typically starts at around 8:00h or slightly earlier in Germany. Fig. 7b shows that German teenagers indeed only reach on average about 7.5h of sleep on school days, while their sleep duration is markedly increased to around 9h on their free days. Similar numbers were also reported for example in Uruguay, where only 20% of surveyed students reported school day sleep durations >8h, with average sleep durations of 6.5h on school days but 8.2h on weekends.[138] A worldwide comparison from 2011 also showed that in 53% of samples, total sleep time was insufficient on schooldays (*i.e.* <8h), while sleep on weekends was reported to be optimal (>9h) in 71% of samples and no study reported insufficient sleep.[148] The situation in Korea is especially bad, where sleep duration was only 4.9-5.5h respectively in 11-12th graders.[137] While there are variations between countries, this points towards an alarming and general shortening of sleep and schooldays across the globe, with typical longer catch up sleeps on the weekends. This shift in sleep timings between a more biological sleep on free days, and a more socially determined sleep on school or work days is called social jetlag (SJL), since it describes a jetlag caused by social schedules[42,149]. It is computed as the absolute difference between midsleep on free days (MSF) and the midsleep on work/school days (MSW) and measured in hours. If sleep times were kept constant, social jetlag would be equal to 0h. German teenagers, however, live on average with a social jetlag of about 2.5h (Fig. 7d)

[h] It should be noted, however, that caffeine sensitivity depends on the genetic makeup. Polymorphisms of the adenosine A_{2A} receptor gene explain differential efficacy, *i.e.* individual sensitivity to caffeine[233].

[i] There are still chronobiological differences between individuals also during puberty. Not all teenagers become very late, some still exhibit early sleep-wake timings, *i.e.* they tend to stay larks. Still, the evidence shows that on average teenagers tend to delay more or less during puberty, exposing all individuals to sleep restrictions when school starts too early.

that is driven by their early rise times on school days – a typical shift that international comparisons between North and South America, Europe and Asia confirm[134,138,148,150–152]. Importantly, advancing sleep times on week days, as sometimes suggested by parents or educators, would actually aggravate social jetlag, and health problems associated with social jetlag will be described in the next chapter (1.3.3). To reduce social jetlag, some believe that shortening one's sleep on the weekend is thus the answer: this is also not very sensible since it would drastically reduce sleep duration, depriving teenagers from healthy catch up sleep, which they only get on the weekends. Early bell times thus have a severe influence on students' sleep, cutting it short below healthy and adequate amounts for this age group, thus contributing to social jetlag.

1.3.3. Consequences of inadequate sleep and social jetlag

As described in the previous chapter and in 1.1.3, social schedules, such as school times, severely influence sleep durations. But how harmful is it really to sleep short? Shouldn't teenagers be prepared for the tough everyday life of adulthood when sleep-ins are frowned upon or simply impossible? Dahl points out, that adolescence is a decisive time enriched with life-changing opportunities that can be used to promote healthy and constructive (social) development and learning[112]. But adolescence also exposes teens to certain risks and increases their willingness to engage in risky behaviours, some of which involve sleep to a greater or lesser extent. Indeed, a large cohort study provided evidence that mortality rates increase if people with short sleep durations on work days do not catch up on sleep on the weekends, making sleep restriction a very costly behaviour in our society.[153] Others have also pointed out that sleep loss is one of society's most pressing problems[154] and that "unnatural" sleep-wake times could be one of the most prevalent high-risk behaviours in modern society.[155] More specific negative consequences are listed below.

Mental and physical health

The relationship between mental health and sleep is complicated and certainly bi-directional[156,157]. Changes in sleep patterns predict depression in general (as reviewed in[156]), and sleep problems already experienced before the onset of puberty were shown to be predictive of anxiety and depression disorders during adolescent years[158]. Inadequate sleep also decreases the ability to regulate mood and increases thoughts of suicide[159], while the other way around, reduced self-esteem during adolescence is also associated with sleep problems[160]. Indeed, sleep disturbances are now viewed as a core feature of depression, with insomnia and hypersomnia even serving as diagnostic criteria for depression[157,161]. Social jetlag has also been linked to depressive symptoms in a rural population in Brazil [162] and irregular sleep patterns were shown to be a risk factor for depression (as reviewed in [163]).

Furthermore, physical health is also compromised by acute and chronic sleep deprivation and social jetlag. Chronic sleep restriction increases the risk to develop several disorders, such as metabolic, cardiovascular, and inflammatory diseases[164,165]. Social jetlag has been linked to metabolic syndrome and obesity[166] and is positively correlated with heart rate in shift workers[167] that is predictive for cardiovascular diseases.

Risk-behaviours

Teenagers in general start to exhibit high-risk behaviours, such as cigarette or marijuana smoking, (unprotected) sexual activity and alcohol consumption during adolescence[168,169]. Later chronotypes in this sub-population are particularly prone to these risk-behaviours: they are more likely to use drugs, drink alcohol[170] and smoke cigarettes[149] compared to their earlier peers. Often, alcohol and drugs are used to self-medicate depressive symptoms and other psychological problems[171], which might have arisen partly from inadequate sleep leading to a vicious cycle of negative reinforcement. By age 16-18, teenagers are also allowed to actively participate in road traffic, exposing them to motor vehicle accidents when attention is reduced and drowsiness increased[172,173]. Reaction times have been shown to improve with more sleep[174,175] while short sleep severely hampers attention and performance, even though this is not necessarily perceived as drastically by the individual[*e.g.* 176–178]. This is particularly insidious because subjectively we might feel able to drive, but objectively we are certainly not. In fact, sleep deprivation affects the body in ways comparable with alcohol intoxication[179]. Teens, who already show an increased readiness to take risks, and who are more likely to be sleep deprived and suffer from social jetlag, are thus potentially even more prone to car accidents than adults.

Cognition and academics

These attention deficits are also problematic in the educational context. Teens with sleep reduction have more difficulty understanding taught material, especially in the early morning classes, regardless of how much time they spend studying[180]. Tardiness and absences[181] are also more likely with increased sleep deprivation reducing students' presence and participation[182] in class and thus simply the exposure to taught material. Furthermore, many higher-order skills such as emotional intelligence, constructive[183] and creative thinking skills[92,184], cognitive ability[185], and verbal fluency[186,187] are reduced with short sleep[185] – all of which contribute to a good social life, learning and academic performance. Alarmingly, the prefrontal cortex, which is associated with many of these executive and high-order functions, seems to be especially susceptible to sleep deprivation[187–189].

Fig. 9 summarises biological and behavioural changes during adolescence that contribute to acute and chronic sleep loss since they clash with early school starts thus leading to higher risks for a variety of different health conditions.

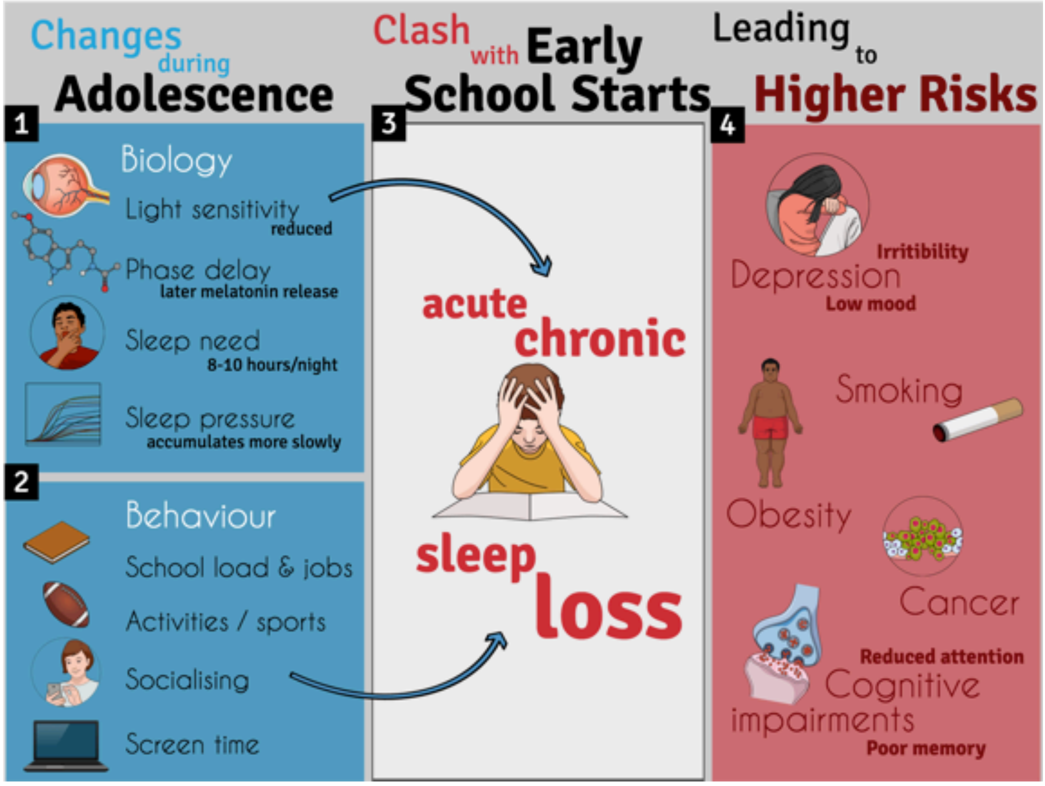

Fig. 9 | **Changes during adolescence clash with social schedules and increase risks for health problems**. There is profound evidence that shows a phase delay and a longer sleep need of 8-10 hours in adolescents compared to adults, while more studies are needed to confirm indicated altered light sensitivity and sleep pressure in teenagers. The altered biology and behaviour during adolescence clash with early school start times leading to acute and chronic sleep restriction in this population. Numerous health risks are associated with acute and chronic sleep loss. Figure created in the Mind the Graph platform (*www.mindthegraph.com*).

1.3.4. Delaying school start times: a viable alternative?

Albeit often unrecognised, sleep deprivation is now widely considered a public health concern[190] and linked to the top 10 leading causes of death in the U.S.[190,191] This has led the American Academy of Sleep Medicine to issue a statement urging for social changes to prioritise sleep to allow optimal functioning and secure health[192]. What could we do as a society to invest in the new generation, to value their altered sleep physiology? It stands to reason that school schedules could be delayed for teenage students to accommodate their delayed circadian rhythms (note that this does not apply to youngsters before puberty) since early bell times are one of the most drastic social alarm clocks teenagers experience. The movement to delay bell times has indeed particularly gained momentum in the U.S., where numerous high-schools have now delayed start times, as well as in South Korea, where a 9 o'clock policy was implemented in 2018 in several provinces. Since then, multiple studies have been conducted to understand whether delaying school start times can resolve the large sleep deprivation of students and maybe even influence other outcomes, such as increased wellbeing, reduced car accidents, and better cognitive and academic performance (tardiness, absences or grades/scores)(as reviewed in [193–195]). While many studies indeed find positive results, several of these unfortunately suffer from methodological drawbacks that limit further generalisability and interpretation (as reviewed in [195]). Most studies that investigated sleep changes, for example, have been conducted cross-sectionally comparing different cohorts of students against each other. This sort of study design does not allow for causal interpretation and is often prone to cohort biases unless covariates are carefully controlled for and participants are randomised, which was not the case in most of these studies. Furthermore, many researchers heavily rely on one-off questionnaires of sleep or other outcomes that are less-sensitive, show low resolutions and are subjective[127]. Especially for performance analyses such as grades or test scores, high-resolution, objective, and intra-individual longitudinal sampling are needed to investigate time trends. However, the evidence is particularly mixed with regards to study designs and results, precluding any clear recommendation for policy makers. Thus, the general evidence concerning school start times changes and associated or causal effects on various outcomes is still weak and warrants further scientific attention.

1.4. Research aims

For my doctoral work, I therefore investigated longitudinal effects of flexible school start times on several outcomes: teenage sleep, subjective psychological functioning and wellbeing, and academic grades. As described in the chapter before, insufficient and partly-mistimed sleep is linked with multiple short and long-term health consequences. Teenager during adolescence, in particular, undergo changes that clash with early school start times in many countries worldwide, exposing them to high levels of sleep reduction[148]. It is thus important to investigate how this health concern can be tackled. In February 2016, a German Gymnasium (the most academic of secondary schools in Germany) changed its school start times (SSTs) for senior students from a mainly 8:00h start to a flexible school start. Students could choose daily whether they would like to begin the first class as usual or delay it to start at 8:50h instead. This opened up a unique and interesting opportunity to study students' sleep in real life in combination with flexi time – a rare opportunity for research since such a system has not been investigated so far. All other studies investigated fixed amount of changes or alternating school shifts.

In **Project 1**, I explain how we followed senior students by means of actimetry and sleep diaries over several weeks in the old SST system during baseline conditions and after flexi time was introduced. We investigated if such a system would allow students to gain an extra hour of sleep, or whether they would simply shift their sleep window, precluding any meaningful sleep improvement and maybe even supporting a sleep worsening. After we observed increased sleep durations in our participants, but only on nights before a delayed school start day, we were curious to know how these sleep changes would manifest themselves after another year since the school maintained the flexible system. In **Project 2,** participants were followed up after exactly one year to again assess their sleep via sleep diaries and also to question them on subjective psychological benefits and wellbeing on early compared to later start days. Here, I also obtained 4 years of quarterly, objective grades of all our participants (provided by the school). This allowed me to tease apart complex interactions between general factors and sleep parameters on students' grades, and to answer another important question: Does a flexible school start system allow students to get better grades? During analyses and writing of Project 2, it became clear that our own and the current evidence concerning altered SSTs and academic performance (grades or test scores) is far from conclusive - a review was warranted. This culminated in **Project 3**, a systematic review of the evidence of altered SSTs on academic performance, including a careful bias assessment of the included research studies since a meta-analysis was not feasible given the heterogenous study designs and outcomes. The result of this review challenges the current opinion held by some researchers within the field that later SSTs definitely lead to improved grades and higher academic performance.

Each manuscript stands for itself with its own Abstract, Introduction, Methods, Results, and References. Appendices that include additional analyses or display items are added at the end of each manuscript. A general discussion (chapter 5) concludes the thesis following the description of the three projects.

2

Project 1

"Later school start times in a flexible system improve teenage sleep"

Winnebeck, E. C., Vuori-Brodowski, M. T., Biller, A. M., Molenda, C., Fischer, D., Zerbini, G., & Roenneberg, T. (2020). Later school start times in a flexible system improve teenage sleep. *Sleep*, *43*(6), zsz307.

Later school start times in a flexible system improve teenage sleep

Authors:

Eva C. Winnebeck[1*], Maria T. Vuori-Brodowski[1], Anna M. Biller[1,2], Carmen Molenda[1], Dorothee Fischer[3,4,5], Giulia Zerbini[1] and Till Roenneberg[1*]

*corresponding author

Affiliations:

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

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

[3] Department of Sleep and Human Factors Research, German Aerospace Center, Cologne, Germany

[4] Division of Sleep Medicine, Brigham and Women's Hospital, Boston, MA, USA

[5] Division of Sleep and Circadian Disorders, Harvard Medical School, Boston, MA, USA

Contact information:

eva.winnebeck@med.uni-muenchen.de

till.roenneberg@med.uni-muenchen.de

Institute for Medical Psychology, Goethestr. 31, 81373 Munich, Germany

Author contributions (CRedIT Taxonomy):

Conceptualisation: TR

Methodology: ECW, MV, LKP, CM, TR

Investigation: ECW, MV, CM, AMB

Data curation: CM, AMB, MV, AKS

Formal analysis: ECW

Validation: MV, DF, CM, AMB, GZ

Supervision: ECW, MV, TR

Visualization: ECW

Writing - original draft: ECW, AMB, GZ, CM

Writing – review and editing: ECW, MV, AMB, CM, GZ, DF, TR

Abstract

Sleep deprivation in teenage students is pervasive and a public-health concern, but evidence is accumulating that delaying school start times may be an effective countermeasure. Most studies so far assessed static changes in schools start time, using cross-sectional comparisons and one-off sleep measures. When a high school in Germany introduced flexible start times for their senior students – allowing them to choose daily between an 8AM or 9AM-start (≥08:50) – we monitored students' sleep longitudinally using subjective and objective measures. Students (10-12th grade, 15-19y) were followed 3 weeks prior and 6 weeks into the flexible system via daily sleep diaries (n=65) and a sub-cohort via continuous wrist-actimetry (n=37). Satisfaction and perceived cognitive outcomes were surveyed at study end. Comparisons between 8AM and ≥9AM-starts within the flexible system demonstrated that students slept 1.1h longer when starting school later – independent of gender, grade, chronotype and frequency of later starts; sleep offsets were delayed but, importantly, onsets remained unchanged. Sleep quality was increased and alarm-driven waking reduced. However, overall sleep duration in the flexible system was not extended compared to baseline – likely because students did not start later frequently enough. Nonetheless, students were highly satisfied with the flexible system and reported cognitive and sleep improvements. Therefore, flexible systems may present a viable alternative for implementing later school starts to improve teenage sleep - if students can be encouraged to use the late-option frequently enough. Flexibility may increase acceptance of school start changes and speculatively even prevent delays in sleep onsets through occasional early starts.

Keywords: *sleep, adolescence, school start time, secondary school*

Significance statement

In many cultures, teenagers are chronically sleep deprived because their typically late sleep times conflict with the relatively early start times of theirs schools. This is a pressing problem since teenage sleep deprivation is linked with reduced performance and substantial long-term health risks. However, the potentially simplest public countermeasure of delaying school starts requires more longitudinal, high-quality evidence. Our study adds important data on later school starts in Europe, using robust longitudinal comparisons and sleep measures, assessing a unique system of flexible start times. Sleep improved substantially and universally on days students opted to start classes at ≥9AM rather than 8AM. Net gains in the flexible system, however, required frequent late starts. Long-term effects of this system are under investigation.

Introduction

Adolescence is a decisive time in life, characterized by important developmental changes that shape individual future trajectories in health, education, social and economic success. A recent review by Dahl and colleagues emphasized the importance of studying these modifications in order to develop policies to support adolescents during such a critical life period[1].

One marked - though often neglected - change during adolescence concerns sleep. Linked with pubertal development, adolescents show a progressive delay in the timing of their sleep until their early 20s when sleep time starts to advance again[e.g. 2–5]. Several biological, environmental, and social reasons have been suggested for explaining the later sleep in adolescents. First of all, the two biological processes regulating sleep – circadian and homeostatic – appear to be altered during adolescence[6]. The circadian system, which promotes wakefulness during the day and sleep at night, shows a later synchronization with the external day compared to children and adults[3,7,8] and thus provides a later circadian sleep window. At the same time, the build-up of sleep pressure appears slower, making adolescents less tired in the evening hours, which further delays their sleep[e.g. 9,10]. This tendency for late sleep (not only on weekends but also throughout the school week) may be increased by external factors such as academic and peer pressure to stay up late studying or socializing online. Concomitantly, adolescents increase their exposure to evening light which results again in later sleep times[11,12] by acutely increasing alertness[13–15] and potently delaying circadian rhythms[13,16,17]. The interplay between all these factors may thus result in a 'vicious cycle of lateness' that exacerbates the natural (biological) tendency of sleeping late during adolescence.

Sleeping late *per se* would not be a problem if school schedules were organized accordingly. However, most schools have early starting times that clash with adolescents' late sleep times. As a result, students accumulate a substantial lack of sleep over the school week[e.g. 18–22]. The consequences for performance and health are evident both in the short and long term. Negative effects of short sleep have been reported, among others, for academic performance[23], absenteeism and tardiness[24], participation and learning in class[25], emotional intelligence and constructive thinking skills[26], and motor vehicle accidents[e.g. 27,28]. Even more worrying are the long-term health consequences of chronic sleep deprivation, such as increased risk for metabolic, cardiovascular and inflammatory diseases[29,30], depressed mood[31–33], and substance use[34,35]. Additionally, students suffer from social jetlag, the mismatch between their circadian clock and their societal schedule[36]. Social jetlag, which is in most instances inherently coupled with sleep deprivation, has been linked with long-term health problems such as obesity and metabolic disorders[37–39].

An obvious solution to the problem of adolescent sleep deprivation is to delay school starting times. Over the last decades, there has been much scientific effort to evaluate the impact of later start times. Most of the studies have been conducted in the US, and they have shown positive outcomes in terms of sleep duration and quality, mood, daytime sleepiness, concentration and attention in class, absenteeism, tardiness, and motor vehicle accidents[40–44]. Still, more studies are required not only in other countries to generalize the results but also to further substantiate the scientific evidence[45]. Given the school setting and research question, study designs are inherently limited and can thus usually not meet highest level evidence criteria such as randomization and double-blind placebo controls. However, so far, the majority of the designs has stopped short of what could be done by using cross-sectional rather than longitudinal comparisons. In addition, outcome parameters (e.g. sleep, mood, academic performance) have often been assessed with just a single-time questionnaire whereas longer monitoring, especially via objective measures such as activity recordings, are rare[42,44,46–50].

We had the opportunity to study the effects of later school starting times when a high school in Germany decided to introduce flexible start times for their senior students. Instead of fixed starts at mostly 8AM, in this new flexible system, the senior students could decide whether to start at 8:00AM or at 8:50AM (referred to as '9AM' herein for convenience) on a daily basis by attending or skipping the first period (a self-study period). We collected daily sleep data via diaries and, in >50% of participants, via objective, continuous activity measures over 9 weeks across systems and across early and late starts. This allowed us i) to compare sleep between alternating early and late school starts within the flexible system in the same individuals without seasonal confounders, and ii) to perform pre-post analyses in the same individuals to assess whether sleep changed from the rigid to the flexible system. This is one of the first studies assessing the effects of delayed school start times conducted in Europe and, to our knowledge, the first to assess the effects of flexible start times[51].

Methods and Materials

Study site

The study was performed at the Gymnasium Alsdorf, a high school in Alsdorf, Germany (50° 53' N, 6° 10' E). Alsdorf is a town of just below 50,000 residents situated in a former coal region in the very West of Germany. A gymnasium is the most academic of several types of secondary schools in the German educational system allowing access to higher education after successful completion. The Gymnasium Alsdorf received the German School Award in 2013 for its innovative teaching[52]. The school operates with a special educational concept called "Dalton plan", which includes daily self-study periods ("Dalton hours") for all students[53,54]. During these self-study periods, students work through their personal 5-week curriculum with a teacher and on a subject of their choice. Each week, students had to fulfill a quota of 10 self-study periods.

School start times at baseline and in the flexible system

In order to address the late sleep times of their adolescent students, the school changed from a conventional school start system with fixed early start times to a new system with flexible start times (flexible system) for their senior students (10th - 12th grade). In the conventional system, senior students started school at times pre-defined by their individual fortnightly schedules. This was usually at 8AM, a typical start time for German high schools, but included a later start on a median of 1 day a week (according to their schedules; cf. Fig. 1, Fig. 6A).

With the introduction of the flexible system on February 1st, 2016, one of the two daily self-study periods was moved into the first period (08:00 - 08:45), and senior students could decide on a daily basis whether to attend this first period or skip it and start school at 08:50 instead (referred to as '9AM' for convenience). Since some students' timetables included days (median ≈ 0.5 d/week) with a free period during the second period, skipping the first period on those days meant a school start at 10:15. Hence, we refer to all later starts as '≥9AM' to include also these cases.

Skipped self-study periods had to be fulfilled at another time during the week in one of the free periods in students' schedules. Although students usually had several free periods per week, there were individual limitations on how often the first self-study period could be skipped without getting home later than individual timetables would otherwise require (see example timetable Table S1). Hardly any timetable allowed making up for 5 skipped self-study periods within its boundaries, however, no student would have had to stay later than the official 4.15 PM end to fulfill the weekly quota.

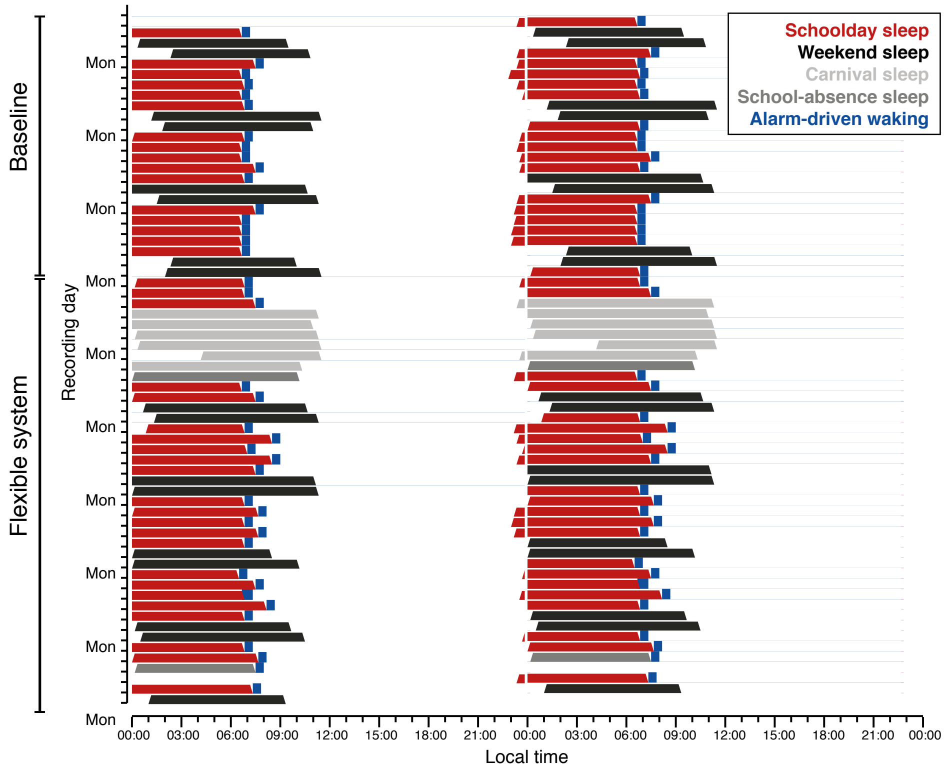

Fig. 1 | Sleep throughout the study period illustrating study design and nature of flexible system. Depicted are sleep-diary-recorded sleep episodes (colored bars) of one participating student over the entire study period. Data are double-plotted. Data on nocturnal sleep episodes were collected over 9 weeks via an online sleep diary and simultaneously via actimeters in >50% of participants. During the first 3 weeks of recording (baseline), students started school at times pre-defined by their individual fortnightly schedules, which was usually at 8AM but included a later start at ≥9AM on around 1 day a week (median across full cohort; see red bars during baseline). Students were then followed 6 weeks into the new flexible system, where they could choose on a daily basis whether to attend the first period at 8AM or start school afterwards at 9AM (08:50) - or occasionally even later on days if and when they had free periods afterwards (see red bars during the flexible system). The holiday period over carnival (light grey bars) was excluded from the analysis.

Study protocol

The recording period lasted from January 8th until March 14th, 2016. We collected daily sleep diary data over 3 weeks before the transition to the flexible system and continued for another 6 weeks after the flexible system was introduced on February 1st, 2016. We also collected objective sleep data via wrist-actimetry throughout the study period in a sub-cohort of students who also filled out daily sleep diaries. For a *status quo* assessment of sleep behavior at the beginning of the study, participants filled out the

Munich ChronoType Questionnaire (MCTQ)[55,56]; at the end of the study period, a purpose-designed survey about the flexible system was also filled out. The holiday period over carnival between February 4th-9th, 2016, was excluded from the analysis.

Participants

We informed all senior students and their parents or guardians via a study leaflet and orally during an information evening. All participants and at least one parent or guardian (when participant was < 18y) had to provide written informed consent. The study was conducted in accordance with the Declaration of Helsinki and approved by the school board, the parent-teacher association and the student association of the school.

We used opportunity sampling without specific exclusion criteria to maximize sample size. Of the 253 students attending 10th- 12th grade (14-19 years) and thus transitioning into the flexible system, 113 (45%) signed up to participate in the study, 93 (82%) students provided at least some data, of which 65 (70%) passed our quantity and quality filter criteria for inclusion in the analysis. These criteria were: i) sleep information for ≥5 schooldays and ≥3 weekend days in each study phase (baseline and flexible system; 27 exclusions); ii) congruent, plausible data (1 exclusion for reported wake-up times that were repeatedly in conflict with reported school start times). The final study cohort of 65 participants was used for all system comparisons. For comparisons between days with an 8AM or ≥9AM start, we additionally required sleep data from at least two 8AM-days and at least two ≥9AM-days per individual to ensure reliable comparisons. After applying this additional filter, a total of 60 participants remained in this sub-cohort. For activity recordings, teachers selected 45 students from all consenting participants who then additionally wore actimeters throughout the study period. After filter application, the actimetry sub-cohort consisted of 34 students also part of the diary cohorts. Cohort characteristics and sample sizes per participant are listed in Table 1.

Out of the 65 students from the full cohort, none reported use of any sleep medication, 3 students (5%) reported to be smokers, 12 (19%) reported weekly alcohol consumption of some sort, and 49 (75%) reported weekly caffeine consumption, with caffeinated drinks as the main caffeine source - not tea or coffee (median of 0.6 drinks/day).

Munich Chronotype Questionnaire (MCTQ)

At the beginning of the study all participants completed the Munich Chronotype Questionnaire (MCTQ) online[55–57]. We used a German version specifically designed for students where all questions pertaining to work were reworked to refer to school, and the formal German "you" (Sie) replaced with the informal "you" (Du)[57]. The MCTQ core module assesses sleep behavior on schooldays and school-free days, and additional modules pose questions about demographics, school times, commute to school, time spent outdoors, and substance use. An estimate of circadian phase of entrainment (chronotype) and a measure of circadian misalignment (social jetlag) are the core variables among the many variables obtainable from the MCTQ (see Data Analysis for formulae). Demographic data were taken from the MCTQ.

By definition, MCTQ-chronotype should only be interpreted if waking on free days is unrestricted, i.e. not alarm-driven; this was not the case for 8 participants in the full cohort and for 4 participants in the sub-cohort 8AM/9AM. To avoid creating additional cohorts, we included these individuals in the analyses but established in sensitivity analyses without these individuals that results were essentially equivalent. For comparisons of sleep behavior between our study cohort and other German adolescents, we randomly drew a 10-fold larger, age- and gender-matched sample of German adolescents from our MCTQ database on August 20th, 2016. Because the study cohort contained 3 additional individuals at that time (n=68 instead of 65; they were later eliminated during a last cleaning round) this database sample contains 680 individuals and not 650.

Tab. 1 | Composition of study cohort and sub-cohorts

		Cohort[a] Diary	**Subcohort**[b] Diary 8AM/≥9AM	**Subcohort**[c] Diary & Actimetry 8AM/≥9AM
Participants				
Total	n	65	60	34
Females	% (n)	62% (40)	63% (38)	65% (22)
Grade (10th/11th/12th)	% (n) per grade	40/35/25% (26/23/16)	42/35/23% (25/21/14)	32/38/29% (11/13/10)
Age (years)	mean (SD, range)	16.5 (1.2, 14-19)	16.5 (1.2, 14-19)	16.7 (1.2, 14-19)
BMI	mean (SD, range)	21.7 (2.9, 16.9-28.9)	21.6 (2.9, 16.9-28.9)	22.2 (3.0, 17.4-28.9)
Chronotype[d] (local time)	mean (SD, range)	5.0 (1.0, 2.7-8.1)	5.0 (1.0, 2.7-8.1)	4.9 (0.88, 3.0-6.6)
Number of sleep diary entries per participant				
Baseline				
Days total (max. 24)	median (IQR, range)	21 (20-23, 10-24)	21 (20-23, 10-24)	22 (21-23, 15-24)
Schooldays (max. 16)	median (IQR, range)	14 (13-15, 6-16)	14 (13-15, 6-16)	14 (13-15, 8-16)
Weekend days (max. 8 + absences)	median (IQR, range)	8 (7-8, 3-8)	8 (7-8, 4-8)	8 (7-8, 5-8)
Flexible system				
Days total (max. 37)	median (IQR, range)	30 (26-33, 9-37)	30 (26-33, 14-37)	32 (27-34, 16-37)
Schooldays (max. 27)	median (IQR, range)	20 (16-22, 6-27)	20 (17-22, 9-27)	21 (19-23, 10-27)
Weekend days (max. 10 + absences)	median (IQR, range)	10 (8-11, 3-15)	10 (8-11, 4-15)	10 (9-11, 4-15)
8AM-days	median (IQR, range)	11 (8-16, 1-23)	11 (8-15, 2-21)	11 (8-15, 2-20)
≥9AM-days	median (IQR, range)	7 (3-11, 0-19)	7 (4-11, 2-19)	9 (5-13, 2-19)

[a] Complete cohort (≥5 schooldays and ≥3 weekend days both at baseline and in flexible system).
[b] Subcohort for 8AM/≥9AM comparisons (additionally ≥2 days per start time in flexible system).
[c] Subcohort for diary/actimetry comparisons (above filters also applied to actimetry data).
[d] MSF_{sc} from MCTQ.

Sleep diary

To obtain daily records of participants' nocturnal sleep, we used a short online sleep diary based on the µMCTQ (a short version of the MCTQ[58] adapted for a German student population. Students were asked to fill it out each morning throughout the study reporting on their past night's sleep. We sent reminder messages around twice a week. If students had missed to fill out the online diary for one or more instances, they were allowed to input their data at a later time point – in most of these instances, students reported to keep an offline log from which they then retrospectively populated the online diary. The sleep diary was provided via LimeSurvey.org. For further details on the diary itself and the data cleaning procedure, please refer to the extended methods in the SI.

Locomotor activity recording (actimetry)

Locomotor activity was recorded continuously over the entire study period in a sub-cohort of 45 participating students via wrist-worn activity-monitoring devices (Daqtometer, version 1.4, Daqtix, Germany). The data analysis pipeline via our in-house analysis program ChronoSapiens[59] entailed averaging activity counts per 30 s into 10-minute-bins, excluding likely off-wrist periods (identified as stretches of 100 min of zero activity or as indicated in actimetry logs) and extracting estimated sleep bouts based on the identification of stretches of relative immobility as detailed in Roenneberg et al. 2015.[59] To allow for sensible comparisons with diary recorded nocturnal sleep, daytime naps (any sleep occurring outside the daily 12-h-trough estimated via cosine fits[59]) were excluded, and bouts <180 min apart were combined into one longer bout. Please refer to the SI for more details.

Final survey

We developed a 12-item self-assessment questionnaire to obtain additional information about the individual use of and satisfaction with the flexible system and the perceived cognitive outcomes. This survey was completed by 56 of the full cohort of 65 students and anonymously by another 82 senior students in the flexible system to assess any selection bias. The participants received the paper-pencil survey in German on the last day of data collection and completed it immediately.

The first 6 items examined the use of the flexible system. The students were asked to indicate i) whether they were satisfied with the new system (yes/no), ii) whether it was difficult for them to start school at 8AM (never/mostly/always), iii) whether it was easier to start school at 9AM compared to 8AM (never/mostly/always), iv) how often (0 days/1-2 days/3-4 days/5 days) and v) on which days of the week they attended the first period at 8AM (Mo/Tu/We/Th/Fr), and vii) reasons for starting school at 8AM. Here, they were given the possibility to state their own reasons or cross at least one of 8 alternatives (easier to study/easier to get to school/additional study time/friends/specific self-study teacher/specific subject/fulfill self-study quota/other).

The final 6 items assessed the behavior and feeling of the students during the baseline and during the flexible system. The first item asked about sleep duration in hours and the second about alarm-driven waking (0-5 days). The last 4 items assessed the quality of sleep, how tired the students felt, ability to concentrate in class, and ability to study at home after school. Each item was scored on a Five-point Likert scale (1 = "bad/poor" to 5 = "good").

Data Analysis

Analyses and visualization were performed in SPSS Statistics (IBM, version 24 and 25) and R[60] (versions 3.5.1 “Feather Spray” and 3.5.3 “Great Truth”) using the R packages effsize,[61] ggplot2,[62] ggpubr,[63] Hmisc[64], lmer4[65], lmerTest[66], PMCMRplus[67], RColorBrewer[68], and reshape2[69].

Data aggregation

For analyses, time course data were aggregated via mean (median for the ordinal variable *sleep quality rating*) to one data point per individual for the 6 conditions of interest. These conditions were i) baseline schooldays, ii) baseline weekends, iii) flexible system schooldays, iv) flexible system weekends, v) flexible system 8AM-days, and vi) flexible system ≥9AM-days. Over the carnival holidays during the flexible system (Feb 5th-9th, 2016), students’ diary compliance was reduced. The remaining entries indicated more irregular sleep, delayed sleep timing and daytime sleep. To minimize any influence on results, we excluded the carnival period from the free-day-aggregates, which are based on fewer data points and can thus be more easily distorted by outliers (Table 1). However, we included the schoolday sleep following the holidays in the schoolday-aggregate measures (as examples of schoolday sleep after a party weekend), where potential outliers are balanced out by more data points (Table 1).

Derived data

From the aggregated measures, the following variables were calculated as per the equations below: average daily sleep duration across the week (SD_{week}); midsleep on schooldays (MSW); midsleep on school-free days (MSF); chronotype as MSF corrected for oversleep (MSF_{sc}); social jetlag (SJL); difference and ratio between ≥9AM-days and 8AM-days for variables of interest (DELTA x; RATIO x, respectively); frequency of ≥9AM-starts (also referred to as 9AM-use) and of alarm-driven waking.

$$SD_{week} = (SD_{schooldays} * 5 + SD_{free\text{-}days} * 2)/7$$

$$MSW = SleepOnset_{schooldays} + \tfrac{1}{2}SD_{schooldays}$$

$$MSF = SleepOnset_{free\text{-}days} + \tfrac{1}{2}SD_{free\text{-}days}$$

$$MSF_{sc} = SleepOnset_{free\text{-}days} + \tfrac{1}{2}SD_{week}$$

$$SJL = MSF - MSW$$

$$DELTA\ x = x_{9AM\text{-}days} - x_{8AM\text{-}days}$$

$$RATIO\ x = x_{9AM\text{-}days}/x_{8AM\text{-}days}$$

$$frequency\ of\ 9AM\text{-}starts = (n_{9AM\text{-}starts_{flex}}/n_{schoolday\text{-}entries_{flex}}) * 100$$

$$frequency\ of\ alarm\text{-}driven\ waking = (n_{alarm\text{-}driven\ waking_{flex}}/n_{schoolday\text{-}entries_{flex}}) * 100$$

Statistical analysis

Data analysis was in part hypothesis driven (comparisons between 8AM/≥9AM-days and between systems) and in part exploratory (analysis of benefit and 9AM-use) to identify important unpredicted patterns. All statistical tests were evaluated to a significance level of $\alpha<0.05$ based on two-sided tests. We used parametric tests for all analyses unless data was below interval level or Shapiro-Wilk test indicated non-normal distribution of a variable in at least one group.

Unfortunately, we could not combine analyses of baseline/flexible system and 8AM/≥9AM-starts, since we lacked reliable information on exact school start times during baseline for each day and participant. We had not asked students about their daily school start time during the baseline period, and students did not follow their timetables exactly (due to teacher absences, exams, etc.; identified via clear mismatches between timetable information and reported wake-up times). Hence, we performed separate analyses as detailed below.

For comparison of sleep parameters between 8AM and ≥9AM-days in the flexible system, we performed paired t-tests or Wilcoxon signed rank tests. Effect size was subsequently estimated using either Cohen's d after paired t-tests via the R package effsize[61] or using the procedure described by Rosenthal[70] for Wilcoxon signed rank tests ($r=Z/sqrt(N_{observations})$; $N_{observation}$ was 2*cohort size as data was paired).

For comparison of sleep parameters between baseline and the flexible system, we used two approaches. For variables present for both schooldays and weekends, we performed 2-factorial repeated-measures ANOVAs with system (baseline/flexible system) and weekday (schoolday/weekend) as main effects. When interaction effects system*weekday were statistically significant, we performed *post-hoc* pairwise comparisons via *t*-tests testing for differences between baseline and flexible system. With two t-tests performed per variable, we corrected p-values via the Bonferroni method by multiplication by 2 to control the family-wise error rate. For variables incorporating information from both schooldays and weekends (social jetlag and daily mean sleep duration across week), we performed paired t-tests or Wilcoxon signed rank tests as described above.

Frequency of alarm-driven waking was additionally analyzed via logistic regression because of the large ceiling effect in this variable (cf. Fig. 2H, 4D). To this end, frequency of alarm-driven waking was dichotomized into high and low frequency of alarm-driven waking based on a median split: at 100% alarm-driven waking for 8AM vs. ≥9AM-days; at 93% for baseline vs. flexible system. Results were equivalent in their direction and statistical significance when using two other splits: i) split at 1st quartile (85% alarm-driven waking) ii) discontinuous split below 1st quartile versus 100% (Table S2 and S3). Logistic regression was performed via mixed effects models using the R package lme4[65] to accommodate the repeated measures nature of the data by including ID as a random effect. In the models reported here, we also included gender as covariate, since there was an obvious trend that males were woken more often by an alarm than females in the flexible system. However, neither exclusion of the covariate gender nor inclusion of additional covariates such as age, chronotype (MCTQ-MSF_{sc}) or 9AM-use altered the effect of school start time or school system in a notable way. Also, gender never reached statistical significance at $p<0.05$ in any of the models.

For the exploratory analysis of characteristics associated with a benefit (sleep extension) and 9AM-use, we analyzed data via Pearson or Spearman correlations for continuous variables as well as via unpaired t-tests, Wilcoxon rank sum tests, 1-way ANOVA or Kruskal-Wallis tests for group comparisons.

The correspondence between diary-recorded and actimetry-determined sleep was assessed via Pearson correlations for average sleep onsets or sleep offsets per person and the 5 relevant, non-overlapping conditions (see data aggregation) leaving out the 6th condition "flexible-system schooldays". Differences between the full study cohort (n=65) and the age- and gender-matched MCTQ-database sample were assessed via Wilcoxon rank sum tests.

Results of statistical tests are reported in the main text in brackets, listing the specific test statistic, the p value and, if applicable, the effect size. Where results across similar variables with similar outcomes are provided in the same bracket, we listed the ranges of the above values across variables. The tests statistics indicate the following tests: *t*, t-test; W, Wilcoxon rank sum test; Z, Wilcoxon signed rank test; r, Pearson correlation; rho, Spearman correlation; H, Kruskal-Wallis test.

Results

Study cohort

The total study cohort, after exclusions based on minimum quantity and quality criteria for sleep diary entries, comprised 65 adolescent students aged 14-19 y covering all three school grades that transitioned to the flexible system (Table 1). The median record length per participant was 21 nocturnal sleep episodes in baseline and 30 episodes in the flexible system. Depending on the study question, we also used two sub-cohorts for analyses, both of which were very similar in their characteristics to the main cohort (Table 1).

Sleep of study cohort is similar to that of other German adolescents

To determine how representative the sleep of our participating students was of other German adolescents, we compared key sleep parameters of the study cohort, assessed via the MCTQ at the beginning of the study, to a ~10-fold larger, age- and gender-matched German sample (n=680) from our large MCTQ database. Study participants were indistinguishable from the larger database sample in any of the analyzed parameters. Namely, sleep duration on schooldays and school-free days, chronotype (midsleep on free days; MSF_{sc}) and social jetlag appeared the same (range of W: 19052-24558; range of p: 0.066-0.2592; Fig. S1). Furthermore, study participants also displayed the gender difference in MCTQ-derived chronotype common for this age group[2], with chronotype on average 1.1 h later in male than female participants (t(48.0)=4.628; p<0.0001; d=1.202), altogether indicating that our sample shows sleep behavior typical for German adolescents – late sleep timing, short sleep on schooldays, long sleep on school-free days, and high social jetlag.

Self-reported sleep times match objective sleep data

Based on data from the sub-cohort of participants wearing actimeters and filling out sleep diaries simultaneously (n=34), we found that subjective, self-reported sleep times matched well with objective sleep times determined from the actimetry records. Average sleep onsets and offsets from both measures were highly correlated (r=0.91 and r=0.94, p<0.0001) and also essentially equivalent (Fig. S2), indicating that the cohort faithfully reported sleep times. Based on this validation, we opted for an analysis of the larger cohort with sleep diary data rather than focusing on the smaller actimetry sub-cohort.

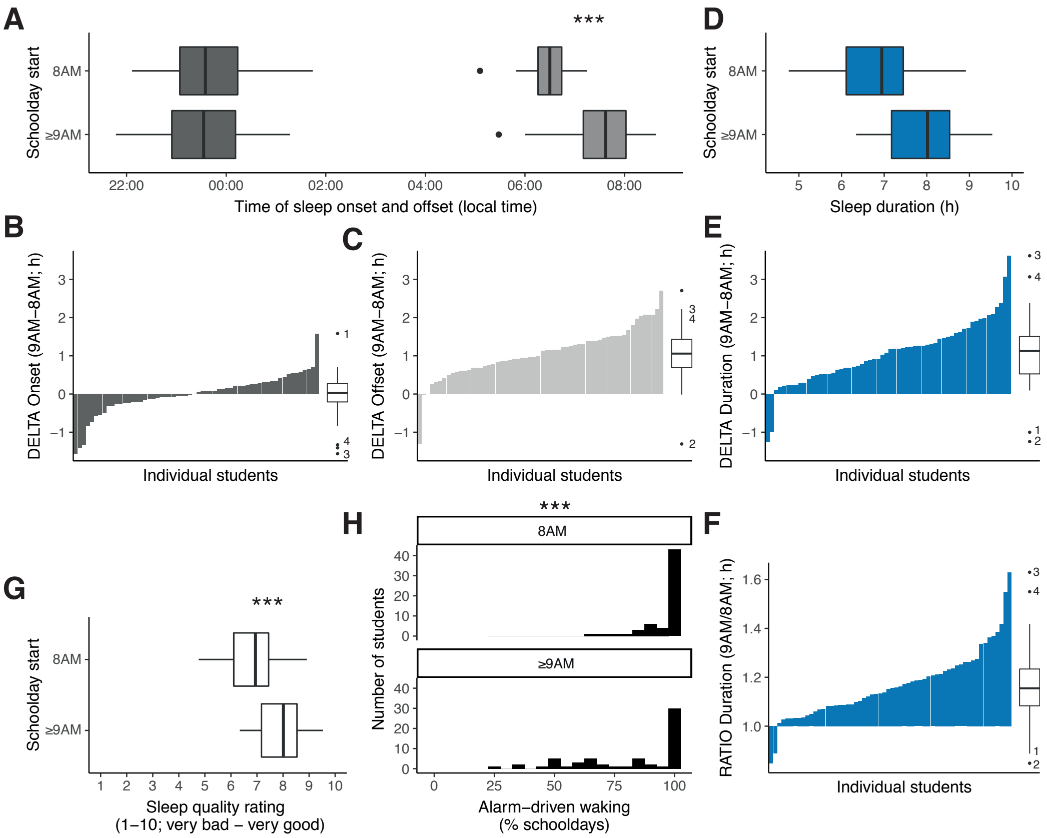

Fig. 2 | **Comparison of sleep parameters between 8AM-days and ≥9AM-days in the flexible system.** Sleep parameters are from sleep diaries of the sub-cohort for 8AM/9AM-comparison (n=60). A) Average sleep onset (dark grey) and offset (light grey) times on 8AM and ≥9AM-days. The average absolute difference in these measures for each individual is depicted in B) for sleep onset times (DELTA Onset) and in C) for sleep offset times (DELTA Offset). Numbers 1-4 identify the 4 over- and under-benefitting students. D) Average sleep duration on 8AM and ≥9AM-days. Each individual's average difference in sleep duration is depicted in E) in absolute terms (DELTA Duration) and in F) in relative terms (RATIO Duration). G) Average sleep quality rating. H) Distributions of individuals' frequency of alarm-driven waking on schooldays. Statistical analysis was performed via paired t-tests or Wilcoxon signed rank tests with *, p<0.05; **, p<0.01; ***, p<0.001; Tukey boxplots. Numbers 1-4 identify the 4 over- and under-benefitting students as per DELTA Duration in E.

Sleep on 8AM versus ≥9AM-days in the flexible system

This section presents our analyses of students' sleep *within* the flexible system. Here, we compared average sleep on nights before a normal 8AM school start (8AM-days) to that on nights when students took advantage of the new option to skip the first period and started school at 9AM - or occasionally even later if they had additional free period(s) in their individual timetables afterwards (≥9AM-days). For simplicity, we henceforth speak of "sleep *on* 8AM-days" or "*on* ≥9AM-days" and mean this to be the nocturnal sleep episodes preceding days with an 8AM or ≥9AM-school start.

Frequency of ≥9AM-starts

Students varied substantially in their use of the 9AM-option, which ranged from 0% to 90% of a student's recorded schooldays (cf. Fig. 6). The median frequency of ≥9AM-starts was 39% (IQR: 20-60%), which amounts to 2 days out of a 5-day school week. For the following analyses, only students that used both the 8AM-option and the 9AM-option at least twice were included (n=60, sub-cohort 8AM/9AM, Table 1).

Sleep onset, offset and duration

As expected, on ≥9AM-days, students woke later than on 8AM-days (t(59)=-13.017; p<0.0001; d=1.68; Fig. 2A). The mean difference in their sleep offset was 1.1 h (SD: 0.64h) - a larger difference than anticipated for a 50-minute delay in school start. There are two likely and additive reasons for this large effect : i) Almost every single student delayed his/her sleep offset time on ≥9AM-days (DELTA Offset >0 h, Fig. 2C). ii) Several students had an additional free period after the skipped first period on several of their ≥9AM-days, allowing them to delay their wake-up times far beyond the expected 50-min difference (DELTA Offset >0.83 h, Fig. 2C). Both communication with the school and our retrospective checks of students' timetables confirmed that this was the case.

Importantly, despite their later sleep offset times, students did not systematically delay their sleep onset times on ≥9AM-days (t(59)=0.0259; p=0.9794; d=0.003; Fig. 2A), illustrated by an even number of students falling asleep either slightly earlier and later on ≥9AM-days compared to 8AM-days (DELTA Onset, Fig. 2B).

Given these stable sleep onsets and markedly delayed offsets, sleep duration was longer on ≥9AM-days than on 8AM-days (Z=6.27, p<0.0001, r=0.57; Fig. 2D). Students extended their sleep by 1.1 h (median; IQR: 0.53-1.5 h) or 15% (median; IQR: 8-23%) from a median of 6.9 h to 8.0 h on ≥9AM days (Fig. 2 D,E,F). Again, the great magnitude of the effect likely results from the 9AM-option sometimes representing a ≥9AM-option as well as almost all students extending their sleep on ≥9AM-days (Fig. 2 E,F).

Subjective sleep quality

On ≥9AM-days, students rated their sleep quality higher than on 8AM-days (Z=-4.435, p<0.0001, r=0.40; Fig. 2G). The median increase was 0.8 points (IQR: 0-1.6) on a 10-point rating scale.

Alarm-driven waking

The proportion of schooldays on which students indicated "woken by alarm clock" was substantial: all students were woken by their alarm more than once a week, and half of the students reported alarm-driven waking on all of their schooldays on both 8AM and ≥9AM-days (Fig. 2H; median in both conditions = 100% of schooldays). Because of this marked ceiling effect in alarm-driven waking, analyses may be less reliable, so we used not only a non-parametric test but also logistic regression to assess potential differences between 8AM and ≥9AM-days. Both analyses indicated that, although the rate of alarm-driven waking was still high on ≥9AM-days, students were woken less often by their alarm than on 8AM-days (Z=4.55, p<0.0001, r=0.42), and the odds for less alarm-driven waking (<100% of schooldays, i.e. alarm-free waking on several schooldays) were increased on ≥9AM-days (OR=3.3; 95% CI =1.28-8.48; Table S2).

Extension of sleep on ≥9AM-days was independent of gender, grade, chronotype and frequency of ≥9AM-starts

To understand which type of student may particularly benefit from later starts, we searched for factors linked with sleep extension on ≥9AM-days, which we considered the core measurable benefit in our study. Sleep extension was quantified as each student's difference in sleep duration between their ≥9AM and 8AM-days, in either absolute terms (DELTA sleep duration 9AM-8AM; Fig. 2E) or relative terms (RATIO sleep duration 9AM/8AM; Fig. 2F). Below, only the result for absolute sleep extension are presented since results for relative sleep extension were essentially equivalent.

The amount of sleep extension on ≥9AM-days showed no systematic relationships with any of the 'key suspects' that we assessed. There was no evidence that genders benefitted differently (t(57.3)=-0.2109; p=0.8337; d=-0.0711; Fig. 3A) or that students from a certain grade (implicitly incorporating the factor age) benefitted more or less (H(2)=2.6445; p-value = 0.2665, Fig. 3B). Notably, also chronotype (either MCTQ or sleep-diary-derived MSF_{sc} at baseline or flexible system) was not associated with the amount of sleep extension (range of r: -0.22 - -0.06; range of p: 0.0845-0.6234; Fig. 3C).

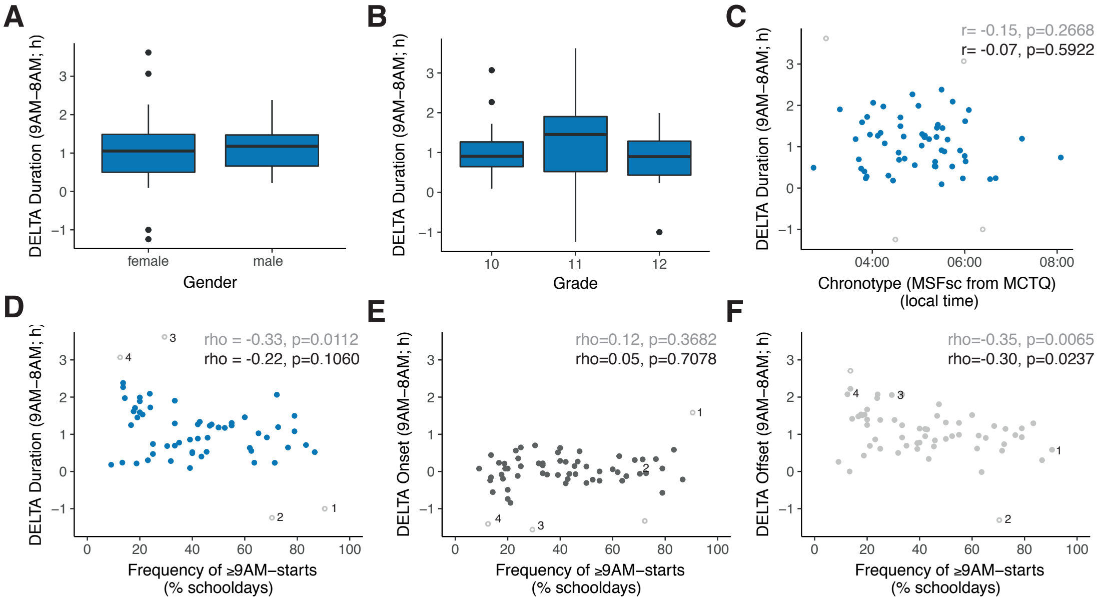

Fig. 3 | Extension of sleep on ≥9AM-days in the flexible system appears independent of gender, grade, chronotype and frequency of ≥9AM-starts. Depicted are the absolute differences (DELTA values) in sleep parameters between ≥9AM and 8AM-days in the flexible system and their relationship to other variables. Sleep parameters are from sleep diaries of the sub-cohort for 8AM/9AM-comparisons (n=60). A,B,C) show the difference in sleep duration between ≥9AM and 8AM-days (sleep extension) against A) gender, B) grade, C) chronotype (midsleep on school-free days corrected for oversleep). D,E,F) show the relationship between frequency of ≥9AM-starts (percentage of schooldays that a student started school at ≥9AM) and the difference between ≥9AM and 8AM-days in D) sleep duration (DELTA Duration = sleep extension), E) in sleep onset (DELTA Onset) and F) in sleep offset (DELTA Offset). Data are color-coded as in Fig. 2 and numbers 1-4 identify the same 4 over- and under-benefitting students. Tukey outliers in the y-axis variable are marked by grey empty circles. Results of Pearson and Spearman correlations are given for data both including outliers (grey) and excluding outliers (black). Statistical analysis for A was via unpaired t-test and for B via Kruskal-Wallis test.

The apparent lack of influence of any of the above factors tallies with the fact that the benefit from ≥9AM-starts was close to universal: virtually all participating students (97%, 58 out of 60) slept longer on ≥9AM-days than on 8AM-days (Fig. 2E). There were only two students that did not benefit (DELTA sleep duration <0 h; outliers 1 and 2 in Fig. 2E), contrasting with two students who benefitted over-proportionally (DELTA sleep duration >3 h; outliers 3 and 4 in Fig. 2E).

What stood out for these negative and positive outliers in sleep extension (Fig. 2E) was that they were at opposite ends in their 9AM-use: the two over-benefiters rarely made use of the 9AM-option whereas the two non-benefiters started quite often at ≥9AM (Fig. 3D). This could have indicated that going more often at ≥9AM reduces the benefit from late starts – a potentially central problem invalidating the flexible system. Indeed, at first sight, this was supported by a negative correlation between 9AM-use and DELTA sleep duration across all students (rho=-0.33, p= 0.0112; Fig. 3D). However, this association was only driven by exactly these 4 outliers. When excluding these from the analysis, the amount of benefit is not associated with the frequency of ≥9AM-starts anymore (rho=-0.22, p=0.1060; Fig. 3D). Furthermore, the most likely mechanism for a smaller sleep extension with greater 9AM-use would be a delay in sleep onsets on ≥9AM-days. However, there is no hint that students with greater 9AM-use had relatively delayed sleep onsets on ≥9AM-days since DELTA onset was not correlated with the frequency of ≥9AM-starts (rho=0.05, p=0.7078; Fig. 3E). However, DELTA offset shows such a correlation (rho=-0.30, p=0.0237; Fig. 3F): students with the least 9AM-use had the greatest delay in offsets. This cross-check shows clearly that a substantial proportion of students with low 9AM-use benefitted over-proportionally through delaying their offsets far beyond the 50-min-extension – likely by starting school much later than 9AM on the few days that they skipped the first period – in contrast to the high users who regularly went at truly 9AM. Hence, the data provide no indication that the frequency of later starts systematically affected the benefit.

Sleep in the flexible system versus baseline with fixed start times

Surprisingly, the switch to the flexible system did not markedly improve students' sleep: Most sleep parameters in the flexible system were not or only minimally different from those reported during the baseline period with fixed school starting times (Fig. 4).

Sleep onset, offset and duration

At first glance, the results are perfectly in line with the positive expectations elicited by the above results comparing sleep on 8AM and ≥9AM-days in the flexible system. Average sleep onset times on schooldays were the same between baseline and flexible system ($t(64) = -0.764$; $p_{bonf} = 0.8956$; post-hoc test to 2-way ANOVA as reported in Fig. 4A), whereas sleep offset times on schooldays were delayed in the flexible system ($t(64) = 2.496$; $p_{bonf} = 0.0303$; post-hoc test, Fig. 4A). However, this delay sports only a small statistical effect size (d=-0.205) and is small also in biological terms at only 6 min (SD: 24 min). Accordingly, average sleep duration on schooldays was indistinguishable between both school start systems ($t(64) = -1.100$; $p_{bonf} = 0.551$; post-hoc test), just like weekend sleep duration and the daily mean duration across the entire week (all $p_{bonf} \geq 0.3649$; Fig. 4 B,C).

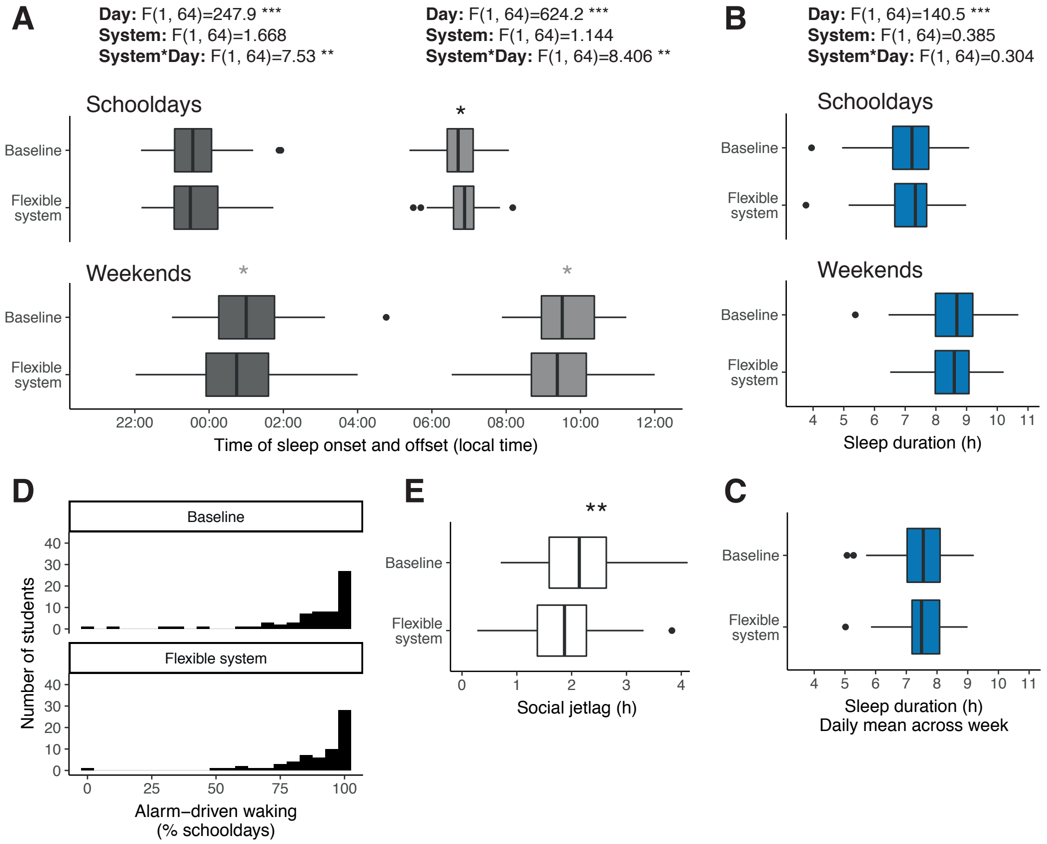

Fig. 4 | Comparison of sleep parameters between baseline and the flexible system. Sleep parameters are from sleep diaries of the full cohort (n=65). A) Average sleep onset (dark grey) and offset (light grey) times, and B) average sleep duration for both study phases on schooldays and weekends. Results of 2-way ANOVAs for the factors weekday (schoolday/weekend) and system (baseline/flexi system) are given above the respective graphs. *, p<0.05 indicates results of *post-hoc* tests before (grey) and after (black) Bonferroni correction. C) Average mean daily sleep duration across a week (weighted for 5 schooldays and 2 weekend days). D) Distributions of individuals' frequency of alarm-driven waking on schooldays. E) Average social jetlag at baseline and in the flexible system. Statistical analysis was performed via paired t-tests or Wilcoxon signed rank tests with *, p<0.05; **, p<0.01; ***, p<0.001; Tukey boxplots.

Alarm-driven waking

The proportion of schooldays on which students were woken by their alarm was as high in the flexible system (median: 93%; IQR: 86-100%) as during baseline (median: 95%; IQR: 85-100%; Z=0.47, p=0.64, r=0.04). The odds ratio from logistic regression also did not indicate a change in odds for alarm-driven waking between school systems (OR=0.69; 95% CI=0.29-1.62; Table S3) (Fig. 4D).

Social jetlag

Overall, the typically large differences in sleep timing and duration between schooldays and weekends were found both during baseline and the flexible system (Fig. 4A,B). However, sleep timing on weekends became slightly earlier in the flexible system, with both onset and offset advanced by a mean of 12 min (SD: 54 min and 48 min) - a difference with small effect size and only statistically significant before correction for multiple testing (onset: $t(64) = 2.092$; $p_{bonf} = 0.0808$; d= 0.200; offset: $t(64) = 2.264$; $p_{bonf} = 0.0539$; d= 0.227; post-hoc tests). In combination with the later offset times on schooldays, however, this trend towards earlier sleep on the weekend resulted in a slight reduction in students' social jetlag by 18 min (SD: 42 min) in the flexible system (*t*(64)=3.309; p=0.0015; d=0.411) (Fig. 4E). Although this suggests a positive effect of the flexible system, we urge for a cautious interpretation, given the small effect sizes and the manifold unsystematic reasons why sleep timing might have become earlier on the weekend days monitored. Furthermore, a systematic advance of sleep timing in spring following the advance of dawn has been found in several studies[71,72] and could explain this effect.

Subjective improvements in the flexible system

Interestingly, although the daily sleep diary entries did not indicate a general improvement in sleep parameters between baseline and flexible system, students nonetheless felt that they were faring better overall in the new system (Fig. 5). In our survey at the end of the study, which was filled out by 56 of the 65 participants, students estimated their sleep times to be 0.5 h longer (median) in the flexible system than at baseline (Z=5.15, p<0.0001, r=0.49) and also rated their sleep quality higher (Z=4.83, p<0.0001, r=0.46) (Fig. 5 A,B). Merely their alarm need was not altered in their view (Z=1.36, p=0.17, r=0.13,; Fig. 5C). In terms of cognitive improvements, students felt that they were less tired (Z=4.67, p<0.0001, r=0.44) and could concentrate better during class (Z=5.07, p<0.0001, r=0.48) and that their ability to study at home after school was improved (Z=3.88, p=0.0001, r=0.37) (Fig. 5 D-F).

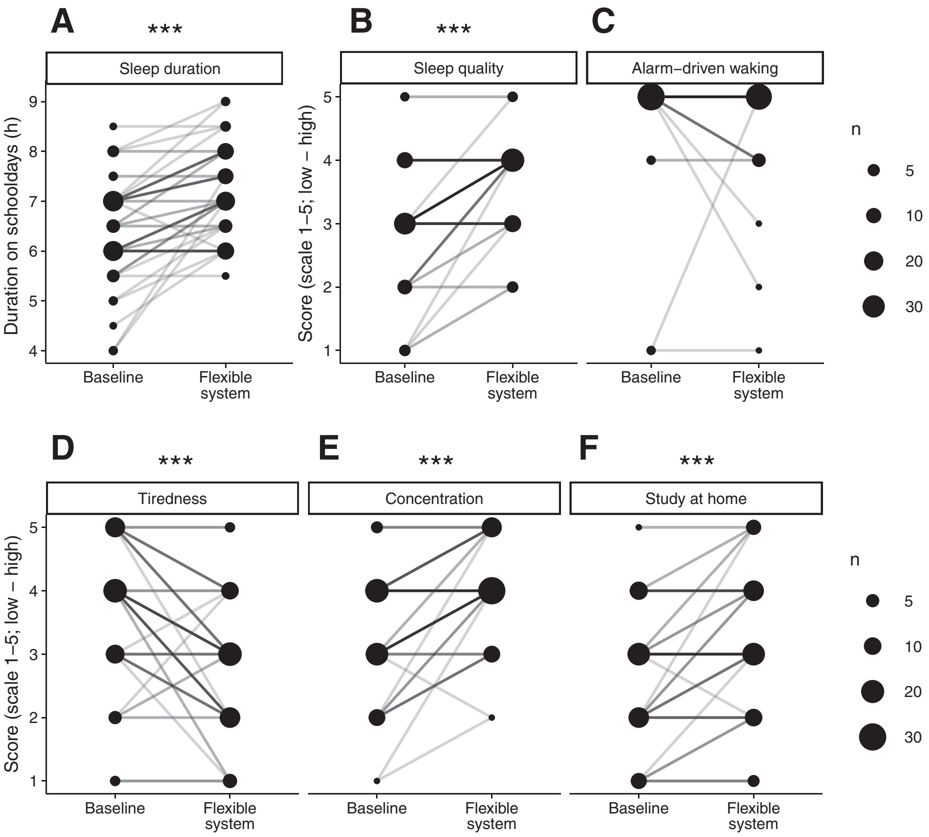

Fig. 5 | Student-rated benefit from the flexible system. Depicted are the results of the final survey in which participants rated their subjective experience in the conventional and flexible system. Data are from 56 students of the main cohort of 65 (9 students did not return the final survey). A) Sleep duration on schooldays in hours. B) Sleep quality on schooldays. C) Frequency of alarm-driven waking per school week. Subjective score for D) tiredness during class, E) ability to concentrate during class, and F) ability to study at home after school. Data are displayed in bubble charts to represent the categorical nature of the data. The area of each circle indicates the number of data points represented. Lines show trajectories of each individual (within-subject trajectories); the darkness of each line illustrates the number of individual trajectories it represents. *, p<0.05; **, p<0.01; ***, p<0.001 for Wilcoxon signed rank tests.

Frequent ≥9AM-starts in the flexible system are associated with longer sleep in the flexible system

The discrepancy of a universal sleep benefit from ≥9AM-starts but not obviously from the flexible system overall might result from students' low use of the 9AM-option in the flexible system.

Frequency of ≥9AM-school starts in the flexible system versus the conventional system

In the flexible system, students used the 9AM-option on average on only 2 days per week (median: 39%, IQR: 20-60% of schooldays, Fig. 6A). The frequency of 9AM-use was not stable across the 6-weeks monitored but tended towards highest values in week 3 and lowest in week 6 (see Fig. S3). In the conventional system at baseline, students had no scheduled first period on ~1 day per week (median: 20%, IQR: 14-27% of schooldays, Fig. 6A), a median difference of only 0.75 days from the flexible system (IQR: 0.2-1.7 days; Z=5.35, p<0.0001, r=0.47). This small increase in the number of later starts in the flexible system might hence be the reason for the lack in measurable sleep benefit in our study.

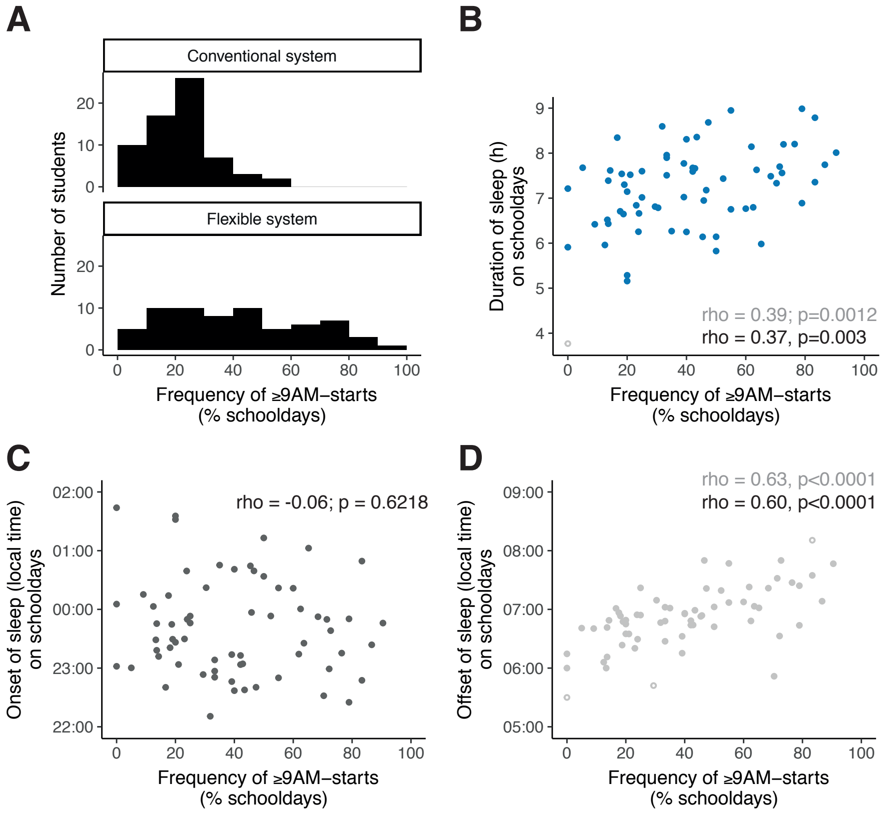

Fig. 6 | Extension of sleep in the flexible system across all schooldays in relation to the frequency of ≥9AM-starts. Sleep parameters are from sleep diaries of the full cohort (n=65). The frequency of ≥9AM-starts is the proportion of schooldays that a student reported to have skipped the first period of i.e. attended school at ≥9AM. A) Distributions of the frequency of ≥9AM-starts in the flexible system in comparison to that in the conventional system as retrospectively extracted from students' timetables (not exactly as in baseline due to teacher absences and exams). Average B) sleep duration, C) sleep onset and D) sleep offset times across all schooldays in the flexible system against frequency of ≥9AM-starts. Data are color-coded as in Fig. 3. Tukey outliers in y-axis values are marked by empty grey circles. Results of Spearman correlations are given for both data including outliers (grey) and excluding outliers (black). *, p<0.05; **, p<0.01; ***, p<0.001 for Wilcoxon signed rank test.

Frequency of ≥9AM-starts in the flexible system - associated with sleep duration but not sleep timing

We therefore sought to determine whether greater use of the 9AM-option was linked with better sleep in the flexible system. Although we could not perform direct comparisons *between* the systems factoring in start times (due to lack of information about exact start times during each day in baseline, see methods), we were able to check for associations *within* the flexible system.

Above, we demonstrated that the benefit of going to school at ≥9AM (DELTA sleep duration) was not affected by how often students actually went at ≥9AM – it was similarly high for all 9AM-use frequencies (Fig. 3D). Our broader analyses here - looking at sleep parameters *across* the flexible system instead of ≥9AM to 8AM-day differences - show that making greater use of the 9AM-option is clearly associated with longer sleep in the flexible system. The more frequently students started school later, the longer was both their average schoolday sleep (rho=0.39; p=0.0012) as well as their average sleep duration across the week (rho=0.31; p=0.0119) (Fig. 6B).

Importantly, this effect appears to be driven solely through later offset times, because 9AM-use was highly correlated with wake-up time (rho=0.63, p<0.0001) but not at all with sleep onset time (rho=-0.06, p=0.6218, Fig. 6C,D). This suggests (although these are just associations) that going to school later more often does not delay overall sleep onsets and, hence, the benefit is maintained. This is further supported by our finding, that weekend sleep timing (onset, offset) and duration were also not associated with the frequency of ≥9AM-starts (weekend onset: rho=0.06, p=0.6584; weekend offset: rho=-0.03, p=0.7944; weekend duration: rho=0.06, p=0.6368).

Discussion

The debate about school start times is currently of very broad scientific and political interest given the widespread problem of teenage sleep deprivation. One of the first observations of a potential relationship between school start times and sleep was made in 1913 by Terman and Hocking[73]. They found that US students slept longer compared to German students; then schools started at 9AM in the US and at 8AM in Germany. This notwithstanding, school starts in the US have since become even earlier than those in Germany and those in Germany were maintained.

Evidence that this trend goes into the wrong direction has been accumulating over the last decades. Numerous studies have documented teenage sleep deprivation[e.g. 19–22,74], linked it with short and long-term performance and health deficits[e.g. 23,40,75,76], and indicated that later school start times are likely an effective public countermeasure[e.g. 43,44,51,77,78]. However, most studies were performed on cross-sectional samples and are limited by nature in their design and thus evidence level[44,45,77] (with randomization and blinding virtually impossible). Hence more studies with different designs and, particularly, better sleep measures are urgently needed.

We had the opportunity to monitor sleep intra-individually over many weeks in a group of German high-school students, whose school system was changed from a rigid one with mainly 8AM-starts to a flexible one with both 8AM- and ≥9AM-starts. Our results are in line with the majority of the other studies on school start times, and support the need for a change in school schedules.

Sleep duration is longer on ≥9AM-days and the benefit is universal

In our study, virtually all participating students (97%) benefitted from later start times, sleeping longer on schooldays with a ≥9AM-start – on average students gained 1 hour of sleep on those days. Importantly, not only was the overall benefit universal but also the magnitude of the benefit was similar across the important factors chronotype, gender, grade, and frequency of later starts. This may seem surprising at first but should actually be alarming: it exemplifies how severe and wide-spread teenage sleep deprivation may be, afflicting practically every single student leading to such ceiling effects. In our study sample, students rarely woke without their alarm clocks on schooldays, indicating that they rose before their sleep need was met and their internal day had started. Indeed, only 18% slept 8 hours or more on their schooldays, the lower bound of the recommended 8-10 hours for this population[79] - only 1 student slept on average over 9 hours (based on sleep diary entries during baseline). These numbers are in line not only with the age and gender-matched adolescents across Germany used in our study (Fig. S1), but also with other studies in Germany[80] and around the world[e.g. 19–22,74] - worrying statistics considering the acute and long-term health and performance detriments linked with teenage sleep deprivation[e.g. 23,40,75,76].

On days with a ≥9AM school start, these statistics looked much less bleak: 52% of students slept more than 8 hours and 13% even more than 9 hours, subjective sleep quality was improved and alarm-free waking was more likely (albeit still rare). However, the delay from an 8AM to a ≥9AM-school start was insufficient to separate the moderately sleep-deprived students from the heavily sleep-deprived students (and to bring out features linked with smaller or greater sleep gain). Similarly, the earlier chronotypes among the students were still quite late in their sleep timing compared to other age groups and thus benefitted as fully from the ~1-hour-delay as the later chronotypes. This suggests that the school start delay from 8AM to 9AM may be at the lower end of the required spectrum to counter teenage sleep deprivation.

Sleep onset does not delay

One of the greatest concerns regarding later school starts is that teenagers might be tempted to stay up even later in the evening either consciously or via delayed circadian rhythms from later exposure to advancing morning light. As a result, they would not gain more sleep but potentially further delay their circadian rhythms through prolonged exposure to evening light. Supporting this line of thought, a recent modelling paper indicated that a delay in wake-up alone may not effectively increase sleep duration or reduce social jetlag long-term (tested for 5 weeks in the model) unless controlling evening light exposure[81]. In our study, however, there was no evidence that sleep onset times differed between ≥9AM-days and 8AM-days. Even the students that went most often at ≥9AM did not show later sleep onsets than those that made less use of the 9AM-option. These findings tally generally with those from many other studies, which also did not observe systematic delays in adolescents' sleep timing after a delay in school starts[51]. However, also the opposite has been reported[51] and direct comparisons are hampered by the fact that many studies were based on cross-sectional data[45] and/or did not distinguish between bedtime (time of going to bed) and sleep onset (time of falling asleep). Importantly, the extended sleep and the stable sleep onset we observed in our study are based on a period of 6 weeks after the change into the new flexible system, suggesting that the sleep benefit might be maintained in the long term. However, longitudinal studies with follow-up assessments are needed to confirm this.

The flexible system: curse or cure?

There are many possible reasons for the absence of a delay in sleep onset. One reason may be the fact that school start choices influenced the opportunity for natural morning light exposure at the geographical location and season (winter/spring) of the study. During the flexible system, most students woke *before* sunrise on their 8AM-days and *after* sunrise on most of their ≥9AM-days. This longer window for natural daylight exposure before school – natural light is a stronger signal for the circadian system than artificial light - might have countered any circadian delay resulting from later timing of artificial light exposure at home. Alternatively, if considering psychological factors leading to stable sleep onsets, students reported to feel more alert and less tired and to sleep better in the flexible system. It is therefore possible that they consciously took advantage of longer sleep opportunities because they felt the benefits of getting more sleep.

However, also the flexibility of the system *per se* could be a reason for the stability in sleep onset. Even just knowing that one could wake up later if required might have improved students' attitude and anxiety around sleep, facilitating an earlier sleep onset and more restful sleep. Furthermore, variable wake up times may positively affect exposure to morning light from artificial sources (independent of sunrise times). *Permanent* later start times generally purport a delay in circadian timing by delaying overall light exposure in the morning for all days of the school week. In contrast, the occasional early start in the flexible system may help to prevent such a delay through ensuring occasional earlier light exposure.

Therefore, one could speculate that providing flexibility may be instrumental in maximizing sleep benefits from later school starts – as long as increased sleep variability on schooldays can be offset by less sleep variability between schooldays and non-schooldays.

What argues against the positive impact of flexibility, however, is our finding that sleep duration was not significantly different between the conventional and the flexible system despite the clear sleep benefit when comparing ≥9AM-days to 8AM-days in the flexible system. The students in our cohort did not make great use of the 9AM-option but started school later on not even one full additional day per week, making it two days per week on average. With this low frequency of later starts, the net gain from the flexible system was negligible to non-detectable. This implies that students, given the choice, may not necessarily opt for what may be best for their sleep.

Why did students not opt for more later starts?

The low use of the 9AM-option greatly surprised us. It is not only at odds with the pervasive sleep deprivation in our sample but also with the results from our final survey where 64% indicated that an 8AM-school start was tough for them (always or most of the time) and 86% that a 9AM-start was actually easier (always or most of the time). Was the low frequency of 9AM-starts incorrectly reported or does it originate from a selection bias? Both seem unlikely: The low diary-reported 9AM-use from our participants tallied with the retrospective survey-reported 9AM-use – not only from the participants but also with that from additional 82 anonymous students that transitioned into the flexible system but did not take part in the study; students across the board did indeed not go later more often. Exploratory analyses did not reveal any stable predictors of 9AM-start frequency from baseline sleep, lifestyle or commuting factors (data not shown). However, given the many factors that can reasonably be assumed to influence the 9AM-use - of which many were not documented in our study (e.g. individual daily timetables, after-school appointments, carpooling, parents' attitude towards later school starts, exams etc.) - our sample was likely insufficient for the complexity of the question.

Asked about the reasons for starting school at 8AM instead of 9AM in our survey, the most frequent answer (75% of students) was "to fulfil the school's quota of 10 self-study hours per week". If not enough free periods existed in a student's schedule, students had to stay longer in the afternoon. It is therefore likely that students opted for early mornings rather than late afternoons and thus made such little use of the 9AM-option. Time management training may help students to better organize their schedules in this regard, whereas the school may want to try to optimize their timetabling. Further frequent reasons for 8AM-starts were "easier logistics to get to school" (40%), an important factor in the implementation of changes in school start times, and "more time to learn (27%)", indicating that students got extra teacher-supervised study time when going to school at 8AM. Follow-up studies will hopefully shed light on this intriguing low use of later start times to guide better implementations of such a flexible system – which was after all liked by 98% of participating students.

Limitations

While selection bias is unlikely to explain the surprisingly low uptake of the 9AM-option as discussed above, it might still have had a systematic effect on some of our other results: of 253 eligible students, only 26% made up the final study cohort. Since the sleep characteristics of the study cohort closely match those of other German adolescents from i) the MCTQ database sample and ii) other published data, the selection bias in this study is likely of a similar magnitude as in other studies.

All sleep durations in our study are based solely on nocturnal sleep of students. Occasional or regular naps were thus not considered in any of the analyses, which may have led to underestimation of total daily sleep duration in some students.

In addition, our assessment of alarm-driven waking might have underestimated the rate of non-natural waking since it did not cover students woken regularly by their parents or siblings. The detected decrease in alarm-driven waking on ≥9AM-days may thus not reflect only increased natural waking but also incorporate a switch from alarm-driven to parent-induced waking.

Finally, seasonal changes in photoperiod and associated changes in sleep timing and duration may have systematically influenced our findings, potentially explaining part of the null-effect of the flexible system on sleep. With our study running from January to March at 50°N, our comparisons between baseline (Jan) and flexible system (Feb-Mar) were likely confounded by the gradual advance of dawn during spring linked with gradually earlier sleep offset times and shortening sleep durations[71,72]. Therefore, seasonal changes in sleep may have offset potential positive (albeit small) effects of the flexible system on sleep rendering them undetectable. Vice versa, the one positive effect detected in the flexible system, the small reduction in social jetlag resulting from earlier timing of weekend sleep, might well be a false positive finding caused by the seasonal trajectory towards earlier sleep. In contrast to these pre-post comparisons, comparisons of 8AM and ≥9AM-school starts within the flexible system were most likely independent of seasonal changes because 8AM and ≥9AM-days occurred interspersed and alternating throughout the flexible system within each individual.

Concluding remarks

Our study is one of the first evaluating the effects of later school start times on sleep and subjective performance parameters in Europe. A flexible system with both early and late start times could be a valid additional solution to the more common policy of delaying school start times outright – if students can be encouraged to use the late option frequently enough.

On days with a later start, students have the opportunity to sleep longer. This should reduce the accumulation of sleep debt during the week. The occasional 8AM-starts could be strategic in avoiding a delay in sleep onset by ensuring that students are exposed to light in the early morning on a weekly basis. In addition, especially important for practical applications, students prefer the flexible system and their subjective parameters are improved.

There are other examples of successful implementations of flexible school systems. In The Netherlands, there are schools where the main subjects are taught in the middle of the day (e.g. from 10AM till 2PM), while students can choose whether to learn minor, facultative subjects earlier in the morning or later in the afternoon. Such a system accommodates the wide distribution of chronotypes in the student population.

In conclusion, our results are in line with the accumulating scientific evidence supporting later school start times as a countermeasure against teenage sleep deprivation. We therefore urge more schools to delay their start times and to collaborate with scientists to increase our knowledge about the (long-term) effects of later starting times on sleep, subjective well-being, health and performance.

Acknowledgements

We thank Dr. LK Pilz for implementing the sleep diary online; A Khan Sullivan for school schedule digitization; the management of the Gymnasium Alsdorf, in particular W Bock and O Vollert, for their open support throughout the study; the student participants for their willingness and effort to share their data.

Financial Disclosure

None

Non-financial Disclosure

None. In particular, none of the authors have any private links with the Gymnasium Alsdorf, and the school was not involved in data analysis or interpretation nor the writing of the manuscript.

References

1. Dahl RE, Allen NB, Wilbrecht L, Suleiman AB. Importance of investing in adolescence from a developmental science perspective. *Nature*. 2018;554:441–50. doi:10.1038/nature25770.
2. Roenneberg T, Kuehnle T, Pramstaller PP, Ricken J, Havel M, Guth A, et al. A marker for the end of adolescence. *Curr Biol*. 2004;14:R1038–9. doi:10.1016/J.CUB.2004.11.039.
3. 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. doi:10.1371/journal.pone.0112199.
4. Crowley SJ, Acebo C, Carskadon MA. Sleep, circadian rhythms, and delayed phase in adolescence. *Sleep Med*. 2007;8:602–12. doi:10.1016/j.sleep.2006.12.002.
5. Fischer D, Lombardi DA, Marucci-Wellman H, Roenneberg T. Chronotypes in the US – Influence of age and sex. *PLoS One*. 2017;12:1–17.
6. 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. doi:10.1016/j.adolescence.2018.06.001.
7. Carskadon MA, Acebo C, Richardson GS, Tate BA, Seifer R. An Approach to Studying Circadian Rhythms of Adolescent Humans. *J Biol Rhythms*. 1997;12:278–89.
8. Carskadon MA, Acebo C, Jenni OG. Regulation of adolescent sleep: Implications for behavior. *Ann N Y Acad Sci*. 2004;1021:276–91. doi:10.1196/annals.1308.032.
9. Jenni OG, Achermann P, Carskadon MA. Homeostatic sleep regulation in adolescents. *Sleep*. 2005;28:1446–54. doi:10.1093/sleep/28.11.1446.
10. Taylor DJ, Jenni OG, Acebo C, Carskadon MA. Sleep tendency during extended wakefulness: Insights into adolescent sleep regulation and behavior. *J Sleep Res*. 2005;14:239–44. doi:10.1111/j.1365-2869.2005.00467.x.
11. Van Den Bulck J. Television viewing, computer game playing, and internet use and self-reported time to bed and time out of bed in secondary-school children. *Sleep*. 2004;27:101–4. doi:10.1093/sleep/27.1.101.
12. Munezawa T, Kaneita Y, Osaki Y, Kanda H, Minowa M, Suzuki K, et al. The Association between Use of Mobile Phones after Lights Out and Sleep Disturbances among Japanese Adolescents: A Nationwide Cross-Sectional Survey. *Sleep*. 2011;34:1013–20. doi:10.5665/SLEEP.1152.
13. Cajochen C. Alerting effects of light. *Sleep Med Rev*. 2007. doi:10.1016/j.smrv.2007.07.009.
14. Souman JL, Tinga AM, te Pas SF, van Ee R, Vlaskamp BNS. Acute alerting effects of light: A systematic literature review. *Behav Brain Res*. 2018. doi:10.1016/j.bbr.2017.09.016.
15. Yang M, Ma N, Zhu Y, Su YC, Chen Q, Hsiao FC, et al. The acute effects of intermittent light exposure in the evening on alertness and subsequent sleep architecture. *Int J Environ Res Public Health*. 2018. doi:10.3390/ijerph15030524.
16. Khalsa SBS, Jewett ME, Cajochen C, Czeisler CA. A phase response curve to single bright light pulses in human subjects. *J Physiol*. 2003. doi:10.1113/jphysiol.2003.040477.
17. Chang A-M, Aeschbach D, Duffy JF, Czeisler CA. Evening use of light-emitting eReaders negatively affects sleep, circadian timing, and next-morning alertness. *Proc Natl Acad Sci*. 2014;112:201418490. doi:10.1073/pnas.1418490112.
18. Carissimi A, Dresch F, Martins AC, Levandovski RM, Adan A, Natale V, et al. The influence of school time on sleep patterns of children and adolescents. *Sleep Med*. 2016. doi:10.1016/j.sleep.2015.09.024.

19. 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 Public Health*. 2006;6:1–9. doi:10.1186/1471-2458-6-116.

20. 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. doi:10.1016/j.smrv.2011.03.005.

21. Keyes KM, Maslowsky J, Hamilton A, Schulenberg J. The Great Sleep Recession: Changes in Sleep Duration Among US Adolescents, 1991-2012. *Pediatrics*. 2015. doi:10.1542/peds.2014-2707.

22. 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. doi:10.1016/j.sleep.2010.11.008.

23. 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. doi:10.1016/j.smrv.2009.10.004.

24. 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 Public Health*. 2015. doi:10.1177/1403494814556647.

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

26. 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;9:517–26. doi:10.1016/j.sleep.2007.07.003.

27. Vorona RD, Szklo-Coxe M, Wu A, Dubik M, Zhao Y, Ware JC. Dissimilar teen crash rates in two neighboring southeastern virginia cities with different high school start times. *J Clin Sleep Med*. 2011;7:145–51.

28. Vaca F, Harris JS, Garrison HG, Vaca F, McKay MP. Drowsy Driving. *Ann Emerg Med*. 2005;45:433–4. doi:10.1016/j.annemergmed.2005.01.015.

29. 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. doi:10.1038/ijo.2011.149.

30. Mullington JM, Haack M, Toth M, Serrador JM, Meier-Ewert HK. Cardiovascular, Inflammatory, and Metabolic Consequences of Sleep Deprivation. *Prog Cardiovasc Dis*. 2009. doi:10.1016/j.pcad.2008.10.003.

31. 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. doi:10.1080/15402002.2015.1120198.

32. Short MA, Gradisar M, Lack LC, Wright HR. The impact of sleep on adolescent depressed mood, alertness and academic performance. *J Adolesc*. 2013. doi:10.1016/j.adolescence.2013.08.007.

33. 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. doi:10.1111/jcpp.12125.

34. Tynjälä J, Kannas L, Levälahti E. Perceived tiredness among adolescents and its association with sleep habits and use of psychoactive substances. *J Sleep Res*. 1997;6:189–98. doi:10.1046/j.1365-2869.1997.00048.x.

35. Pasch KE, Latimer LA, Cance JD, Moe SG, Lytle LA. Longitudinal Bi-directional Relationships

Between Sleep and Youth Substance Use. *J Youth Adolesc*. 2012. doi:10.1007/s10964-012-9784-5.

36. Wittmann M, Dinich J, Merrow M, Roenneberg T. Social Jetlag: Misalignment of Biological and Social Time. *Chronobiol Int*. 2006;23:497–509. doi:10.1080/07420520500545979.

37. Larcher S, Gauchez AS, Lablanche S, Pépin JL, Benhamou PY, Borel AL. Impact of sleep behavior on glycemic control in type 1 diabetes: The role of social jetlag. *Eur J Endocrinol*. 2016;175:411–9. doi:10.1530/EJE-16-0188.

38. Parsons MJ, Moffitt TE, Gregory AM, Goldman-Mellor S, Nolan PM, Poulton R, et al. Social jetlag, obesity and metabolic disorder: Investigation in a cohort study. *Int J Obes*. 2015;39:842–8. doi:10.1038/ijo.2014.201.

39. Roenneberg T, Allebrandt K V., Merrow M, Vetter C. Social jetlag and obesity. *Curr Biol*. 2012;22:939–43. doi:10.1016/j.cub.2012.03.038.

40. Owens J. Insufficient Sleep in Adolescents and Young Adults: An Update on Causes and Consequences. *Pediatrics*. 2014;134:e921–32. doi:10.1542/peds.2014-1696.

41. 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 Heal*. 2017;86:363–81. doi:10.1111/josh.12388.School.

42. Bowers JM, Moyer A. Effects of school start time on students' sleep duration, daytime sleepiness, and attendance: a meta-analysis. *Sleep Heal*. 2017;3:423–31. doi:10.1016/j.sleh.2017.08.004.

43. 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. doi:10.1097/DBP.0000000000000018.

44. 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. doi:10.1016/j.smrv.2015.06.002.

45. 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. *Cochrane Database Syst Rev*. 2017;2017. doi:10.1002/14651858.CD009467.pub2.

46. Lufi D, Tzischinsky O, Hadar S. Delaying school starting time by one hour: Some effects on attention levels in adolescents. *J Clin Sleep Med*. 2011;7:137–43.

47. Nahmod NG, Lee S, Master L, Chang AM, Hale L, Buxton OM. Later high school start times associated with longer actigraphic sleep duration in adolescents. *Sleep*. 2019;42:1–10. doi:10.1093/sleep/zsy212.

48. 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. doi:10.1126/sciadv.aau6200.

49. Lo JC, Lee SM, Lee XK, Sasmita K, Chee NIYN, Tandi J, et al. Sustained benefits of delaying school start time on adolescent sleep and well-being. *Sleep*. 2018. doi:10.1093/sleep/zsy052.

50. Carskadon MA, Wolfson AR, Acebo C, Tzischinsky O, Seifer R. Adolescent sleep patterns, circadian timing, and sleepiness at a transition to early school days. *Sleep*. 1998;21:871–81.

51. 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 Heal*. 2016;86:363–81. doi:10.1111/josh.12388.School.

52. Der Deutsche Schulpreis 2019. https://www.deutscher-schulpreis.de/preistraeger/gymnasium-der-stadt-alsdorf.

53. Parkhurst H. Education on the Dalton Plan. New York: E.P. Dutton & Company; 1922.

54. Dalton International n.d. https://daltoninternational.org/ (accessed June 11, 2019).

55. Roenneberg T. What is chronotype? *Sleep Biol Rhythms*. 2012;10:75–6. doi:10.1111/j.1479-8425.2012.00541.x.

56. Roenneberg T, Kuehnle T, Juda M, Kantermann T, Allebrandt K, Gordijn M, et al. Epidemiology of the human circadian clock. *Sleep Med Rev*. 2007;11:429–38. doi:10.1016/j.smrv.2007.07.005.

57. The WeP - the worldwide experiment platform n.d. https://www.thewep.org/documentations/mctq (accessed June 11, 2019).

58. Ghtobi N, Pilz LK, Winnebeck EC, Vetter C, Zerbini G, Lenssen D, et al. The µMCTQ - an ultra-short version of the Munich ChronoType. *J Biol Rhythms*. 2019;in press.

59. Roenneberg T, Keller LK, Fischer D, Matera JL, Vetter C, Winnebeck EC. Human activity and rest in situ. Methods Enzymol., vol. 552, Academic Press; 2015, p. 257–83. doi:10.1016/bs.mie.2014.11.028.

60. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2019.

61. Torchiano M. R package effsize: Efficient Effect Size Computation. Version 0.7.1 2017.

62. Wickham H. ggplot2: Elegant Graphics for Data Analysis 2016.

63. Kassambara A. R package ggpubr: “ggplot2” Based Publication Ready Plots. Version 0.2 2018.

64. Harrell Jr FE. R package Hmisc: Harrell Miscellaneous. Version 4.1-1 2018.

65. Bates D, Maechler M, Bolker B, Walker S. R package lme4: Linear Mixed-Effects Models using “Eigen” and S4. Version 1.1-18-1 2018.

66. Kuznetsova A, Brockhoff PB, Christensen RHB. R package lmerTest: Test in Linear Mixed Effects Models. Version 3.0-1 2018.

67. Pohlert T. R package PMCMRplus: Calculate Pairwise Multiple Comparisons of Mean Rank Sums Extended. Version 1.4.1 2019.

68. Neuwirth E. R package RColorBrewer: ColorBrewer Palettes. Version 1.1-2 2014.

69. Wickham H. R package reshape2: Flexibility Reshape Data: A Reboot of the Reshape Package. Version 1.4.3 2017.

70. Rosenthal R. Meta-analytic procedures for social research. 2nd. CA:Sage; 1991.

71. Kantermann T, Juda M, Merrow M, Roenneberg T. The Human Circadian Clock’s Seasonal Adjustment Is Disrupted by Daylight Saving Time. *Curr Biol*. 2007;17:1996–2000. doi:10.1016/j.cub.2007.10.025.

72. Hashizaki M, Nakajima H, Shiga T, Tsutsumi M, Kume K. A longitudinal large-scale objective sleep data analysis revealed a seasonal sleep variation in the Japanese population. *Chronobiol Int*. 2018;35:933–45. doi:10.1080/07420528.2018.1443118.

73. Terman LM, Hocking A. The sleep of school childeen: Its distribution according to age and its relation to physical and mental efficiency. *J Educ Psychol*. 1913;4:199–208.

74. Leger D, Beck F, Richard JB, Godeau E. Total Sleep Time Severely Drops during Adolescence. *PLoS One*. 2012. doi:10.1371/journal.pone.0045204.

75. Roberts RE, Roberts CR, Duong HT. Sleepless in adolescence: Prospective data on sleep deprivation, health and functioning. *J Adolesc*. 2009. doi:10.1016/j.adolescence.2009.03.007.

76. Dahl RE, Lewin DS. Pathways to adolescent health: Sleep regulation and behavior. *J Adolesc Heal*. 2002. doi:10.1016/S1054-139X(02)00506-2.

77. Wahlstrom KL, Owens JA. School start time effects on adolescent learning and academic performance, emotional health and behaviour. *Curr Opin Psychiatry*. 2017;30:485–90. doi:10.1097/YCO.0000000000000368.

78. Short MA, Gradisar M, Lack LC, Wright HR, Dewald JF, Wolfson AR, et al. A Cross-Cultural Comparison of Sleep Duration Between U.S. and Australian Adolescents: The Effect of School Start Time, Parent-Set Bedtimes, and Extracurricular Load. *Heal Educ Behav*. 2013;40:323–30. doi:10.1177/1090198112451266.

79. Paruthi S, Brooks LJ, D'Ambrosio C, Hall WA, Kotagal S, Lloyd RM, et al. Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine. *J Clin Sleep Med*. 2016;12:785–6. doi:10.5664/jcsm.5866.

80. Loessl B, Valerius G, Kopasz M, Hornyak M, Riemann D, Voderholzer U. Are adolescents chronically sleep-deprived? An investigation of sleep habits of adolescents in the Southwest of Germany. *Child Care Health Dev*. 2008;34:549–56. doi:10.1111/j.1365-2214.2008.00845.x.

81. Skeldon AC, Phillips AJK, Dijk DJ. The effects of self-selected light-dark cycles and social constraints on human sleep and circadian timing: A modeling approach. *Sci Rep*. 2017. doi:10.1038/srep45158.

APPENDIX - Project 1

Supplementary information for

“Later school start times in a flexible system improve teenage sleep”

Authors:

Eva C. Winnebeck, Maria T. Vuori-Brodowski, Anna M. Biller, Carmen Molenda, Dorothee Fischer, Giulia Zerbini & Till Roenneberg

Contact information

eva.winnebeck@med.uni-muenchen.de
till.roenneberg@med.uni-muenchen.de

Methods

Sleep diary details

The online sleep diary was set up the following: After prompts for personal ID and date of the wake-up day, students were asked about their previous night's sleep: 1. the time they fell asleep (sleep onset; with a specific note that this is the time when they fell asleep and not when they went to bed); 2. the time they woke up (sleep offset; with a specific note that this is the time when they woke up and not got up); 3. if they were woken by their alarm clock (yes/no); 4. if the wake-up day was a school day (yes/no); 5. in case question 4 was answered with "yes", the students had to indicate if they participated in the first lesson at 8AM (yes/no; this question was only added to the questionnaire with the introduction of flexible system); 6. their subjective sleep quality on a 10-Point-Likert-Scale (1= "very bad", 10="very good"). They were asked to specify all times in hh:mm on a 24-hour scale. The questionnaire did not cover any naps during the day. To illustrate the difference between going to bed/waking up as well as falling asleep/getting up, the sleep diary was headed by an infographic. Sleep duration was subsequently calculated using sleep onset and sleep offset.

Data cleaning

To ensure reliable and congruent data, the following corrections and exclusions were carried out on the diary entries. We searched for obvious errors in ID (e.g. IDs outside the range assigned), wake-up date (e.g. incorrect year, month or confusion day/month) and sleep times due to AM/PM confusion (using negative or very short sleep duration as warning signs). Erroneous entries were manually corrected if the original meaning was clearly identifiable, otherwise they were discarded. Any daytime naps were eliminated as not asked for and thus not provided by all participants.

Subsequently, the record of each individual participant was examined for duplicate entries. Multiple identical entries for the same night were reduced to a single entry. When multiple, non-identical diary entries for one night were identified, the following cleaning rules were applied: i) Non-overlapping entries were interpreted as reports of a nocturnal sleep episode with wake interruption(s). Hence, we fused these sleep bouts into one sleep episode with a single onset and offset and a combined sleep duration corrected for the wake interruption in between. ii) For overlapping entries, we based our cleaning on the timespan elapsed between these entries (calculated from the time stamp of the entries). If entries were made within seconds or minutes of each other, we treated them as correction of the former entry(s), and hence we kept the last entry and removed the former ones. If entries were made with large distance from each other, the first entry was kept and the later entry(s) was deleted assuming that it was an erroneous retrospective entry that the participant had made thinking they had not provided data for that night.

Locomotor Activity Recording

Locomotor activity was recorded over the entire study period in a sub-cohort of 45 participating students via wrist-worn activity-monitoring devices (actimeters). Students were free to choose on which wrist to wear the device but had to keep this constant over the recording period. If actimeters were not worn for more than 30 minutes, students had to indicate this in actimetry logs. We used Daqtometers (Version 1.4, Daqtix, Germany), which are dual-axis accelerometers that detect both static and dynamic acceleration, i.e., motion and changes of position. Acceleration is converted to activity counts by summation of the linear differences of subsequent readings for each axis. Devices were set to sample acceleration every 1s and to store activity counts every 30s as the mean of all samples within the storage interval.

Activity data processing

We used our in-house analysis program ChronoSapiens for data analyses. All activity records were averaged into 10-min bins. We identified episodes when actimeters were not worn as episodes of at least 10 consecutive bins (100 min) of zero activity as well as from subjects' self-reports (actimetry logs). These were excluded from the analysis. In several cases, the 100-min rule was not applied when the stretch of 100-min-zero-activity was detected in the middle of the night and was judged as extremely little movement during sleep.

If actimetry records showed ≥1h of missing data between 9PM (10PM on weekends) and 9AM (10AM on weekends), any sleep bouts of that night were excluded. The same applied if >4h of missing data appeared between 9AM and 9PM the following day (i.e. over a period of 36 h). Missing data render the sleep detection method less reliable by altering the 24-h centered moving average and hence must be accounted for. We only included data that fulfilled our quantity and quality criteria (see *Participants* in main manuscript). After this quality check, 39 out of 45 original records were kept for further analyses.

Identification of sleep bouts

Identification of sleep bouts in the activity records relied on the identification of stretches of relative immobility as detailed in Roenneberg et al. 2015.[1] In this approach, 10-min-bins with activity counts below 20% of the 24-hr centered moving average were classified as potential sleep and then consolidated into longer bouts via a correlation procedure. The resulting comprehensive list of sleep bouts was filtered according to the following criteria to allow for sensible comparison with diary-reported data: i) Individual sleep bouts were fused into one longer bout if they were not separated by >180 min of wake (if the minimum length of the second bout was ≥30 min and the length of the combined bout was <900 min). Sleep duration of the combined bout was calculated as time between onset of the first bout and offset of the second bout minus time awake in between the two combined bouts. ii) Sleep bouts outside the range of 3-15 h in duration were excluded. iii) Any naps (sleep bouts occurring outside the daily 12 h of lowest activity[1]) were excluded since diary reports did not include information on naps.

References

1. Roenneberg T, Keller LK, Fischer D, et al. Human activity and rest in situ. In *Methods in Enzymology* 2015; 552: 257–283. Academic Press. https://doi.org/10.1016/bs.mie.2014.11.028

Supplementary Figures

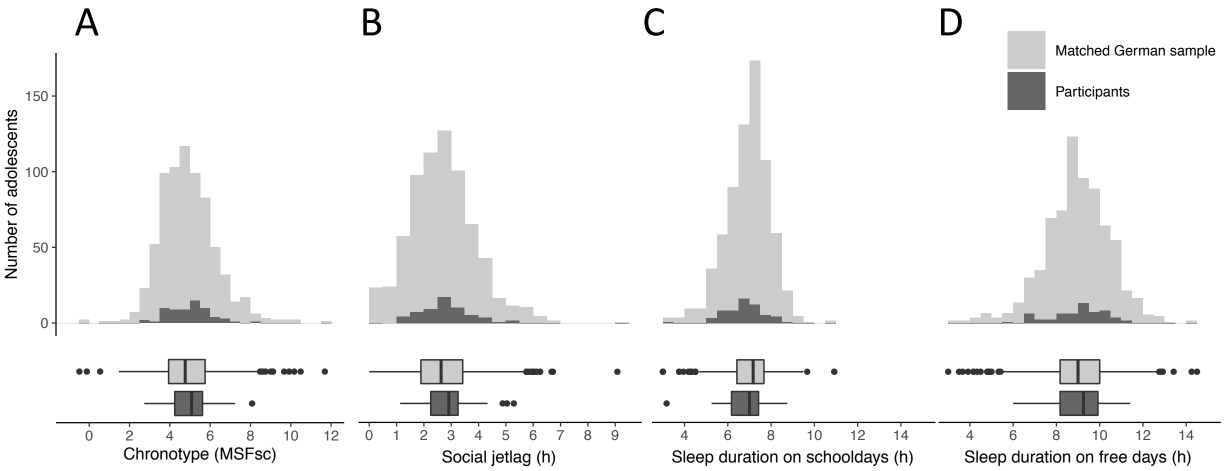

Fig. S1 | Sleep parameters in study participants versus age- and gender-matched German sample. Self-reported sleep parameters from the Munich ChronoType Questionnaire (MCTQ) of the full study cohort (n=65, dark grey) in comparison to a >10-fold larger German sample from the MCTQ database matched in age and gender (n=680, light grey). A) Chronotype (midsleep time on school-free days corrected for oversleep, MSFsc), B) social jetlag, C) sleep duration on schooldays, D) sleep duration on school-free days. Samples were compared via Wilkoxon rank sum tests (all comparisons non-significant).

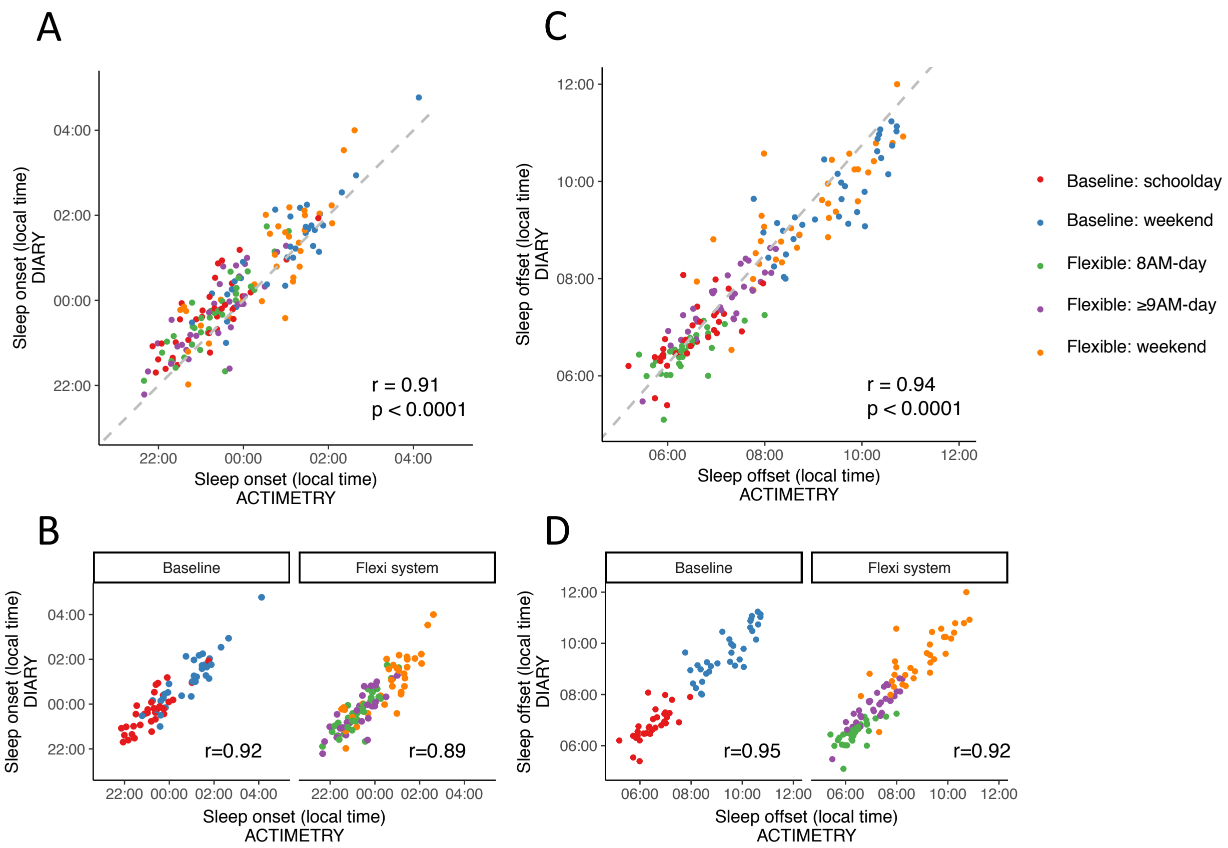

Fig. S2 | Comparison between subjective and objective sleep measures.

Comparison of average sleep onset and offset times per individual from sleep diary and actimetry. Data are from the sub-cohort assessed simultaneously with both techniques (n=34). A) Sleep onset times for all day types and B) split by study phase. C) Sleep offset times for all day types and D) split by study phase. Results of Pearson correlations are indicated in the figures. Dashed lines indicate 1:1 relationship.

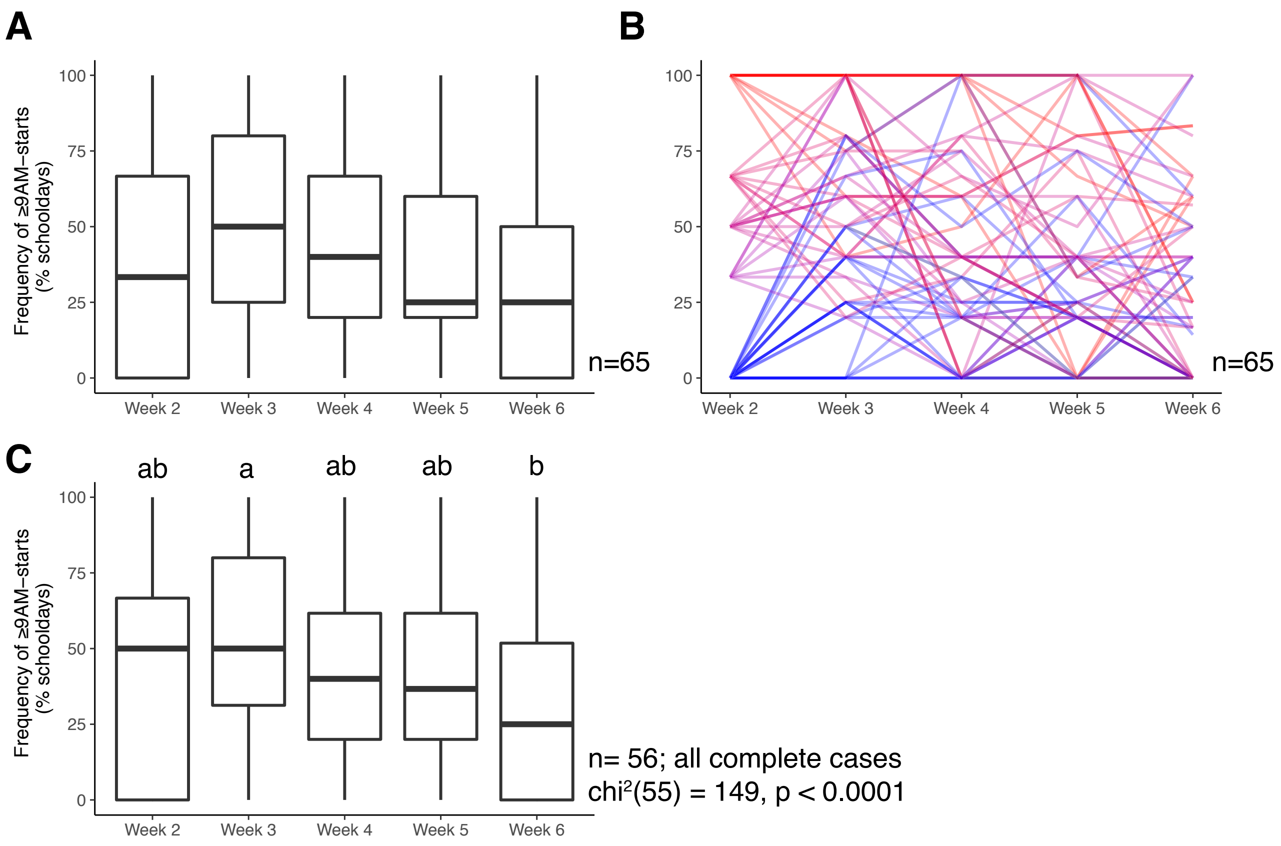

Fig. S3 | Progression of the frequency of 9AM-use.

Frequency of ≥9AM-starts from week 2 to week 6 of the flexible system A),B) for all participants (n=65), C) for all complete cases (n=56) used in the statistical analysis. Lines in B) represent individual trajectories color-coded according to individuals' starting frequency; line intensity indicates the number of trajectories represented. Week 1 was omitted because diaries did not include an 8AM/9AM differentiation for the first 2 days and the end of the week was marked by holidays. Results of Friedman rank sum test are provided for C; letters indicate results from posthoc pairwise comparisons using the Nemenyi-Wilcoxon-Wilcox all-pairs test with single-step p-value adjustment performed via the R-package PMCMRplus (version 1.4.1): weeks marked by different letters were statistically different.

Supplementary Tables

Tab. S1 | Typical timetable during the flexible system. A fortnightly personal timetable of a student attending 11th grade at the Gymnasium Alsdorf after the flexible system was introduced.

Period	Time	Week I					Week II				
		Mon	Tue	Wed	Thu	Fri	Mon	Tue	Wed	Thu	Fri
1	0800 - 0845	Self-study period (flexible)									
2	0850 - 0950	Class	Class	Class	Class	Class	Class	Class	Class	Class	Free
3	1015 - 1115	Free	Class	Class	Class	Class	Free	Class	Class	Class	Class
4	1120 - 1205	Self-study period (mandatory)									
5	1210 - 1310	Class	Class	Free	Class	Class	Class	Class	Free	Class	Class
6	1315 - 1415	Class	Class	Class	Class		Class	Class	Class	Class	
7	1415 - 1515		Class					Class			
8	1515 - 1615										

Tab. S2 | Logistic regression analysis for frequency of alarm-driven waking on 8AM-days versus ≥9AM-days (pertaining to Figure 2H). Logistic mixed effects regression models were performed on the sub-cohort for 8AM/9AM comparison (n=60 students) using 3 dichotomization strategies for low and high alarm-driven waking.

Alarm-driven waking (1 = low, 0 = high) dichotomized at 100% of schooldays (median)

Random effects			Variance	SD	
Student	(Intercept)		1.05	1.03	
Fixed effects	OR	95%CI lower	95%CI upper	z-value	p-value
(Intercept)	0.449	0.201	0.999	-1.96	0.050(*)
School start time (1 = ≥9AM, 0 = 8AM)	3.29	1.28	8.48	2.47	0.014*
Gender (1 = male, 0 = female)	0.344	0.111	1.06	-1.86	0.063

Alarm-driven waking (1 = low, 0 = high) dichotomized at 85% of schooldays (1st quartile)

Random effects			Variance	SD	
Student	(Intercept)		3.07	1.75	
Fixed effects	OR	95%CI lower	95%CI upper	z-value	p-value
(Intercept)	0.059	0.006	0.445	-2.75	0.006**
School start time (1 = ≥9AM, 0 = 8AM)	14.2	2.15	94.1	2.76	0.006**
Gender (1 = male, 0 = female)	0.236	0.038	1.48	-1.54	0.124

Alarm-driven waking (1 = low, 0 = high) dichotomized at <85% and 100% of schooldays

Random effects			Variance	SD	
Student	(Intercept)		2.81	1.68	
Fixed effects	OR	95%CI lower	95%CI upper	z-value	p-value
(Intercept)	0.094	0.016	0.563	-2.59	0.010**
School start time (1 = ≥9AM, 0 = 8AM)	13.3	2.03	87.2	2.70	0.007**
Gender (1 = male, 0 = female)	0.189	0.028	1.30	-1.69	0.090

SD, standard deviation; CI, confidence interval; OR, odds ratio

Tab. S3 | Logistic regression analysis for frequency of alarm-driven waking at baseline versus in the flexible system (pertaining to Figure 4D). Logistic mixed effects regression models were performed on the full cohort (n=65 students) using 3 dichotomization strategies for low and high alarm-driven waking.

Alarm-driven waking (1 = low, 0 = high) dichotomized at 93% of schooldays (median)

Random effects			Variance	SD	
Student	(Intercept)		2.09	1.46	
Fixed effects	OR	95%CI lower	95%CI upper	z-value	p-value
(Intercept)	1.50	0.652	3.46	0.954	0.340
School start system (1 = flexible, 0 = baseline)	0.689	0.292	1.62	-0.853	0.394
Gender (1 = male, 0 = female)	0.306	0.089	1.05	-1.88	0.060

Alarm-driven waking (1 = low, 0 = high) dichotomized at 85% of schooldays (1st quartile)

Random effects			Variance	SD	
Student	(Intercept)		2.22	1.49	
Fixed effects	OR	95%CI lower	95%CI upper	z-value	p-value
(Intercept)	0.254	0.089	0.725	-2.56	0.011*
School start system (1 = flexible, 0 = baseline)	1.12	0.442	2.83	0.236	0.813
Gender (1 = male, 0 = female)	0.501	0.136	1.85	-1.03	0.301

Alarm-driven waking (1 = low, 0 = high) dichotomized at <85% and 100% of schooldays

Random effects			Variance	SD	
Student	(Intercept)		8.59	2.93	
Fixed effects	OR	95%CI lower	95%CI upper	z-value	p-value
(Intercept)	0.583	0.101	3.35	-0.605	0.545
School start system (1 = flexible, 0 = baseline)	0.861	0.198	3.73	-0.201	0.841
Gender (1 = male, 0 = female)	0.149	0.005	4.42	-1.101	0.271

SD, standard deviation; CI, confidence interval; OR, odds ratio

3

Project 2

“One year later: longitudinal effects of flexible school start times on teenage sleep, psychological benefits, and academic grades”

One year later: longitudinal effects of flexible school start times on teenage sleep, psychological benefits, and academic grades

Authors:

Anna M. Biller[#1,2], Carmen Molenda[#1], Giulia Zerbini[1,3], Fabian Obster[4], Christian Förtsch[5], Till Roenneberg[1] & Eva C. Winnebeck[1*]

[#]shared first authorship

*corresponding author

Affiliations:

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

[2] Graduate School of Systemic Neurosciences, LMU, Germany

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

[4] Department of Statistics, Statistical Consulting Unit, Ludwig Maximilian University Munich, Munich, Germany

[5] Biology Education, Faculty of Biology, Ludwig Maximilian University Munich, Munich, Germany

Contact information:

eva.winnebeck@med.uni-muenchen.de

Institute of Medical Psychology, Goethestrasse 31, 81373 Munich, Germany

Author contributions (CRedIT Taxonomy):

Conceptualisation: TR, ECW

Methodology: ECW, AMB, CM

Investigation: AMB, CM, ECW

Data curation: AMB, CM

Formal analysis: AMB, CM, FO, ECW

Validation: ECW, AMB, CM, FO, GZ

Supervision: ECW

Visualization: AMB, CM, ECW

Writing – original draft: AMB, ECW

Writing – review and editing: AMB, ECW, CM, GZ, FO, CF

Abstract

Early school times fundamentally clash with the late sleep of teenagers. This mismatch results in chronic sleep deprivation, which poses acute and long-term health risks and impairs students' learning and career prospects. Despite conclusive evidence that delaying school times has immediate benefits for sleep, the medium and long-term effects on sleep and academic achievement are unresolved due to a shortage of longitudinal data, short follow-up times and the many factors influencing sleep and academic grade trajectories. Here, we studied whether a flexible school start system, with the daily choice of an 8AM or 08:50AM-start, allowed high school students to improve their sleep, psychological functioning and academic grades in a longitudinal pre-post design over up to 1.5 years. Based on 2 waves with ≥6 weeks of daily sleep diary, we found that students maintained their 1-hour-sleep gain on later schooldays longitudinally (n=28) and cross-sectionally (n=79). Notably, girls were particularly successful in keeping early sleep onsets despite later sleep offsets. Students also reported psychological benefits (n=93). However, our regression analyses of ≤16,724 official grades over 4 year (n=65-157) detected no meaningful grade improvements in the flexible system *per se* or with sleep improvements – academic quarter, discipline or grade level had a greater, more systematic effect. Our findings thus contradict high-held hopes of large grade improvements from later school starts – at least at the ‘dose’ received in our sample. Importantly, students may nonetheless enjoy cognitive improvements alongside their psychological and sleep gains that simply do not translate into detectable grade changes.

Keywords: *sleep, adolescence, school start time, grades, academic performance*

Significance statement

Teenage sleep becomes progressively later during adolescence but school starts do not accommodate this shifted sleep window. This mismatch results in chronic sleep deprivation in teenagers worldwide, which is a pervasive public health concern. Since sleep, well-being, and performance are tightly linked, there is the strong expectation that counteracting sleep deprivation with later school starts results in improved learning and thus better academic grades. Although our students slept persistently longer and felt better when school started later, our results do not support that improved sleep leads to detectable grade changes within 1.5 years. Importantly, this does not preclude that better sleep and well-being facilitate better learning but suggests that grades are suboptimal to measure later school start effects on performance.

Introduction

Teenagers around the world are chronically sleep deprived because their late sleep timing often clashes with early school starts forcing them to get up long before their sleep has come to a natural end. Sleep is timed progressively later during adolescence because teenagers' internal circadian phase (chronotype) markedly delays[1–3]. At the same time, sleep pressure (the homeostatic load) accumulates more slowly over the day compared to adults or younger children, making teenagers less tired in the evening[4,5]. These biological tendencies are exacerbated by non-biological factors, such as academic pressure or cultural influences to stay up late[6,7]. Evening activities then lead to longer exposure to artificial light at night which increases alertness[8–10] and further delays circadian rhythms resulting in later sleep timings. Consequently, many students do not get enough sleep during the school week and compensate their sleep loss by oversleeping on weekends. This is often accompanied by a delay of sleep timing on free days - a phenomenon called "social jetlag"[11]. Yet, even with these weekend lie-ins, most teenagers do not achieve weekly sleep durations of at least 8 hours each night[e.g., 12,13], the recommended minimum sleep amount at this age[14].

The consequences of short sleep are numerous biological and psychological health compromises. In the long-term, chronic sleep deprivation has been linked to metabolic, cardiovascular, and inflammatory diseases[15,16], to depressed mood and worsened emotional regulation[17–19], as well as substance use[20,21]. Social jetlag, too, has been described to increase the risk for metabolic syndrome and obesity[22–24]. Prioritising other activities over sleep is short-sighted not only in terms of health, but especially academic performance: sleep deprivation has been related to absenteeism and tardiness[25], reduced participation and learning in class[26], worsened mood and emotional regulation[19], compromised emotional intelligence and constructive[27] and creative thinking skills[28,29], verbal fluency[30,31], and decreased academic performance[32].

The obvious solution, to simply delay school start times, has gained much scientific and public attention over the past decades. Positive associations were found for sleep and sleep quality, daytime sleepiness, concentration and attention in class, absenteeism and tardiness, and even motor vehicle accidents[33–37]. Nonetheless, policy-uptake is still rare, also invoking the low evidence level of the findings[38,39] as a reason. Indeed, the vast majority of studies used a cross-sectional design, which does not allow to track individual changes over time and is prone to cohort effects if not randomized or very carefully adjusted[35,40]. Double-blinding, the gold standard in terms of evidence level, is, of course, inherently unfeasible in this context and it seems almost impossible to convince schools to participate in randomization[41]. Although there are some real-life settings, such as in Uruguay or Argentina, where students are randomly assigned to morning, middle, and afternoon school shifts[42,43], this is not the case in most other countries around the world. The few longitudinal studies that exist often covered ≤6 months in their follow ups[35] (but see [44–48]), and are thus prone to seasonal confounding. Furthermore, sleep, mood, and performance have often been assessed via one-off questionnaires, while continuous sleep recordings via daily sleep diaries and especially objective actimetry measures are scarce[35,37,46,49–52]. One notable exception is a recent study by Widome and colleagues who followed students over two years and found persisting extended sleep durations (measured with one week of actimetry) in students from schools who delayed bell times compared to students in schools which did not change[48]. The evidence concerning effects on grades has similar caveats: Although much furor has been caused by studies suggesting grade improvements[51], most of these lacked basic statistical adjustments essential for the mainly cross-sectional data and/or used coarse or single grade measures. Robust, adjusted effects were only found based on very large samples with rather small effect sizes [43,53–56]. Nonetheless, due to the severity of the problem of teenage sleep deprivation and the many positive indications, clear recommendations that later school start times are beneficial for students' performance have been formulated[57,58] but may run the danger of raising false hopes in parents and teachers about resulting grade improvements.

To address the need for longitudinal studies and add high-quality data on academic performance, we investigated changes in sleep, psychological benefits, and official grades in a high school in Germany which permanently switched from a fixed start at 8AM to a flexible school start. Senior students could now choose *daily* whether to attend school at 8AM or skip the first class and start at 8:50AM. We used daily sleep diaries (and actimetry in the first year, see[59]) to monitor sleep in detail for several weeks during baseline, immediately after the change, and again after one year at the same photoperiod to circumvent the pitfalls of one-off questionnaires and optimally control for seasonal effects. Did students maintain their sleep extension of 1 hour on days with later starts[59] also after one year? Or did they adjust to the flexible system and delay their sleep times? At the end of the study period, we also retrospectively surveyed subjective wellbeing and psychological functioning on days with early versus later starts. To examine effects of the flexible system on academic grades in various disciplines, we additionally analysed official, quarterly grades across four years. With 2.5 years of data prior and 1.5 years after the introduction of the flexible school start, we could control for important confounders and address trends and complex interactions which started long before the system was changed.

Methods and Materials

Study Site

The study took part at the Gymnasium Alsdorf (50° 53' N, 6° 10' E), a high school in a town of ~45.000 residents in the West of Germany. A gymnasium is the most academic of secondary schools in Germany and grants access to higher education after 8-9 years of study and successful completion of the final exam. The school received the German School Award in 2013 for its innovative teaching[60]. It follows an educational system called "Dalton plan" that incorporates daily self-study periods called "Dalton hours" during which students work through a personal 5-week curriculum with a teacher and on a subject of their own choice.

Change in School Start Times

The school changed permanently from a fixed start ("conventional system") to a flexible start ("flexible system") for senior students (grades 10th-12th) on February 1st, 2016. In the conventional system, the first period started at 8AM. On a median of 1 day/week, depending on students' individual timetables, classes started with the second period at 9AM.

In the flexible system, one of the two daily self-study periods was advanced into the first period (lasting 08:00-08:45AM) and made optional to attend for senior students (for an example timetable see[59]). Senior students could thus choose daily whether to start at 8AM with the first self-study period or skip it and start at 08:50AM instead (called "9AM"). On a median of 1 day/fortnight, students also had a scheduled free second period (08:50-09:50AM), i.e. the chance to turn the 08:50-start into an 10:15-start when skipping the first period (">9AM"). In our analyses, we grouped these two later school starts into "≥9AM-days" and compared those with 8AM-days.

Students had to make up for the skipped first periods throughout the week, using gap periods or adding study time after their last classes (up to the official school closing at 4:15 PM). To be able to start later on all 5 schooldays/week, most students had to make use of both options since their individual schedules did not provide 5 gap periods and 5 early class ends per week.

Study Design

Data were collected in two waves that were exactly one year apart (Fig. 1a). Baseline data collection (=t_0) took place in winter 2016, covering 3 weeks in January (Jan 8th to 31st, 2016) in the conventional system with mainly 8AM-starts. This was followed by wave 1 data collection for 6 weeks (Feb 1st to Mar 14th, 2016) in the flexible system right after its introduction on Feb 1st, 2016 (=t_1). For the follow-up study (wave 2) we chose the matching photoperiod and time of $t_{1,}$ lasting from Feb 2nd to Mar 20th, 2017 (= t_2). As the school had remained in the flexible system ever since the introduction, no second baseline just before t_2 was carried out.

The holiday periods over carnival between February 4th-9th, 2016 and February, 23rd-28th 2017 were excluded from the analyses.

Participants

Written informed consent was obtained from all participants (or their parents/guardians if <18y). The study was conducted according to the Declaration of Helsinki and approved by the school board, the parent-teacher association, the school's student association and the ethics committee of the Medical Faculty of the LMU Munich (#774-16). We used opportunity sampling without specific exclusion criteria. In the first year (t_0+t_1), 113 (45%) out of 253 possible senior students attending 10th-12th grade (14-19 years) signed up, 83 (73%) students provided some data (response rate), of which 65 (70%) passed our minimal quantity and quality filter criteria (cohort 1). In the second year (t_2), 162 (71%) out of 227 possible students signed up, 137 (85%) provided data (response rate), of which 105 (77%) passed the minimal filter (cohort 2). Across both years, 33 students passed the minimal filter, hence forming the longitudinal cohort. To determine the longitudinal attrition rate, one needs to note that of the 65 students in cohort 1, 16 students graduated after t_1 and hence could not participate at t_2 (scheduled attrition rate of 34%). Of the 49 students that could have partaken again in t_2, 16 provided no or insufficient data at t_2 (attrition rate of 33%). Differences in baseline characteristics between the 33 and 16 students were tested and not significant (chronotype, social jetlag, gender, grade level; all p>0.05), except for age (t(47)=-2.933, p=0.005) with the missing students on average 0.8 years older.

Minimal filter criteria were: i) sleep information for ≥5 schooldays and ≥3 weekend days at each time point and ii) congruent, plausible data (more detailed information in[59]). For 8AM or ≥9AM-start

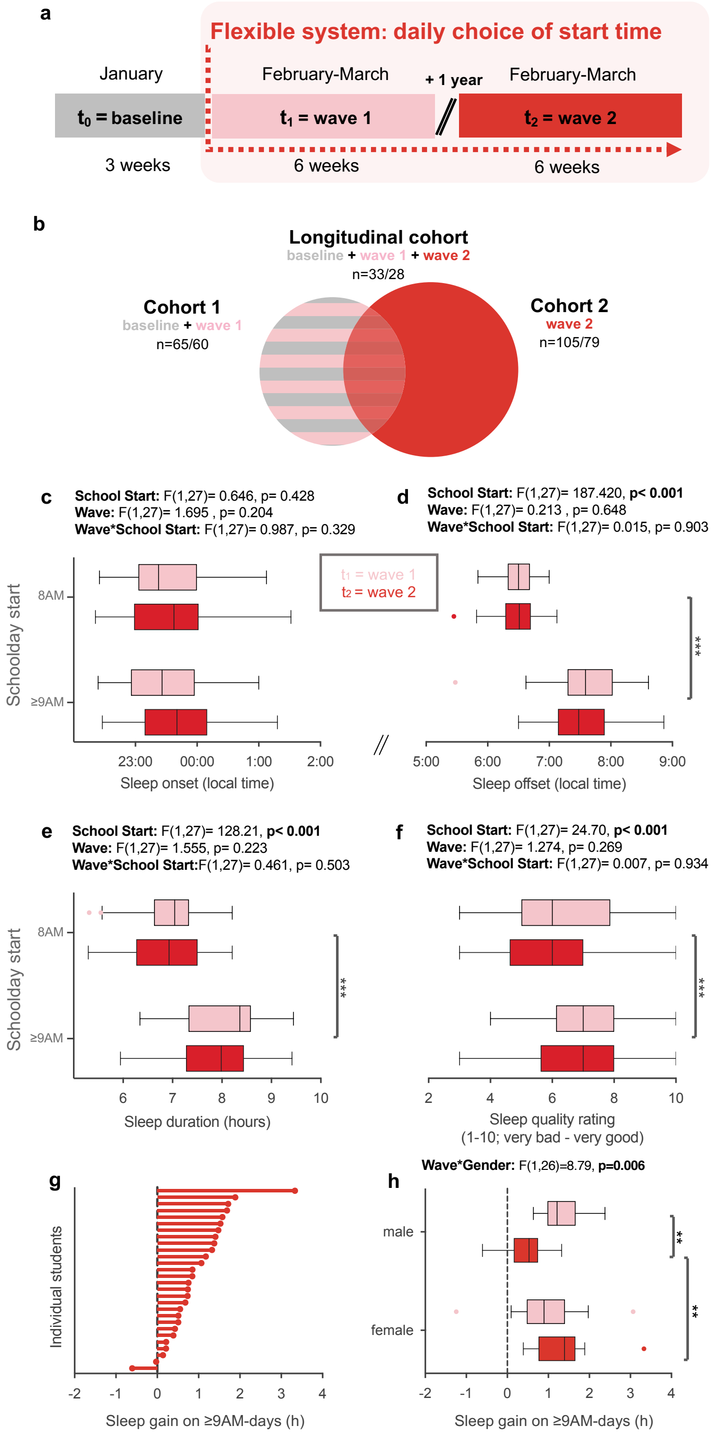

Fig. 1 | Comparison of sleep parameters between 8AM-days and ≥9AM-days in the flexible system.

a, Schematic of longitudinal study design including one baseline assessment and 2 waves in the flexible system. **b,** Schematic of study cohorts and sample sizes**. c-h,** Sleep parameters from the longitudinal cohort (n=28) comparing 8AM and ≥9AM-days as well as wave 1 (t_1, light red) and wave 2 (t_2, dark red) intra-individually. **c,** Average sleep onset, **d,** offset, **e,** duration, and **f,** quality on 8AM versus ≥9AM-days in the flexible system across waves. Results of two-way repeated measures ANOVA with the within-subject factors school start (8AM/≥9AM) and wave (wave 1/2) are reported above each graph. Brackets indicate statistically significant post-hoc comparisons. **g-h,** Sleep gain on ≥9AM-days as the average absolute difference in sleep duration between 8AM and ≥9AM-days. Positive values mean longer sleep on ≥9AM-days. **g,** Average sleep gain on ≥9AM-days during wave 2 for each participant. **h**, Average sleep gain on ≥9AM-days during both waves comparing female and male participants. Results of two-way mixed ANOVA with the between-subject factor gender (female/male) and the within-subject factor wave (wave 1/2). Given the significant interaction effect, main effects are not reported, instead subsequent post-hoc pairwise comparisons are indicated. All boxplots are Tukey boxplots.
*p<0.05, **p<0.01, ***p<0.001.

comparisons, we additionally filtered for at least two 8AM-days and at least two ≥9AM-days per person. After this additional filter, a total of 60 participants remained in cohort 1, 79 in cohort 2, and 28 in the longitudinal cohort. All students from the longitudinal cohort were granted promotion to the next grade level from wave 1 to wave 2.

Outcome measures

Sleep Diary

We used a daily sleep diary (provided online via LimeSurvey.org) based on the µMCTQ[61] (a short version of the Munich Chronotype Questionnaire) and adapted it to a German student population by changing the formal you ("Sie") to the informal you ("Du") and work days to schooldays. Students provided sleep onset (note: not bedtime) and offset (wake time) of their past night's sleep, whether they were woken by their alarm clock (yes/no), the type of day they woke up (schoolday or free day), when they started school (8AM, 9AM or >9AM), and their subjective sleep quality (rated on a 10-point-Likert scale from 1="very bad" to 10="very good"). The questionnaire did not cover any naps during the day. Although daily population of the online sleep diary was encouraged, students could fill in data in retrospect also if they had missed a day or more (they reported to have kept an offline log from which they copied their sleep timings). For more details see[59].

Survey

We developed a 17-item paper-pencil survey about the flexible system, which was distributed at the end of wave 2 and filled out by ~90% of cohort 2. Because some students did not answer all questions on the survey, the sample size ranged from 91 to 93 depending on the item. The first 7 items of the survey asked whether i) students were satisfied with the flexible system (yes/no), ii) they would rather have the old system with fixed school starts back (yes/no), iii) it was difficult for them to go to school at 8AM (never/most of the time/always), iv) it was easier to go to school at 9AM compared to 8AM (never/most of the time /always), v) how often (0 days/1-2 days/3-4 days/5 days) and vi) on which days of the week they attended the first period at 8AM (Mo/Tu/We/Th/Fr), and vii) reasons for starting school at 8AM. Answer options for vii) were to mark at least one of nine alternatives (easier to study/easier to get to school/additional study time/friends/specific teacher/specific subject/fulfill self-study quota/parents/late school end) and/or to name other reasons.

The last 10 items asked for ratings on 8AM versus ≥9AM-days. Questions were about i) sleep duration (h), ii) sleep quality (1=bad, 5=good), iii) number of schooldays with alarm-driven waking (0-5 days), iv) how tired the students felt (1=not at all, 5=very), v) ability to concentrate in class (1=bad, 5=good), vi) ability to study at home after school (1=bad, 5=good), vii) motivation to actively take part in class (1=not at all, 5=very), viii) how well they remembered new class content (1=not at all, 5=very), and ix) attitude towards school (1=negative, 5=positive). Items ii) and iv)-ix) were scored on a Five-point Likert scale.

Academic Grades

The school registry provided official quarterly grades obtained between summer 2013 and summer 2017. Of the 170 students from cohort 1 and 2 qualifying for analysis, 13 students had grades missing, thus resulting in a maximum sample of 157 students for the grade analyses. For the majority of these students (62%) grade data span 2.5 years in the conventional and 1.5 years in the flexible system; for those in grade level 10 at wave 2 (18%) it was 3 and 1 years, and for those at grade level 12 at wave 1 (15%) it was 2.5 and 0.5 years. The grades were provided for all academic subjects taken by a student, but we only included subjects in our analyses that most students took and clustered them into three disciplines: Sciences (Biology, Chemistry, Maths, Physics, Natural Sciences), Social Sciences (Geography, History), and Languages (English, German, Spanish, French, Latin). Provided grades were averages per academic quarter per academic subject over a mixture of written and oral examinations, course work and participation in class. The school year lasted from the end of August to mid-July divided into the following quarters: quarter 1 until end of October, quarter 2 until third week of January, quarter 3 until third week of April and quarter 4 until first week of July.

In grade levels 7-10, the grading scale ranged from 1 (best) to 6 (worst) with grades ≥4 considered passing grades. This scale was additionally broken down into plus (+) and minus (-) for all but grade 6. In grade levels 11 and 12, the scale ranged from 0 (worst) to 15 (best) with ≥4 considered passing. Both scales were combined by transforming the 1-6 scale to a 0-15 scale based on its finer plus/minus system and then transformed to a more universal 0%-100% scale.

Data Analysis

Analyses were performed in SPSS Statistics (IBM, versions 24 and 25), R (versions 3.6.1 and 3.6.3) and R studio (versions 1.1.463, 1.2.1335 and 1.2.5042). Graphs were produced using Graph Pad Prism (versions 6 and 7) and *ggplot2*[62] R.

Sleep Data

Daily sleep data from diaries were aggregated as mean per person for 10 conditions: at t_0 (baseline conventional system) for schooldays and weekends; and at t_1 and t_2 (flexible system wave 1 and wave 2) for schooldays, weekends, 8AM-days, and ≥9AM-days. From these aggregates, we derived the following variables as per equations below for t_0-t_2: average daily sleep duration during the week (SD_{week}), chronotype as midsleep on free days (MSF) corrected for oversleep (MSF_{sc}), and social jetlag (SJL); for t_1 and t_2 only: absolute difference between ≥9AM-days and 8AM-days for variables of interest (DELTA x), frequency of alarm-driven waking, and frequency of ≥9AM-starts. For the linear mixed models 3a-d of the grade analyses, we additionally calculated the absolute differences between t_0 and t_1 (i.e., from baseline to the flexible system during wave 1) for variables of interest (X change).

$$\mathrm{SD_{week}} = (\mathrm{SD_{schooldays}} * 5 + \mathrm{SD_{free\ days}} * 2)/7$$

$$\mathrm{MSW} = \mathrm{SleepOnset_{schooldays}} + \frac{1}{2}\mathrm{SD_{schooldays}}$$

$$\mathrm{MSF} = \mathrm{SleepOnset_{free\ days}} + \frac{1}{2}\mathrm{SD_{free\ days}}$$

$$\mathrm{MSF}_{sc} = \mathrm{SleepOnset_{free\ days}} + \frac{1}{2}\mathrm{SD_{week}}$$

$$\mathrm{SJL} = \mathrm{MSF} - \mathrm{MSW}$$

$$\mathrm{DELTA\ x} = \mathrm{x_{9AM\text{-}days}} - \mathrm{x_{8AM\text{-}days}}$$

$$\text{Frequency of alarm-driven waking} = \left(\mathrm{n_{alarm\text{-}driven\ waking_{flex}}} / \mathrm{n_{schoolday\text{-}entries_{flex}}}\right) * 100$$

$$\text{Frequency of} \geq \text{9AM-starts} = (\mathrm{n}_{\text{9AM-starts}_{\text{flex}}} / \mathrm{n}_{\text{schoolday-entries}_{\text{flex}}}) * 100$$

$$\text{X change} = \mathrm{x}_{\mathrm{t1}} - \mathrm{x}_{\mathrm{t0}}$$

Statistical analyses

Unless indicated otherwise, descriptive statistics are reported as mean ± standard deviation and test statistics are abbreviated as follows: *t*, t-test; *Z*, Wilcoxon signed-rank test; F, ANOVA; *r*, Pearson correlation; *rho*, Spearman rank correlation; b, unstandardized coefficient of linear regression or linear mixed models; $\mathrm{b}_{\text{flex*change}}$, unstandardized coefficient of the interaction of linear mixed models; *p*, significance level. Significant levels were set to *p*<0.05 for all statistical analyses. All data were tested on normality (histograms, QQ plots, Shapiro-Wilk's test) and sphericity (Mauchley's test; Greenhouse-Geisser corrections if violated). If normality was violated, non-parametric tests were used (Spearman rank correlations and Wilcoxon signed rank test; no non-parametric test was used if normality was violated for ANOVA analysis since violations were marginal. Group difference for attrition groups were tested via independent t-test (chronotype, social jetlag, age) or Chi squared test (gender, class).

For sleep variables in the longitudinal cohort, we performed 1-way repeated measures ANOVAs with the factor time point ($t_0/t_1/t_2$), 2-way repeated measures ANOVAs with the factors wave (t_1/t_2) and school start (8AM/≥9AM-days), and with the factors time point ($t_0/t_1/t_2$) and day (schooldays/weekend). For sleep variables in cohort 2, paired t-tests (two-sided) were run for school start (8AM/≥9AM-days) and days (schooldays/weekend), and Wilcoxon signed rank test for sleep quality and survey items. Gender differences in sleep variables were assessed via 2-way mixed ANOVA with gender (female/male) and wave (t_1/t_2), and via linear regression (including the covariates grade level, chronotype and frequency of ≥9AM-starts) for DELTA sleep duration/onset/offset using the *nlme* package in R[63]. ANOVA results are presented above each graph (main effects and interaction). If the main interaction was significant, we reported simple main effects in the main text or in Tab. S2. Post hoc tests were carried out using Bonferroni corrections if interactions were significant.

Pearson and Spearman rank correlations were run for chronotype and frequency of ≥9AM-starts with DELTA sleep duration respectively. Three Tukey outliers were identified for frequency of ≥9AM-starts and DELTA sleep duration during wave 1 (rho=-5.45, p=0.003 before removal) and were subsequently removed (rho= -0.37, p=0.064 after removal). Frequency of alarm driven waking was analyzed using logistic regression (Fig. S1; *lme4* package R[64]). Due to a large ceiling effect, we dichotomized this variable based on a median split at 100%-use (<100%: "less use") and accommodated the repeated measures nature of the data by including ID as a random effect. Gender was included as covariate (males were woken more often by an alarm than females in the flexible system) but gender did not reach statistical significance.

For simple grade analyses comparing aggregated grade point averages in the conventional vs the flexible system, a two-sided paired t-test was used. For more complex grade analyses, we calculated linear mixed-effects regression models (*lme4* and *lmer test* package[64,65] in R). In total, 4 different models (plus submodels) were performed to answer different questions based on different fixed effects, interaction terms and subcohorts (see overview Tab. 2). Student ID was added as random effect to all models to incorporate unsystematic differences between individuals. In all models, the outcome (dependent variable) was quarterly grades per discipline per student; the fixed effects (independent variables) were system (conventional/flexible), gender (female/male), grade level (7-12), academic quarter (1-4), and academic discipline (Sciences/Social Sciences/Languages), all entered as categorical variables. Model 1 additionally included interaction terms between discipline and gender to assess general grade influences, model 2 included interaction terms between school start system and gender, and system and discipline to assess system effects per discipline and gender. In models 3, we included one of the sleep-change variables (see above; mean-centred) as additional fixed effects, each in interaction with system (conventional/flexible): chronotype change (model 3a), sleep duration on schooldays change (model 3b), social jetlag change (model 3c) or frequency of ≥9AM-starts (model 3d). In model 4, we included instead the absolute value of chronotype, sleep duration on schooldays, social jetlag, and frequency of ≥9AM-starts for the flexible system only (from t_2 if available, else from t_1 to maximize the sample size) As chronotype, sleep duration on schooldays, social jetlag, and frequency of ≥9AM-starts were prone to collinearity, we first assessed their correlations before adding them into the models (Fig. S4). Only chronotype and social jetlag were highly correlated (rho = 0.65, p<.001; Fig. S4), and results from all models including just one of these variables each (4a-d) were essentially similar to model 4e which includes all sleep variables together (Tab. S7). The variance inflation factor (*car* package in R[66]) also indicated no problematic collinearity for model 4e. Marginal means of model estimates were calculated using *emmeans* in R[67] in models where interactions were significant. All linear mixed models were visualised in tables using the *sjPlot* and *sjmisc* packages [68,69] and in figures as marginal means via the *ggeffects* package[83][70] in R. Simple contrast results from interactions in linear mixed models were averaged over the levels of system or gender (depending on the model), grade level, and quarter; degrees of freedom method used was Kenward-Rogers.

Results

The flexible system, established and retained at the school since 2016, provides flexibility on the school start time on a daily basis. This means that every single senior student decides each day if they attend the first period at 8AM or if they skip the first period and start at 08:50AM instead. In the rare case of a scheduled free second period, skipping the first period leads to a 10:15AM-start. Non-attended first periods have to be made up for within the same week during free periods or after classes.

Students participating in our study kept daily sleep diaries for three weeks in the conventional system (t_0; baseline), for six weeks after the introduction of the flexible system (t_1; wave 1), and after exactly one year for another six weeks (t_2; wave 2) in the same photoperiod as wave 1 (Fig. 1a). We allowed students to take part during all time points irrespective of their participation beforehand, so our study eventually consisted of three cohorts (Tab. 1 and Fig. 1b) : (i) cohort 1 provided sleep data at t_0 and t_1 (n=60-65), (ii) cohort 2 provided sleep and survey data only at t_2 (n=79-105), and (iii) the longitudinal cohort provided sleep data throughout from t_0-t_2 (n=28-33; Tab. 1 and Fig. 1b). The samples sizes within each cohort varied due to different filters employed for different analysis questions (see methods). Students of all cohorts (n=63-157) contributed their quarterly grades (≤16,724) through official school records. Please see Table 1 for detailed cohort characteristics.

Notably, our participants accumulated fewer late starts per week (“≥9AM-days”) than expected. We had observed this for cohort 1[59], but now saw this confirmed in cohort 2, where participants (n=105) chose to skip the first period only on a median of 24% of their schooldays (IQR: 10-47), which equates to 1.2 later starts per 5-day school week (Tab. 1). Similarly, the longitudinal cohort (n=33) had a median frequency of late starts (“≥9AM-use”) of 39% (20-51) and 22% (11-46) during wave 1 and 2, with no systematic difference between the waves (Z=-1.653, p=0.098). Importantly, ≥9AM-use varied drastically between individual participants from 0% to 100% of their schooldays, with 8:50AM-starts making up the majority of later starts per person and 10:15AM-starts only 25% (median, IQR: 6.3-60).

Sleep on days with later school starts

In the following, we present analyses *within* the flexible system comparing days with early school starts (“8AM-days”) to those with later starts (“≥9AM-days”).

Student slept longer and better on days with later school starts – an improvement persisting over one year

How was students’ sleep altered by later school start times in the flexible system over one year? We showed previously that right after the introduction of the flexible system students from cohort 1 slept about one hour longer on ≥9AM-days by maintaining their sleep onset but delaying their sleep offset[59]. After one year, we found the same for cohort 2 and, importantly, also in the longitudinal cohort across both waves.

Repeated measures ANOVAs in the longitudinal cohort (n=28) showed that sleep onsets did not differ with start time or wave (Fig. 1c), whereas sleep offsets were 61 min (± 47) later on average, and students hence slept 62 min (± 47) longer on ≥9AM-days compared to 8AM-days across both waves (Fig. 1d-e, full statistics in figures). Findings from cohort 2 (n=79) tally with this pattern: sleep onsets on 8AM and ≥9AM-days were comparable (t[78]=-1.87, p=0.065), while wake up times were significantly later on ≥9AM-days (t[78]=-19.75, p<0.001), which resulted in 60 min longer sleep durations on those days (t[78]=-10.83, p<0.001). This large sleep gain likely reflects the combination of 08:50 and 10:15-starts at around 4:1.

Furthermore, subjective sleep quality was improved on ≥9AM-days by 1 point on a 10-point Likert scale for cohort 1[59] and cohort 2 (n=79, Z=-5.874, p<0.001), and also longitudinally across waves (n=28, Fig. 1f). In addition, the extensive use of alarm clocks remained slightly reduced on ≥9AM-days also one year into the system (Fig. S1). Just as in cohort 1[59], the odds for less alarm-driven waking were increased in cohort 2 (n=79, OR = 1.9, 95% CI = 1.3-4.1) and showed a similar qualitative pattern also in the longitudinal cohort (n=28; Fig. S1), demonstrating that a natural waking was more likely when school started later.

Tab. 1 | Composition of study cohorts. Displayed are cohort characteristics after standard filter criteria. An additional filter (see methods) was applied for comparisons between 8AM and ≥9AM-days which reduced cohort 1 to 60 students, the longitudinal to 28 students, and cohort 2 to 79 students. Abbreviations: n, number of individuals; SD, standard deviation; IQR, interquartile range; conv., conventional.

		Cohort 1	Longitudinal cohort		Cohort 2
Waves		wave 1	wave 1	wave 2	wave 2
Participants					
Total	n	65	33		105
Females	n (%)	40 (62%)	20 (60%)		73 (70%)
Grade level	n (%) per level 10th/11th/12th	26/23/16 (40/35/25%)	20/13/0 (60/40/0%)	0/20/13 (0/60/40%)	29/38/38 (28/36/36%)
Age	mean (SD, range)	16.5 (1.2, 14–19)	15.8 (0.9, 14-17)	16.9 (0.9, 15-18)	16.7 (1.1, 15-21)
Chronotype (MSF_{sc}; time in h)	mean (SD, range)	4.6 (0.9, 2.1–7.0)	4.3 (0.7, 2.1-5.9)	4.6 (0.9, 0.8-6.2)	4.7 (1.0, 0.2-8.6)
Social jetlag (h)	mean (SD, range)	1.8 (0.7, 0.3-3.8)	1.7 (0.6, 0.3-3.1)	1.9 (0.6, 0.5-3.3)	2.0 (0.8, 0.2-6.0)
Sleep duration (h)	mean (SD, range)	7.6 (0.8, 5.2-8.9)	7.7 (0.8, 5.2-8.8)	7.6 (0.7, 6.1-9.0)	7.7 (0.7, 6.1-9.3)
Proportion of schooldays with later starts					
≥9AM-use	median (IQR)	32% (19-55)	39% (20-51)	22% (11-46)	24% (10-47)
Students reaching ≥8 hours of sleep in the flexible system					
8AM-days	%	15.4%	12.0%	3.0%	6.8 %
≥9AM-days	%	50.0%	59.4%	45.2%	47.3%
Schooldays	%	18.5%	18.2%	9.1%	13.3%
Weekends	%	73.8%	84.8%	69.7%	74.3%
Academic grades per discipline on a scale from 0% (worst) – 100% (best)					
Languages	median (IQR)	53% (47-70)	–	–	53% (47-73)
Sciences	median (IQR)	60% (47-73)	–	–	67 % (53-73)
Social Sciences	median (IQR)	60% (53-73)	–	–	67% (53-73)
Number of students per outcome					
Sleep	8AM vs. ≥9AM-days	n=60	n=28		n=79
	conv. vs. flexible system	n=65	n=33		n=105
Psychological benefits					n=91-93
Academic grades		n=63-157			

Students reported profound improvements in cognitive and psychological parameters on later school days

To assess psychological benefits, we used survey data from the end of wave 2, which were provided by 90% of cohort 2. Students' subjective ratings of their sleep, cognition and well-being on 8AM-days compared to ≥9AM-days showed statistically significant improvement in all areas assessed (n=91-93; full statistics in Fig. 2). On days with later starts, students felt generally better, less tired during class, more motivated to actively take part in class, and were better able to concentrate. Students also reported a more positive attitude towards attending school and higher quality of self-study after school. Altogether this shows that students clearly preferred the late-start-option.

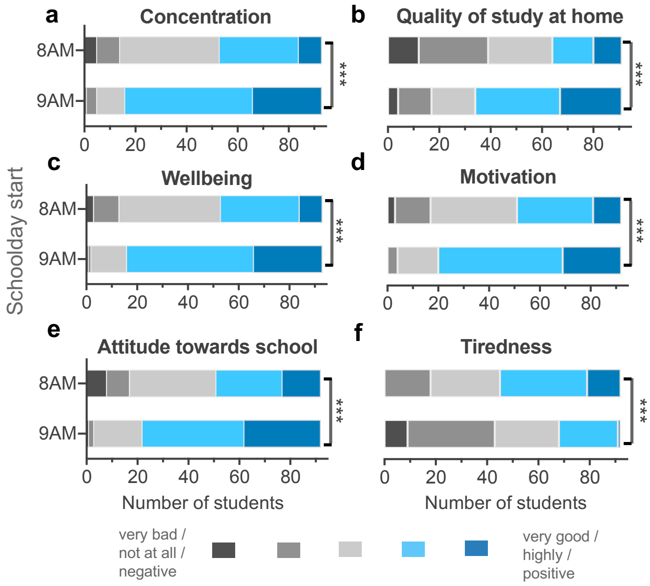

Fig. 2 | Comparison of subjective psychological benefits between 8AM-days and ≥9AM-days in the flexible system. Results from the survey at end of wave 2 asking cohort 2 for the following ratings: **a,** ability to concentrate during class (Z= -6.419, n=93), **b,** quality of study at home after school (Z= -6.055, n=91), **c,** general wellbeing (Z=-6.559, n=93), **d,** motivation to attend school (Z= -5.927, n=92), **e,** attitude towards school (Z= -5.896, n=92), and **f,** tiredness during class (Z=-5.419, n=92). Wilcoxon signed rank test.
*p<0.05, **p<0.01, ***p<0.001.

Girls maintained their sleep benefit from later school starts more than boys after one year in the flexible system

We wondered whether particular students benefitted more or less than others from later starts. Therefore, we assessed the relationship of chronotype, ≥9AM-use and gender with the core sleep benefit, the sleep gain on ≥9AM-days (the difference in sleep duration between ≥9AM- and 8AM-days). In the longitudinal cohort (n=28), 93% of students experienced a sleep gain on ≥9AM-days in both waves (Fig. 1g), so the sleep benefit was close to universal. Chronotype did not correlate with sleep gain (wave 1: r=-0.024, p=0.903; wave 2: r=-0.091, p=0.647), i.e. both early and late chronotypes appear to have benefitted equally from later starts (Fig. S2). We already observed this in cohort 1[59] and interpreted it as the consequence of the severe sleep deprivation in adolescent students afflicting even earlier chronotypes. Similarly, no matter how often the students attended school later, their sleep gain on ≥9AM-days was not systematically affected. Although there was a slight trend towards smaller gains with more frequent ≥9AM-use in wave 1 (rho= -0.37, p=0.064; 3 outliers removed), it was driven – just like in cohort 1 - by a few over-benefitting individuals with low ≥9AM-use and it was absent during wave 2 (rho=0.028, p=0.889; Fig. S2). In contrast, gender showed a clear effect on sleep gain after one year: both genders enjoyed similar sleep gains during wave 1, as also found in cohort 1[59], but boys clearly reduced their sleep gain during wave 2 from 1.3h (± 0.53) to 0.5h (± 0.53; detailed statistics in Fig. 1h). Follow-up analyses revealed that the reduced sleep gain in boys resulted from a delay in their sleep onsets on ≥9AM-days compared to 8AM-days, while their offset times were unaltered during wave 2 (n=28; Fig. S3, statistics in figure).

The bigger sample size of cohort 2 (n=79) allowed us to address the above relationships together in single regression models. Besides gender, chronotype and ≥9AM-use, we also included grade level (inherently incorporating age) as predictors modelling sleep gain, sleep onset delay and sleep offset delay (the differences in onset/offset between ≥9AM and 8AM-days; Tab. S1) to address again the reasons for the gender disparity. The regression results corroborated all observations from the longitudinal cohort showing that only gender had a significant influence on any of the outcomes, namely sleep gain and sleep onset delay (Tab. S1). Boys reduced their sleep gain on average by 0.52 h (b=-0.52, p=0.010), which was driven by a delay in their onset on ≥9AM-days by 0.53h (b=0.53, p<0.001), while their offset was unchanged (b=0.01, p=0.942). Sensitivity analyses indicated that this effect was not just driven by the longitudinal cohort comprising 35% of cohort 2. Taken together, while most inter-individual differences did not systematically influence sleep gains, boys showed a delay in sleep onset and thus displayed a smaller sleep gain on ≥9AM-days after one year in the flexible system.

Sleep in the flexible system versus baseline

Despite obvious improvements in sleep and subjective parameters on ≥9AM-days also after one year, it is essential to determine if these actually translated into better sleep in the flexible system overall. Based on our analyses of cohort 1[59], this was largely not the case during the first six weeks after the introduction of the flexible system. Most likely the limited ≥9AM-use in combination with occasional late starts during baseline reduced improvements by the flexible system compared to the conventional system. But did long-term effects emerge after one year?

Students did not extend their sleep in the flexible system overall

Analyses in the longitudinal cohort (n=33) revealed that students' sleep was not improved compared to baseline even after 1 year in the flexible system. Despite small delays in sleep offset on schooldays (Fig. 3a, detailed statistics in Fig. 3 and Tab. S2), sleep duration on schooldays and across the week were not significantly increased at t_1 or t_2 compared to t_0 (Fig. 3b). Students still only slept 7.6 h (± 0.65) on a daily average across the week (including weekend catch-up sleep) at t_2, a sleep duration below the recommended 8-10 h for this age group[71]. Students' chronotype remained expectedly late across all time points (Fig. 3c), and there was still a substantial difference between sleep timing on schooldays and weekends (Fig. 3a; Tab. S3 for similar results in cohort 2). Students' social jetlag, which quantifies this typical shifting between the 'schoolday-time zone' and the 'weekend-time zone', although reduced at t_1 by 30 min (± 0.62, p=0.002), was indistinguishable from baseline after one year (p=0.256; Fig. 3d).

So, the mild reduction in social jetlag experienced immediately after entering the system was lost later on, emphasizing that there was no widespread improvement in sleep under the low ≥9AM-use in the flexible system.

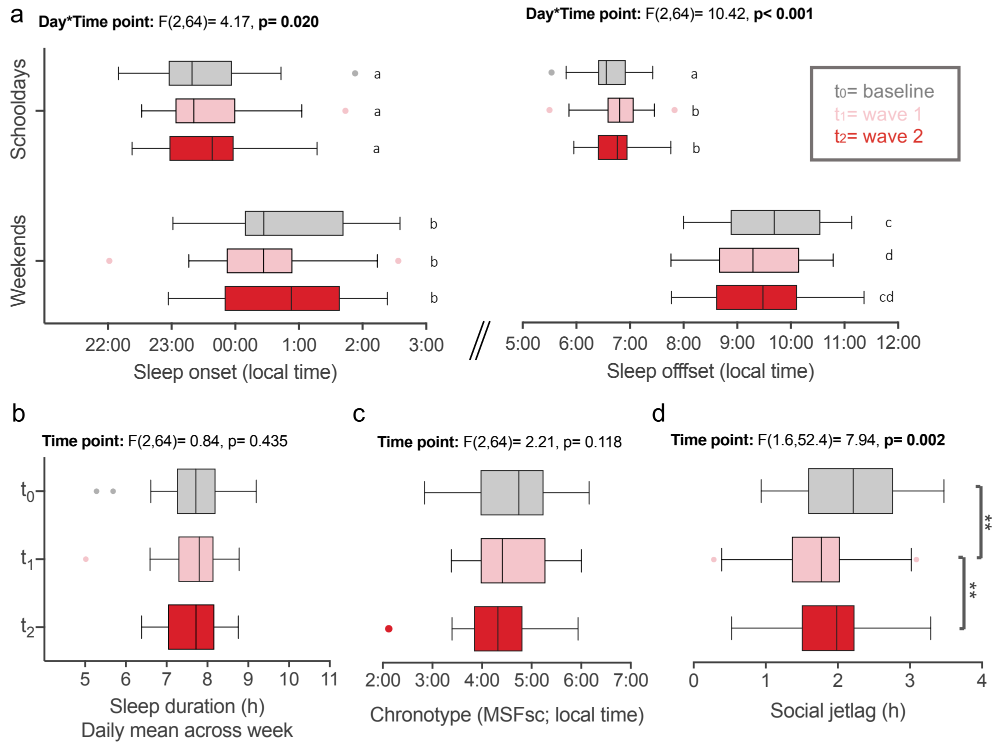

Fig. 3 | Comparison of sleep parameters across school start systems.
Sleep parameters from the longitudinal cohort (n=33) comparing the conventional start system at baseline (t_0, grey) with the flexible system during wave 1 (t_1, light red) and wave 2 (t_2, dark red). **a,** Average sleep onset and offset on schooldays and weekends. Results of two-way repeated measures ANOVAs with the factors day (schooldays/weekends) and time point (t_0/t_1/t_2) are provided. Given the significant interaction effect, main effects are not reported. Letters indicate results of post-hoc tests on simple contrasts, with data marked by different letters demonstrating significant differences. **b,** Average daily sleep duration across the week (weighted for 5 schooldays and 2 weekend days), **c,** average chronotype, **d,** average social jetlag. Results of one-way repeated measures ANOVAs across time points are presented above each graph. Brackets indicate statistically significant post-hoc comparisons. All boxplots are Tukey boxplots. *p<0.05, **p<0.01, ***p<0.001.

Do students receive better grades in the flexible system per se or with improved sleep?

In the second part of our study, we analysed the longitudinal development of official grades in our sample (all cohorts). Specifically, we investigated i) whether the change in school start system, and ii) any resulting individual sleep benefits allowed students to improve their grades, and iii) whether students with disadvantaging sleep and circadian characteristics (shorter sleep duration, later chronotype, higher social jetlag) exhibited lower academic grades overall. In total, we analysed 16,724 school-reported quarterly grades from 157 students (mean of 107 grades per student) that students received in 12 academic subjects over 2.5 years in the conventional system and 1.5 years in the flexible system (Tab. 2). Grades were provided by the school and subsequently transformed to a 0%-100% scale and grouped into three disciplines (Languages, Sciences, and Social Sciences).

Tab. 2 | Overview of linear mixed model analyses on official, quarterly grades.
Four different models (and various submodels) were calculated, each with a different aim and including appropriate predictors (fixed effects) and interaction terms. All models included ID as a random intercept to incorporate random inter-individual differences. Abbreviations: conv, conventional school start system; flex, flexible school start system.

	Model 1	Model 2	Model 3a-d	Model 4a-e
Outcome	Official grades (per quarter)	Official grades (per quarter)	Official grades (per quarter)	Official grades (per quarter)
Aim	General effects	System effects	Effect of sleep changes	Sleep effects in flexible system only
Fixed effects	System (conv/flex) Gender Grade level Academic quarter Academic discipline	System (conv/flex) Gender Grade level Academic quarter Academic discipline	System (conv/flex) Gender Grade level Academic quarter Academic discipline Change[a] in... a. ... chronotype b. ... sleep duration c. ... social jetlag d. 9AM-use[b]	- Gender Grade level Academic quarter Academic discipline - a. Chronotype[c] b. Sleep duration[c,d] c. Social jetlag[c] d. 9AM-use e. all of the above
Interactions	Gender * Academic discipline	System * Academic discipline System * Gender	System * Chronotype change / Sleep duration change / Social jetlag change / ≥9AM-use	-
Random intercept	ID	ID	ID	ID
Cohort	Cohort 1 & 2	Cohort 1 & 2	Cohort 1	Cohort 1 & 2
N	157	157	63	129
Number of observations	16724	16724	6683	5111
Data span	4 years: 2.5 y conv & 1.5 y flex	4 years: 2.5 y conv & 1.5 y flex	4 years: 2.5 y conv & 1.5 y flex	1.5 years: only flex

[a] Change refers to the absolute difference between the respective variable at t_1 (flexible system wave 1) minus t_0 (baseline). Positive values indicate higher numbers at t_1.

[b] Since the exact frequency of 9AM-starts during baseline (t_0) is not known, 9AM-use was added as an absolute value rather than as the change from t_0 to t_1. Students attended school at ≥9AM at a median of 1 day per week in the conventional system.

[c] From t_2 if possible, else from t_1.

[d] Duration on schooldays.

School start system showed no systematic effect on academic grades overall

A simple comparison of overall grades yielded a small improvement in grade point average from 58.2% (± 2.1) in the conventional to 59.6% (±2.0) in the flexible system (Fig. 4a; t[154]=-2.15, p=0.033). However, attributing this improvement to the flexible system is likely unwarranted. From the educational literature it is known that grades are influenced by a multitude of factors, and comparisons that do not account for these can be misleading. We therefore applied linear mixed-effects regression models adjusting for potential confounders (Tab. 2). When incorporating gender, grade level (inherently including age), academic quarter, and discipline in addition to school start system in the analysis, the flexible system showed no systematic impact on students' grades (Fig. 4b; b= -0.10, p=0.815, Model 1, Tab. S4), i.e. the flexible system was not associated with students receiving better or worse grades overall in our sample.

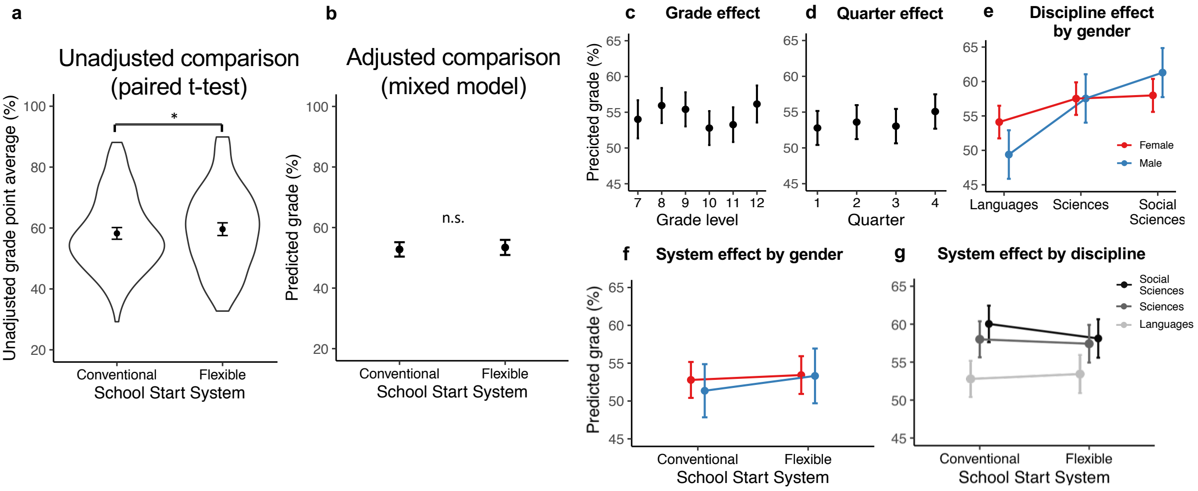

Fig. 4 | Longitudinal analysis of official quarterly grades - effects of school start system and general predictors. Quarterly grades (0%-100%) were sampled from cohort 1 and 2 across 12 academic subjects of 3 disciplines for 4 years i.e., for most students this was 2.5 years before and 1.5 years after the flexible school start was introduced (n=157 students; 16,724 grades; 107 grades per student on average). **a**, Simple, unadjusted comparison of average grades across all disciplines in the conventional and the flexible school start system via paired t-test (n_{ID}=157). Mean and 95% CI are indicated as well as the distribution of the underlying raw data (violin plots). The apparent increase in grades in the flexible system could not be confirmed in detailed analyses using linear mixed models. **b-g,** Visualization of mixed-model-determined influences on grades. Plots show marginal means from models 1 and 2 (Tab. S4), i.e. the estimated grade and 95% CI for the reference situation (female student, class level 10, quarter 1, languages, conventional system). Statistical significance is only indicated in the simple cases (b), results for more complex cases can be gleaned from the text and Tab. S4 and Tab. S5. **b,** Effect of school start system (model 1). **c**, Effect of grade level (model 1). **d**, Effect of academic quarter (model 1). **e**, Effect of academic discipline by gender (model 1). **f**, Effect of school start system by gender (model 2). **g,** Effect of school start system by academic discipline (model 2).

Grades varied systematically with grade level, academic quarter, discipline and gender

But what drives better grades in the unadjusted comparison if not the flexible system itself? The same factors that we adjusted for in the regressions also stood out as major predictors (Model 1, Tab. S4): Students in 12th grade (the last year in high school) did consistently better compared to their peers across all other grade levels (Fig. 4c) – a sort of "leavers effect" that has already been observed before[72]. Moreover, we found that students enjoyed a bump in grades in the last quarter of the school year with an estimated improvement of 2.3 percentage points compared to the first quarter (Fig. 4d; b=2.34, p<0.001; Model 1, Tab. S4). The combination of these two effects might explain the positive finding of the unadjusted comparison: the flexible system replaced the conventional system mid-year between quarter 2 and 3, so quarter 4 and higher grade levels were overrepresented in the flexible system, which the t-test could not account for.

The mixed models also revealed other strong systematic influences on grades in our sample. Firstly, we observed a clear difference between the disciplines: students performed generally best in Social Sciences, followed by Sciences and then Languages (Model 1, Tab. S4). Post-hoc tests (Fig. 4e) showed that these differences were highly significant for both genders (all p<0.001; post-hoc to Model 1, Tab. S5), except for girls' grades in Sciences and Social Sciences, which were indistinguishable (b=-0.47, p=0.3895; post-hoc to Model 1, Tab. S5).

Gender has been reported as another driving force for higher grades[73]. However, girls in our sample did not outperform boys overall (Model 2, Tab. S4 and Tab. S5). Girls were significantly better in Languages (Fig. 4e; b=4.72, p=0.0284; post-hoc to Model 1, Tab. S5), while boys surpassed them in the Social Sciences (b=-3.31, p=0.1269; post-hoc to Model 1, Tab. S5), and both genders did equally well in Sciences (b=-0.00, p=0.9915; post-hoc to Model 1, Tab. S5).

The flexible system was linked with subtle improvements in Languages and subtle drops in Social Sciences grades

Although we did not find evidence that the flexible system was linked with better grades overall (Model 1, see above), the flexible system might be linked with grade improvements in certain disciplines and genders. To assess this, we looked at the interaction between i) school start system and discipline, as well as ii) school start system and gender in a second model (Model 2; Tab. S4). Neither females nor males significantly improved their overall grades from the conventional to the flexible system (Fig. 4f; post-hoc to Model 2, Tab. S5). In terms of discipline effects, we found that grades in Social Sciences slightly dropped (b=1.26, p=0.0384; post-hoc to Model 2, Tab. S5), Science grades remained unchanged (b=-0.07, p=0.8849; post-hoc to Model 2, Tab. S5), and Language grades slightly improved (b=-1.30, p=0.0168, post-hoc to Model 2, Tab. S5) in the flexible system. Notably, these changes were subtle but reduced the grade differences between the academic disciplines (Fig. 4g, Tab. S5). These small changes in opposite directions likely explain the absence of a net effect of the flexible system on overall grades.

Improvements in chronotype, sleep duration, and social jetlag did not systematically improve grades

What was the role of sleep parameters on grade developments? We speculated that students who showed greater improvements in the flexible system (i.e., advanced chronotype, lengthened sleep duration, and lowered social jetlag) also received better grades in the flexible system. Thus, we computed changes in sleep from t_0 to t_1 based on the subpopulation of students with sleep parameters during these time points (n=63, ~cohort 1). Adding these parameters separately into a third model (Models 3a-c, Tab. S6, Fig. 5a), we found that neither changes in chronotype (flex*chronotype change: b=0.10, p=0.845) nor sleep duration (flex*sleep duration change: b= -0.77, p=0.352) were systematically associated with changes in grades. Surprisingly, however, students who increased their social jetlag in the flexible system obtained slightly better grades (flex*social jetlag change: b=1.28, p=0.027), which was contrary to our hypothesis. Therefore, our analyses in this subsample suggest that sleep improvements experienced immediately after transitioning to the flexible system did not result in detectable higher academic achievement.

If not linked to sleep improvements, were grades nonetheless linked with the choice of more later school starts? The results of Model 3d suggests that higher 9AM-use was associated with worse grades in the conventional system (b=-3.04, p=0.015), a link reversed partly – albeit not significantly – in the flexible system (flex*9AM-use: b=0.59, p=0.101; Fig. 5a and Tab. S6). Hence, students who made high use of the late-option in the flexible system were predominantly the lower-achievers, but they tended to profit at least slightly from more later starts.

Absolute effects of sleep characteristics (not their changes) on grades within the flexible system (Model 4, n=129 students) also showed no systematic influence on grades, independent of whether they were added separately into the model (4e) or together (4a-d; Fig. 5f, Tab. S7). However, higher 9AM-use was associated with lower grades (Model 4d: b=-2.12, p=0.022; combined Model 4e: b=-1.32, p=0.272) which tallies with results from Model 3d: it seemed that mainly lower-achieving students liked to use the 9AM-option.

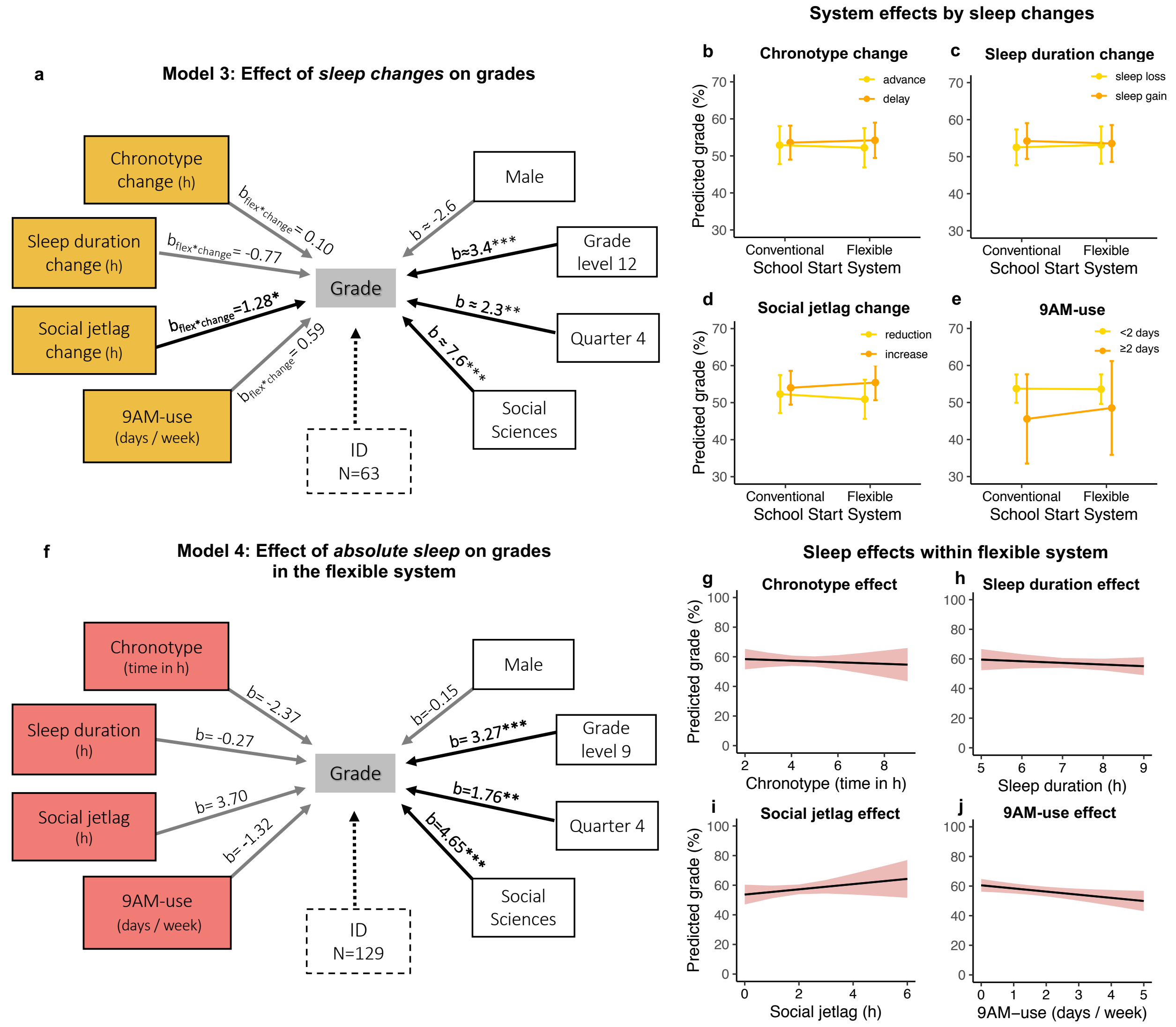

Fig. 5 | Longitudinal analysis of official quarterly grades - effects of sleep and 9AM-use.
Results from linear mixed model analyses of quarterly grades (0-100%) considering sleep variables as well as the frequency of ≥9AM-starts in suitable subcohorts (cohort 1 in model 3 and cohort 1 and 2 in model 4). **a,f,** Schematic of the structure and results from models 3 and 4 (Tab. S6 and Tab. S7) showing the outcome, official quarterly grades (center), all predictors (black-framed boxes), the statistical significance of their effect (arrows; black: $p<0.05$, grey: $p\geq0.05$), the unstandardized regression coefficients (b-values) and ID as random intercept (dashed box). General predictors (white) are categorical variables, so the levels with the highest impact are shown compared to their reference (female, grade level 10, quarter 1, languages). b-values are approximate in **a,** indicated by ≈, as representing results from models 3a-d.

a, Effect of changes in sleep and of 9AM-use on grade improvements from the conventional to the flexible system. Summarized results from models 3a-d (n_{ID}=63; Tab. S6) where each submodel included a different yellow predictor in interaction with school start system (conventional/flexible; $b_{flex*change}$) to model effects of sleep changes on grade changes. **b-e,** Visualization of the yellow interaction effects from (a) via marginal means, i.e. grade estimates and 95% CI for the reference situation (female student, class level 10, quarter 1, languages) and categorical splits in the continuous sleep change variables to facilitate display. The effect of school start system on grades by **b,** chronotype change (advance/delay), **c,** sleep duration change (sleep loss/sleep gain), **d,** social jetlag change (reduction/increase) from the conventional to the flexible system, and by **e,** the frequency of 9AM-use (<2days/≥2 days) in the flexible system.

f, Effect of absolute sleep characteristics on grades in the flexible system. Summarized results from model 4e (n_{ID}=129; Tab. S7) predicting grades only for the flexible system, i.e., 1.5 years post-change, including the red sleep predictors in one common model after running separate models (4a-d) to check for collinearity. **g-j,** Visualization of the red effects from (f) via marginal means, i.e. grade estimates and 95% CI for the reference situation (female student, class level 10, quarter 1, languages). *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$.

Discussion

Teenagers show restricted sleep on school days and oversleep on weekends. Early school starts are a major determinant of this pattern, thereby impacting on students' daily lives and their future trajectories. Most studies that looked at delayed school starts and sleep improvements were cross-sectional and thus could not track individual differences over time. We investigated whether a flexible school start system allows teenagers to reduce their sleep deprivation long-term, and whether changes in sleep translated to higher academic achievement in such a system.

The few studies that recorded sleep changes longitudinally after a delay in school start times reported mixed results. Bowers and Moyer (2016) determined in a meta-analysis[35] that all five longitudinal studies examined showed sleep extensions after a school start delay, and this benefit persisted until the follow-up period at 0.25 to 6 months after the delay[36,49,52,74,75]. Lo et al. (2018) also tracked sleep after a 45-min delay and found a delay in bedtime of 23 min which was sustained after 9 months[46]. In contrast, Thacher and Onyper (2016) showed a 20-min sleep extension after 45-min delay disappeared after 1 year because students delayed their sleep times[45]. Das-Friebel et al. (2020) also provided evidence that students merely shifted their sleep timing to later and thus did not benefit from their 20-min school delay after 1 year[47].

Here, in the flexible start system compared to the conventional start system, we found no shift in sleep timing but also no net sleep gains, which is probably connected to the low uptake of later starts of only 1-2 days per week on average and occasional later starts already during the conventional system. We identified three main reasons for this low uptake via survey answers during wave 1: students could not fulfil their quota of 10 self-study periods per week without otherwise getting home later in the afternoon (75%), it was easier to get to school for the 8AM-start (40%), and students wanted to have more time to study (27%)[59]. During wave 2, these reasons remained the most common ones (54%, 37%, 50% respectively), although yet another year later the uptake apparently rose to a median of 79% (IQR=70-86), i.e. 4 days per week, according to the school. It is therefore likely that the temperate use of the flexible starts during our recording period underlies the persistent absence of sleep benefits in the flexible system in our sample. Thus, many more late starts are probably required to translate into net sleep benefits in a flexible system. Alternatively – or in addition – the flexible system might have compensated a potential deterioration in sleep with increasing age or adolescence[76,77] and the absence of a net change in sleep between all time points is actually a success as it prevented a worsening. Longitudinal observational data, however, are unfortunately not suited to answer this question.

Within the flexible system, our results demonstrate that sleep length on ≥9AM-days in the flexible system remained increased on average by 1 hour even after one year, and that ≥9AM-starts were subjectively helpful for students across many psychological domains. The sleep and psychological effects might be either downstream of each other (e.g. longer and better sleep improving well-being and concentration or vice versa) or parallel improvements (e.g. more self-determination in the flexible system improving both sleep and psychological aspects in day-time functioning). The finding that almost every single student profited from a later start highlights the pervasiveness and severity of sleep deprivation in this age group.

Importantly, however, while girls' sleep benefit on ≥9AM-days was completely sustained over the follow-up period, boys' sleep gain was reduced after 1 year since they fell asleep later on ≥9AM days than on 8AM-days at t_2. This could have been a cohort effect but the larger cohort 2, which had a similar gender ratio, showed the same pattern. The delay in sleep onsets for boys but not for girls is a central finding, since avoiding delays in sleep onsets is key to long-term success of later school start times, both flexible and fixed. Our analyses revealed no effects of chronotype or frequency of later starts on this delay. We can thus only speculate about the possible biological, psychological and behavioral reasons explaining the observed gender difference, ranging from different circadian light sensitivities to (un)consciously differing sleep hygiene or pre-bed activities. For example, girls may have remained careful with their sleep hygiene since they continuously appreciated the extra sleep, while boys may have started to exploit their "longer" evenings for activities increasingly important to them. Boys could have also engaged in more screen-activities such as video games on school nights[78] before later school starts and thus received a stronger blue-light stimulus delaying sleep onsets on those nights. Alternatively, over time, boys may have developed a different strategy than girls about why and when they chose later starts. They may have chosen later starts predominantly on mornings after late nights *because* they fell asleep late. From our speculation it is clear that this gender difference after 1 year raises many central questions and might underlie the contradictory findings from the few previous longitudinal studies (with e.g. all-girls samples[46] or few gender analyses), highlighting the urgent need for long-term follow-ups of sleep timing adjustments.

The benefits of later school starts are also reflected by the fact that 45% to 59% of students across all cohorts enjoyed at least 8 hours of sleep on ≥9AM-days (Tab. 1), while numbers looked worrying on 8AM-days, when only 3% to 15% of students reached the minimal amount of 8h required for healthy sleep in teenagers[71]. Although students still did not get the recommended 8-10h on school days overall, this demonstrated that later starts are beneficial for teenage sleep and constitute a move in the right direction. Sleep lengths on ≥9AM-days got closer to more optimal levels which were otherwise only observed on weekends when 70-85% of students reached at least 8h of sleep.

Another bonus is that students themselves liked the new system. They were more motivated to go to school, they rated their concentration and motivation higher during class, and generally felt better on ≥9AM-days. These are all prerequisites for good academic achievement. However, we report here that these sleep and psychological benefits were only associated with higher grades at first sight. When not adjusting for confounding factors, we observed a small improvement of grades in the flexible system, which would be in line with some previous studies[e.g., 34,51]. We argue, however, that such simple pre-post analysis of aggregated grades is not suited to answer this complex question – although this has been frequently done despite the use of cross-sectional data. Few other studies on grades performed proficient analyses, such as mixed regression models[45], quantile regression models[53] or difference-in-difference approaches[54,55,79] accounting for available confounders. Often, self-reported grades, grades from a single academic subject[51] or coarse categories such as "mostly As/mostly Bs" were used as outcomes precluding sensitive or detailed analyses[34]. Results from these investigations included many null-findings[34] and the positive effects were rather small with changes ranging mainly from 1-4 percentage points on a 0%-100% scale. Nonetheless, positive effects of delayed school start times on academic performance have been widely proclaimed, bound to raise falsely high expectations in parents and teachers.

When we considered grade level, discipline and quarter in our mixed model analyses of our rich longitudinal dataset, we find that the flexible system is clearly not associated with overall grade improvements except for subtle increases in Languages and subtle decreases in Social Sciences. In fact, the "confounders" weighed much stronger in our sample than any system effects on individual disciplines: graduating students did constantly better, highest grades were given in the final quarter of the year, and students were most successful in Social Sciences. Furthermore, the interplay between gender, discipline and school start system on grades is complex.

Importantly, we also did not find any expected relationships between chronotype, social jetlag, or sleep duration with grades in our sample. Neither changes in these sleep parameters from the conventional to the flexible system nor their absolute values in the flexible system showed any link with grades - except for changes in social jetlag. Surprisingly, an increase, not a decrease, in social jetlag in the flexible system was predictive of higher grades in the flexible system. We have not been able to identify obvious explanations for this finding in exploratory analyses, except for the fact that weekend sleep was much more variant and backed by fewer data points than schoolday sleep, pointing towards a potential chance finding. A likely explanation for our null-finding for the other sleep parameters is a possible lack of power in our sample of 157 students (even though we have >16,000 longitudinal grades) given the small effect sizes previously identified. Thus, we cannot preclude a subtle effect in our sample but any such effect is likely extremely small and thus rather meaningless for real-life.

There is a substantial body of evidence supporting that both acute and chronic sleep loss compromises alertness, cognitive performance and memory, and reduces engagement to perform well (performance effort)[27,80,81]. Thus, improving sleep in sleep-deprived teenagers is very likely to improve their learning[82–84]. The question is whether better learning mediated by improved sleep actually translates into better grades. Students' grades are known to be strongly affected by many factors beyond those captured in our study or others' investigations on this topic. Models of teaching and learning include factors of individual students, such as motivation and prior knowledge; factors of the learning environment, such as learning atmosphere or class mates; and factors of instruction, such as teachers and instructional quality[85,86]. Additionally, research has shown that especially factors of instruction greatly influence students' learning[87,88]. Furthermore, grades are not always valid measures of students' academic performance, as teacher include other factors such as compliance, effort, attitude, or behavior in their assessment[89]. Therefore, it may be a big ask and possibly naive to expect grades to improve noticeably and within a few months after delays in school start times have been affected. Rather, we should acknowledge maintained achievements (under potentially less effort) besides the gift of more sleep and better well-being.

Our study has several limitations that have not yet been mentioned. Sleep analyses were solely based on subjective diaries entries. However, importantly, diary data corresponded very well to objective activity data in cohort 1 [59], and other studies report similar correlations[90,91], so we assume faithful reporting from our sample. Furthermore, our sleep calculations did not consider potential naps and hence might underestimate the total sleep duration in some students. We could not obtain information about teaching quality and classroom atmosphere but accounted for gender, quarter, grade level, and discipline - factors that are often overlooked in the field. Finally, we also did not have data on the socioeconomic background of our participants but students attending high school (the most academic type of school in Germany) tend to be from families with higher socio-economic status and often at least one parent has a similar educational level (65.9% of parents have A-levels, and 22.2% a General Certificate of Secondary Education equivalent[92]).

In conclusion, one should bear in mind that what matters most is the mental and physical health of our students. Even the most motivated students cannot learn when they are busy trying to keep their eyes open. Most importantly, teaching students to take responsibility, which incorporates to decide for themselves when to learn and to some extent when to start school, increases their motivation, investment, and wellbeing, and can thus have potential indirect effects on their sleep quality. These factors together form a profound basis for good academic achievement – a development that might take much longer than the time frames generally investigated, and might not necessarily translate into meaningful grade improvements. Nonetheless, maintained grades in addition to better sleep and well-being is already a central achievement.

Acknowledgements

We thank the students for their participation and the management of the Gymnasium Alsdorf, in particular W Bock and O Vollert, for their open support throughout the study. We also thank Dr. M Vuori-Brodowski for her contributions during wave 1, general advice and statistical support.

Data availability

Data were collected with a consent form that prohibits online deposition of data for open access sharing. This prohibition was implemented in order to protect participants' privacy in a cohort where most individuals are well-acquainted with each other and peers or teachers might identify participants. Data are available from the corresponding author upon reasonable request.

Code availability

We did not develop any custom code or algorithms for data analyses. The code for grade analyses can be found here on github:
https://github.com/annambiller/Schoolstudy/blob/13bfb38c35332fe285b3cab738b621a4083df7f6/Linear%20mixed%20model%20code%20for%20grade%20analyses.txt

Competing interest statement

None of the authors have any private links with the Gymnasium Alsdorf, and the school was not involved in data analysis or interpretation nor the writing of the manuscript. AMB received a travel grant from the ESRS to present parts of this study and research and travel funds from the Graduate School of Systemic Neurosciences Munich. CM, GZ and FO report no funding in relation to the study and outside the submitted work. CF reports no funding in relation to the study and receiving funds from the German Research Foundation (DFG), Joachim Herz Stiftung and a travel grant from the LMU Excellence Grant outside the submitted work. TR reports no funding in relation to the study and receiving funding from the DAAD outside the submitted work. ECW reports receiving travel funds from the German Dalton Society to present results of this study as well as from the LMU Excellence Grant, Gordon Research Conference and Friedrich-Baur-Stiftung outside the submitted work.

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. doi:10.1371/journal.pone.0112199.

2. Carskadon MA, Acebo C, Richardson GS, Tate BA, Seifer R. An Approach to Studying Circadian Rhythms of Adolescent Humans. *J Biol Rhythms*. 1997;12:278–89.

3. Carskadon MA, Acebo C, Jenni OG. Regulation of adolescent sleep: Implications for behavior. *Ann N Y Acad Sci*. 2004;1021:276–91. doi:10.1196/annals.1308.032.

4. Jenni OG, Achermann P, Carskadon MA. Homeostatic sleep regulation in adolescents. *Sleep*. 2005;28:1446–54. doi:10.1093/sleep/28.11.1446.

5. Taylor DJ, Jenni OG, Acebo C, Carskadon MA. Sleep tendency during extended wakefulness: Insights into adolescent sleep regulation and behavior. *J Sleep Res*. 2005;14:239–44. doi:10.1111/j.1365-2869.2005.00467.x.

6. Van Den Bulck J. Television viewing, computer game playing, and internet use and self-reported time to bed and time out of bed in secondary-school children. *Sleep*. 2004;27:101–4. doi:10.1093/sleep/27.1.101.

7. Munezawa T, Kaneita Y, Osaki Y, Kanda H, Minowa M, Suzuki K, et al. The Association between Use of Mobile Phones after Lights Out and Sleep Disturbances among Japanese Adolescents: A Nationwide Cross-Sectional Survey. *Sleep*. 2011;34:1013–20. doi:10.5665/SLEEP.1152.

8. Cajochen C. Alerting effects of light. *Sleep Med Rev*. 2007. doi:10.1016/j.smrv.2007.07.009.

9. Souman JL, Tinga AM, te Pas SF, van Ee R, Vlaskamp BNS. Acute alerting effects of light: A systematic literature review. *Behav Brain Res*. 2018. doi:10.1016/j.bbr.2017.09.016.

10. Yang M, Ma N, Zhu Y, Su YC, Chen Q, Hsiao FC, et al. The acute effects of intermittent light exposure in the evening on alertness and subsequent sleep architecture. *Int J Environ Res Public Health*. 2018. doi:10.3390/ijerph15030524.

11. Wittmann M, Dinich J, Merrow M, Roenneberg T. Social Jetlag: Misalignment of Biological and Social Time. *Chronobiol Int*. 2006;23:497–509. doi:10.1080/07420520500545979.

12. Kuula L, Pesonen AK, Merikanto I, Gradisar M, Lahti J, Heinonen K, et al. Development of Late Circadian Preference: Sleep Timing From Childhood to Late Adolescence. *J Pediatr*. 2018;194:182-189.e1. doi:10.1016/j.jpeds.2017.10.068.

13. Short MA, Gradisar M, Lack LC, Wright HR, Dewald JF, Wolfson AR, et al. A Cross-Cultural Comparison of Sleep Duration Between U.S. and Australian Adolescents: The Effect of School Start Time, Parent-Set Bedtimes, and Extracurricular Load. *Heal Educ Behav*. 2013;40:323–30. doi:10.1177/1090198112451266.

14. Hirshkowitz M, Whiton K, Albert SM, Alessi C, Bruni O, DonCarlos L, et al. National sleep foundation's sleep time duration recommendations: Methodology and results summary. *Sleep Heal*. 2015;1:40–3. doi:10.1016/j.sleh.2014.12.010.

15. 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. doi:10.1038/ijo.2011.149.

16. Mullington JM, Haack M, Toth M, Serrador JM, Meier-Ewert HK. Cardiovascular, Inflammatory, and Metabolic Consequences of Sleep Deprivation. *Prog Cardiovasc Dis*. 2009. doi:10.1016/j.pcad.2008.10.003.

17. 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. doi:10.1080/15402002.2015.1120198.

18. Short MA, Gradisar M, Lack LC, Wright HR. The impact of sleep on adolescent depressed mood,

alertness and academic performance. *J Adolesc*. 2013. doi:10.1016/j.adolescence.2013.08.007.

19. 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. doi:10.1111/jcpp.12125.

20. Tynjälä J, Kannas L, Levälahti E. Perceived tiredness among adolescents and its association with sleep habits and use of psychoactive substances. *J Sleep Res*. 1997;6:189–98. doi:10.1046/j.1365-2869.1997.00048.x.

21. Pasch KE, Latimer LA, Cance JD, Moe SG, Lytle LA. Longitudinal Bi-directional Relationships Between Sleep and Youth Substance Use. *J Youth Adolesc*. 2012. doi:10.1007/s10964-012-9784-5.

22. Larcher S, Gauchez AS, Lablanche S, Pépin JL, Benhamou PY, Borel AL. Impact of sleep behavior on glycemic control in type 1 diabetes: The role of social jetlag. *Eur J Endocrinol*. 2016;175:411–9. doi:10.1530/EJE-16-0188.

23. Parsons MJ, Moffitt TE, Gregory AM, Goldman-Mellor S, Nolan PM, Poulton R, et al. Social jetlag, obesity and metabolic disorder: Investigation in a cohort study. *Int J Obes*. 2015;39:842–8. doi:10.1038/ijo.2014.201.

24. Roenneberg T, Allebrandt K V., Merrow M, Vetter C. Social jetlag and obesity. *Curr Biol*. 2012;22:939–43. doi:10.1016/j.cub.2012.03.038.

25. 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 Public Health*. 2015. doi:10.1177/1403494814556647.

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

27. 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;9:517–26. doi:10.1016/j.sleep.2007.07.003.

28. Horne JA. Sleep loss and “divergent” thinking ability. *Sleep*. 1988;11:528–36. doi:10.1093/sleep/11.6.528.

29. Randazzo AC, Muehlbach MJ, Schweitzer PK, Walsh JK. Cognitive Function Following Acute Sleep Restriction in Children Ages 10–14. *Sleep*. 1998;21. doi:10.1093/sleep/21.8.861.

30. Harrison Y, Horne JA. Sleep deprivation affects speech. *Sleep*. 1997;20:871–7. doi:10.1093/sleep/20.10.871.

31. Harrison Y, Horne JA. Sleep loss impairs short and novel language tasks having a prefrontal focus. *J Sleep Res*. 1998;7:95–100. doi:10.1046/j.1365-2869.1998.00104.x.

32. 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. doi:10.1016/j.smrv.2009.10.004.

33. Owens J. Insufficient Sleep in Adolescents and Young Adults: An Update on Causes and Consequences. *Pediatrics*. 2014;134:e921–32. doi:10.1542/peds.2014-1696.

34. 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 Heal*. 2017;86:363–81. doi:10.1111/josh.12388.School.

35. Bowers JM, Moyer A. Effects of school start time on students’ sleep duration, daytime sleepiness, and attendance: a meta-analysis. *Sleep Heal*. 2017;3:423–31. doi:10.1016/j.sleh.2017.08.004.

36. 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. doi:10.1097/DBP.0000000000000018.

37. 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. doi:10.1016/j.smrv.2015.06.002.

38. 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. *Cochrane Database Syst*

Rev. 2017;2017. doi:10.1002/14651858.CD009467.pub2.

39. Troxel WM, Wolfson AR. The intersection between sleep science and policy: introduction to the special issue on school start times. *Sleep Heal*. 2017;3:419–22. doi:10.1016/j.sleh.2017.10.001.

40. Levin KA. Study design III: Cross-sectional studies. *Evid Based Dent*. 2006;7:24–5. doi:10.1038/sj.ebd.6400375.

41. Illingworth G, Sharman R, Jowett A, Harvey CJ, 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. doi:10.1016/j.sleep.2018.10.021.

42. Estevan I, Silva A, Vetter C, Tassino B. Short Sleep Duration and Extremely Delayed Chronotypes in Uruguayan Youth: The Role of School Start Times and Social Constraints. *J Biol Rhythms*. 2020:1–14. doi:10.1177/0748730420927601.

43. Goldin AP, Sigman M, Braier G, Golombek DA, Leone MJ. Interplay of chronotype and school timing predicts school performance. *Nat Hum Behav*. 2020:43–7. doi:10.1038/s41562-020-0820-2.

44. Wahlstrom K. Changing Times: Findings From the First Longitudinal Study of Later High School Start Times. *NASSP Bull*. 2002;86:3–21. doi:10.1177/019263650208663302.

45. Thacher P V., Onyper S V. Longitudinal Outcomes of Start Time Delay on Sleep, Behavior, and Achievement in High School. *Sleep*. 2016;39:271–81. doi:10.5665/sleep.5426.

46. Lo JC, Lee SM, Lee XK, Sasmita K, Chee NIYN, Tandi J, et al. Sustained benefits of delaying school start time on adolescent sleep and well-being. *Sleep*. 2018. doi:10.1093/sleep/zsy052.

47. Das-Friebel A, Gkiouleka A, Grob A, Lemola S. Effects of a 20 minutes delay in school start time on bed and wake up times, daytime tiredness, behavioral persistence, and positive attitude towards life in adolescents. *Sleep Med*. 2020;66:103–9. doi:10.1016/j.sleep.2019.07.025.

48. Widome R, Berger AT, Iber C, Wahlstrom K, Laska MN, Kilian G, et al. Association of Delaying School Start Time with Sleep Duration, Timing, and Quality among Adolescents. *JAMA Pediatr*. 2020;174:697–704. doi:10.1001/jamapediatrics.2020.0344.

49. Lufi D, Tzischinsky O, Hadar S. Delaying school starting time by one hour: Some effects on attention levels in adolescents. *J Clin Sleep Med*. 2011;7:137–43.

50. Nahmod NG, Lee S, Master L, Chang AM, Hale L, Buxton OM. Later high school start times associated with longer actigraphic sleep duration in adolescents. *Sleep*. 2019;42:1–10. doi:10.1093/sleep/zsy212.

51. 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. doi:10.1126/sciadv.aau6200.

52. Carskadon MA, Wolfson AR, Acebo C, Tzischinsky O, Seifer R. Adolescent sleep patterns, circadian timing, and sleepiness at a transition to early school days. *Sleep*. 1998;21:871–81.

53. Edwards F. Early to rise? The effect of daily start times on academic performance. *Econ Educ Rev*. 2012;31:970–83. doi:10.1016/j.econedurev.2012.07.006.

54. Kim T. The Effects of School Start Time on Educational Outcomes: Evidence From the 9 OOClock Attendance Policy in South Korea. *SSRN Electron J*. 2018;2019:1–26. doi:10.2139/ssrn.3160037.

55. Shin J. Sleep More , Study Less ? The Impact of Delayed School Start Time on Sleep and Academic Performance 1 Introduction 2018:0–52.

56. 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;7:1–10. doi:10.1038/s41598-017-04076-y.

57. Wahlstrom KL, Owens JA. School start time effects on adolescent learning and academic performance, emotional health and behaviour. *Curr Opin Psychiatry*. 2017;30:485–90. doi:10.1097/YCO.0000000000000368.

58. Dunster GP, Crowley SJ, Carskadon MA, de la Iglesia HO. What Time Should Middle and High School

Students Start School? *J Biol Rhythms*. 2019;34:576–8. doi:10.1177/0748730419892118.

59. Winnebeck EC, Vuori-Brodowski MT, Biller AM, Molenda C, Fischer D, Zerbini G, et al. Later school start times in a flexible system improve teenage sleep. *Sleep*. 2019:1–17. doi:10.1093/sleep/zsz307.

60. Der Deutsche Schulpreis 2019. https://www.deutscher-schulpreis.de/preistraeger/gymnasium-der-stadt-alsdorf.

61. Ghotbi N, Pilz LK, Winnebeck EC, Vetter C, Zerbini G, Lenssen D, et al. The µMCTQ: An Ultra-Short Version of the Munich ChronoType Questionnaire. *J Biol Rhythms*. 2020;35:98–110. doi:10.1177/0748730419886986.

62. Wickham H, Chang W, Henry L, Pederson TL, Takahashi K, Wilke C, et al. R package ggplot2: Elegant Graphics for Data Analysis. Version 3.2.1. 2019.

63. Pinheiro J, Bates D, DebRoy S, Sarkar D. R package nlme: Linear and Nonlinear Mixed Effects Models 2020.

64. Bates D, Mächler M, Bolker B, Walker S. R package: Fitting Linear Mixed-Effects Models Using lme4. Version 1.1-8. 2015.

65. Kuznetsova A, Brockhoff PB, Christensen RHB. R package lmerTest: Test in Linear Mixed Effects Models. Version 3.0-1 2018.

66. Fox J, Weisberg S, Price B, Adler D, Bates D, Baud-Bovy G, et al. R package car: Companion to Applied Regression. Version 3.0-8. 2020.

67. Lenth R, Singmann H, Love J, Buerkner P, Herve M. R package emmeans: Estimated Marginal Means, aka Least-Square Means. Version 1.4.6. 2020.

68. Lüdecke D, Bartel A, Schwemmer C, Powell C, Djalovski A. R package sjPlot: Data Visualization for Statistics in Social Science. Version 2.8.4. 2020.

69. Lüdecke D, Giné-Vásquez I, Bartel A. R package sjmisc: Data and Variable Transformation Functions. Version 2.8.4. 2020.

70. Lüdecke D, Aust F. R package ggeffects: Create Tidy Data Frames of Marginal Effects for “ggplot” from Model Outputs. Version 0.15.0. 2020.

71. Paruthi S, Brooks LJ, D’Ambrosio C, Hall WA, Kotagal S, Lloyd RM, et al. Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine. *J Clin Sleep Med*. 2016;12:785–6. doi:10.5664/jcsm.5866.

72. Thacher P V., Onyper S V. Longitudinal Outcomes of Start Time Delay on Sleep, Behavior, and Achievement in High School. *Sleep*. 2016;39:271–81. doi:10.5665/sleep.5426.

73. Voyer D, Voyer SD. Gender differences in scholastic achievement: A meta-analysis. *Psychol Bull*. 2014;140:1174–204. doi:10.1037/a0036620.

74. Owens JA, Belon K, Moss P. Impact of delaying school start time on adolescent sleep, mood, and behavior. *Arch Pediatr Adolesc Med*. 2010;164:608–14. doi:10.1001/archpediatrics.2010.96.

75. Wolfson A, Tzischinsky O, Brown C, Darley C, Acebo C, Carskadon M. Sleep, behavior, and stress at the transition to senior high school. *Sleep Res*. 1995;24:115.

76. Bai S, Karan M, Gonzales NA, Fuligni AJ. A daily diary study of sleep chronotype among Mexican-origin adolescents and parents: Implications for adolescent behavioral health. *Dev Psychopathol*. 2020:1–10. doi:10.1017/S0954579419001780.

77. 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. doi:10.1016/j.adolescence.2018.06.001.

78. Desai RA, Krishnan-Sarin S, Cavallo D, Potenza MN. Video-gaming among high school students: Health correlates, gender differences, and problematic gaming. *Pediatrics*. 2010;126. doi:10.1542/peds.2009-2706.

79. Jung H. A late bird or a good bird? The effect of 9 o’clock attendance policy on student’s achievement.

Asia Pacific Educ Rev. 2018;19:511–29. doi:10.1007/s12564-018-9558-1.

80. Lim J, Dinges DF. A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychol Bull*. 2010;136:375. doi:10.1037/a0018883.

81. Engle-Friedman M. The effects of sleep loss on capacity and effort. *Sleep Sci*. 2014;7:213–24. doi:10.1016/j.slsci.2014.11.001.

82. Tarokh L, Saletin JM, Carskadon MA. Sleep in adolescence: physiology, cognition and mental health. *Neurosci Biobehav Rev*. 2016;70:182–8. doi:10.1016/j.neubiorev.2016.08.008.

83. Sadeh A, Gruber R, Raviv A. The Effects of Sleep Restriction and Extension on School-Age Children: What a Difference an Hour Makes. *Child Dev*. 2003;74:444–55. doi:10.1111/1467-8624.7402008.

84. Potkin KT, Bunney WE. Sleep improves memory: The effect of sleep on long term memory in early adolescence. *PLoS One*. 2012;7:8–11. doi:10.1371/journal.pone.0042191.

85. Löwen K, Baumert J, Kunter M, Krauss S, Brunner M. Cognitive activation in the mathematics classroom and professional competence of teachers. Results from the COACTIV project. In: Kunter M, Baumert J, Blum W, Klusmann U, Krauss S, Neubrand M, editors. COACTIV Res. Progr. Methodol. Fram., New York: Springer; 2013, p. 79–96.

86. Neumann K, Kauertz A, Fischer HE. Quality of instruction in science education. In: Fraser BJ, Tobin KG, McRobbie CJ, editors. Second Int. Handb. Sci. Educ., Berlin: Springer; 2012, p. 247–58. doi:10.1007/978-1-4020-9041-7.

87. Seidel T, Shavelson RJ. Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Rev Educ Res*. 2007;77:454–99. doi:10.3102/0034654307310317.

88. Hattie J. Visible learning: A synthesis of over 800 meta-analyses relating to achievement. 2008. doi:10.4324/9780203887332.

89. Allen JD. Grades as Valid Measures of Academic Achievement of Classroom Learning. *Clear House A J Educ Strateg Issues Ideas*. 2005;78:218–23. doi:10.3200/tchs.78.5.218-223.

90. Campanini MZ, Lopez-Garcia E, Rodríguez-Artalejo F, González AD, Andrade SM, Mesas AE. Agreement between sleep diary and actigraphy in a highly educated Brazilian population. *Sleep Med*. 2017;35:27–34. doi:10.1016/j.sleep.2017.04.004.

91. Tremaine RB, Dorrian J, Blunden S. Subjective and objective sleep in children and adolescents: Measurement, age, and gender differences. *Sleep Biol Rhythms*. 2010;8:229–38. doi:10.1111/j.1479-8425.2010.00452.x.

92. Statistisches Bundesamt. Verteilung der Schüler auf Gymnasien nach dem höchsten Bildungsabschluss der Eltern im Jahr 2018. *Statista*. 2019. https://de.statista.com/statistik/daten/studie/162247/umfrage/besuch-des-gymnasiums-nach-abschluss-der-eltern/ (accessed May 24, 2020).

APPENDIX - Project 2

Supplementary information for

One year later: longitudinal effects of flexible school start times on teenage sleep, psychological benefits, and academic grades

Authors:

Anna M. Biller, Carmen Molenda, Giulia Zerbini, Fabian Obster, Christian Förtsch, Till Roenneberg & Eva C. Winnebeck

Contact information:

eva.winnebeck@med.uni-muenchen.de

Supplementary Figures

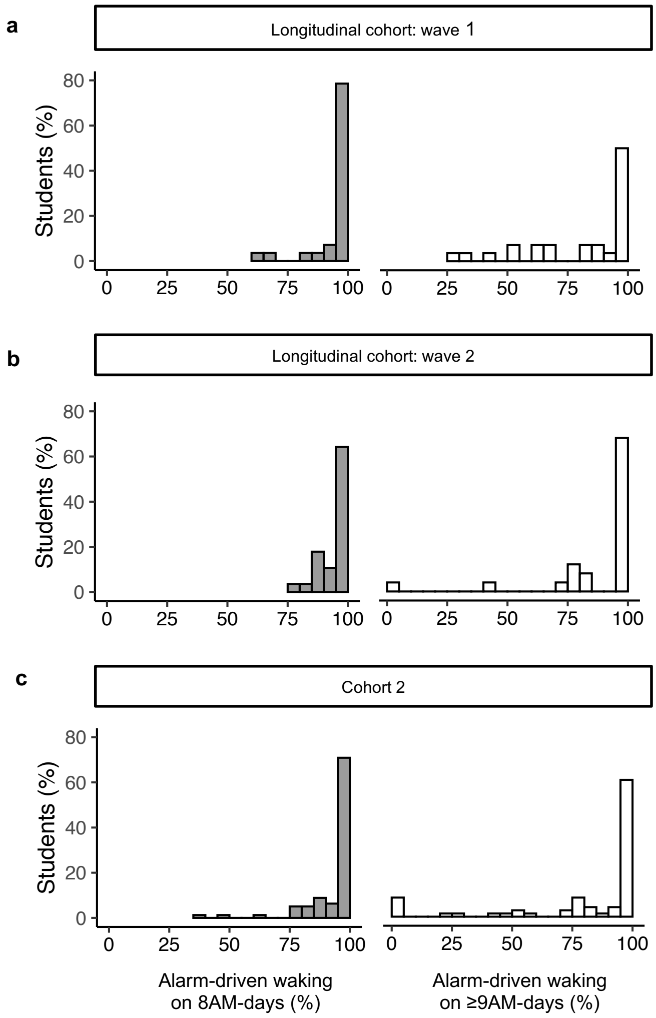

Fig. S1 | Alarm-driven waking on 8AM-days and ≥9AM-days in the flexible system. Histograms displaying the distribution of alarm-driven waking on schooldays (% of recorded schooldays) for 8AM-days (left panels) and ≥9AM-days (right panels) **a,** Longitudinal cohort (n=28) during wave 1, **b,** during wave 2, and **c,** cohort 2 (n=79).

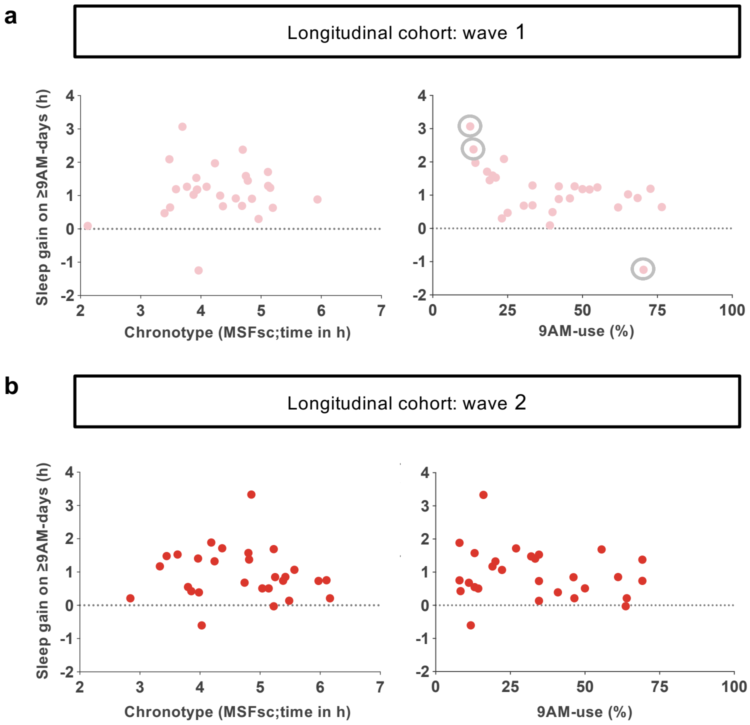

Fig. S2 | Inter-individual differences in sleep gain on ≥9AM-days. Shown are relationships between chronotype (MSF_{sc}; local time) or frequency of ≥9AM-starts (% of schooldays with later starts) with sleep gain on ≥9AM-days. Sleep gain was quantified as the absolute difference in sleep duration between ≥9AM and 8AM-days, with positive numbers indicating longer sleep duration on ≥9AM-days. Data are from the longitudinal cohort (n=28) during **a,** wave 1 (light red) and **b,** wave 2 (red). Results of Pearson correlations are indicated in figures; grey circles indicate Tukey outliers that were removed for statistical analysis.

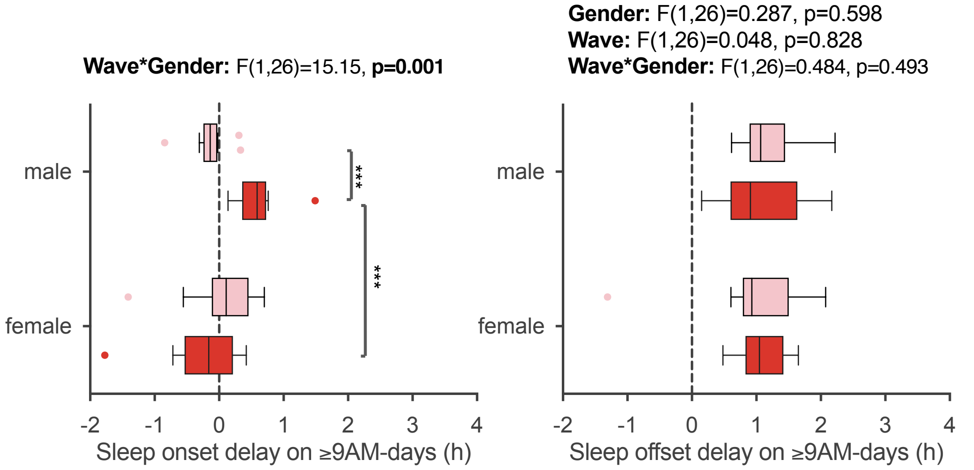

Fig. S3 | Gender differences in sleep onset and offset on ≥9AM-days versus 8AM-days in the flexible system. Depicted is the average absolute difference in **a,** sleep onset (sleep onset delay) and **b,** sleep offset (sleep offset delay) between 8AM and ≥9AM-days for the longitudinal cohort (n=28), with higher numbers indicating later times on ≥9AM-days. Results of two-way mixed ANOVAs with the between-subjects factor gender (female/male) and the within-subjects factor wave (light red=wave 1/red= wave 2) are reported above each graph. Given the significant interaction effect on sleep onset delay, main effects are not reported, instead statistically significant post-hoc comparisons are indicated. All boxplots are Tukey boxplots. *, p<0.05; **, p<0.01; ***, p<0.00

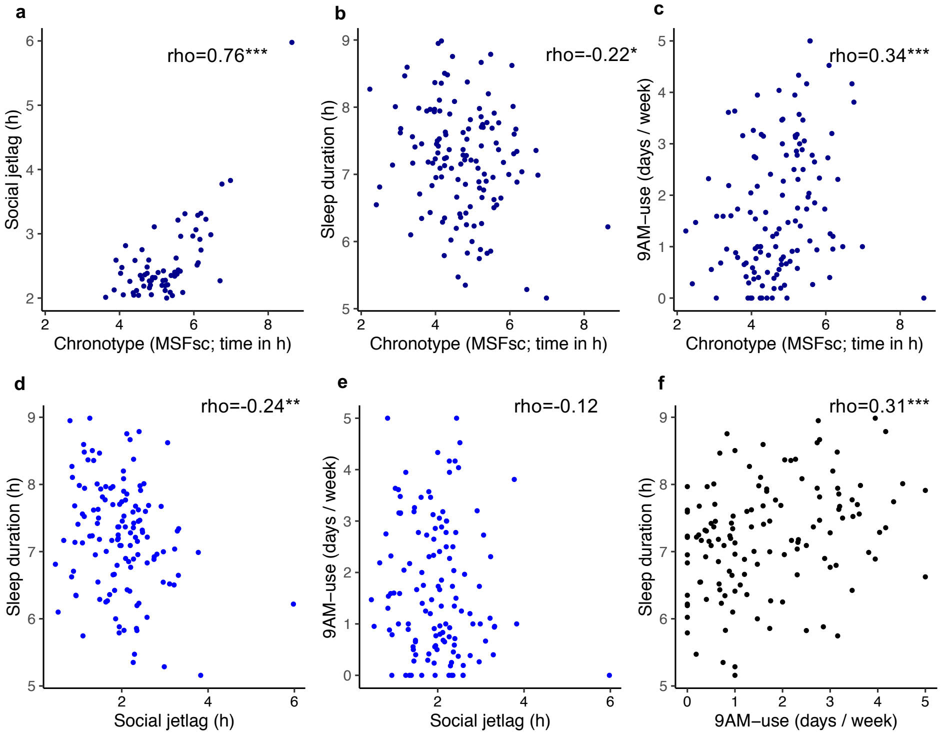

Fig. S4 | Correlations between sleep variables. Spearman rank correlations between the sleep variables social jetlag, chronotype (MSF_{sc}), and sleep duration on schooldays, as well as frequency of ≥9AM-starts (n=129). *, p<0.05; **, p<0.01; ***, p<0.001

Supplementary Tables

Tab. S1 | Individual differences in sleep gain on ≥9AM-days. Linear regression analyses on sleep gain, sleep onset delay and sleep offset delay on ≥9AM-days compared to 8AM-days in cohort 2 (N=79). Abbreviations: b, unstandardized coefficient; std. error, standard error; t, t-statistic; p, p-value. R^2 describes the explanatory power of the model (how much variance is explained). R^2 adjusted is the explanatory power accounted for the number of predictors in the model.

	Sleep onset delay				Sleep offset delay				Sleep gain			
Predictors	b	std. error	t	p	b	std. error	t	p	b	std. error	t	p
(Intercept)	0.25	0.34	0.72	0.471	1.27	0.31	4.10	<0.001	1.03	0.49	2.09	**0.040**
Gender: Male[a]	**0.53**	0.14	3.90	**<0.001**	0.01	0.12	0.07	0.942	-**0.52**	0.20	-2.64	**0.010**
Grade level: 11[b]	-0.03	0.17	-0.16	0.872	-0.21	0.16	-1.33	0.186	-0.19	0.25	-0.73	0.468
Grade level: 12[b]	0.01	0.18	0.08	0.933	-0.10	0.16	-0.62	0.538	-0.11	0.26	-0.45	0.655
Chronotype (MSF_{sc}; time in h)	-0.08	0.07	-1.06	0.295	0.03	0.07	0.51	0.611	0.11	0.11	1.05	0.298
9AM-use (schooldays/ week)	0.04	0.06	0.55	0.583	-0.11	0.06	-1.77	0.081	0.14	0.09	-1.50	0.139
Observations	79				79				79			
R^2 / R^2 adjusted	0.175 / 0.119				0.072 / 0.008				0.113 / 0.052			

[a] Reference is female.
[b] Reference is grade level 10.

Tab. S2 | Sleep differences between school start systems and type of day in the longitudinal cohort (post-hoc comparisons relating to Fig. 3a). Two-way repeated measures ANOVAs were run for sleep onset, sleep offset, and sleep duration with the within-factors day (schooldays/weekends) and time point (t_1/t_2/t_3) (see Fig. 3a). Because of significant interaction of both factors, simple main effects were carried out as post-hoc tests for time point comparisons and paired t-tests for day comparisons. Data presented are mean ± standard deviation from the longitudinal cohort (n=33) for these factors and the post-hoc results.

Sleep onset

School System	t_0	t_1	t_2	Simple main effect	Post hoc paired t-tests
Schooldays	-0.54h ± 0.79	-0.43h ± 0.75	-0.39h ± 0.73	F(2,31)= 1.61 p=0.217	-
Weekends	0.74h ± 0.94	0.46h ± 0.91	0.81h ± 1.00	F(2,31)= 2.681 p=0.084	-
Simple main effect	F(1,32)=153.70 **p<0.001**	F(1,32)=62.57 **p<0.001**	F(1,32)=72.88 **p<0.001**	Main interaction Day*System: F(2,64)= 0.4.17 **p<0.020**	

Sleep offset

School System	t_0	t_1	t_2	Simple main effect	Post hoc paired t-tests
Schooldays	6.62h ± 0.45	6.80h ± 0.47	6.76h ± 0.44	**F(2,31)= 9.029** **p=0.001**	t_0-t_1: **p<0.001** t_0-t_2: **p=0.025**
Weekends	9.69h ± 0.97	9.31h ± 0.87	9.42h ± 0.93	**F(2,31)= 4.882** **p=0.014**	t_0-t_1: **p=0.004**
Simple main effect	F(1,32)=294.21 **p<0.001**	F(1,32)=240.44 **p<0.001**	F(1,32)=322.85 **p<0.001**	Main interaction Day*System: F(2,64)= 10.418 **p< 0.001**	

Sleep duration

School System	t_0	t_1	t_2	Simple main effect	Post hoc paired t-tests
Schooldays	7.10h ± 1.00	7.14h ± 0.56	7.16h ± 0.47	F(2,31)=0.539 p=0.588	-
Weekends	8.57h ± 0.42	8.50h ± 0.43	8.36h ± 0.55	F(2,31)=2.700 p=0.083	-
Simple main effect	F(1,32)=120.238 **p<0.001**	F(1,32)=96.519 **p<0.001**	F(1,32)=59.066 **p<0.001**	Main Interaction Day*System: **F(2,64)= 3.880** **p=0.026**	

Tab. S3 | Sleep differences between school start systems and type of days in cohort 2. Sleep data from cohort 2 (n=105) are presented as mean ± standard deviation, and were analysed via paired t-test.

	Sleep onset	Sleep offset	Sleep duration
Schooldays	-0.38h ± 0.81	6.83h ± 0.55	7.21 ± 0.76
Weekends	0.85h ± 1.11	9.57 ± 1.13	8.72h ± 0.95
Paired t-test	t(104)=-14.757, **p<0.001**	t(104)=-26.471, **p<0.001**	t(104)=-14.230, **p<0.001**

Tab. S4 | Linear mixed regression models 1 and 2: General and system effects on grades. Predicted outcomes are quarterly grades (0%-100%) in 12 academic subjects from students of cohort 1 and 2 (n=157). Abbreviations: b, unstandardized coefficient; se, standard error; t, t-statistic; p, p-value; σ^2, variance of residuals of random effects; τ_{00}, variance of ID intercepts of random effects; ICC, intra-class correlation coefficient (describes how much variance is explained by the random effects); N, number of participants; Marginal R^2 describes the amount of variance explained by the fixed effects (predictors); Conditional R^2 describes the amount of variance explained by the full model.

	Model 1				Model 2			
Predictors	b	std. error	t	p	b	std. error	t	p
(Intercept)	54.11	1.21	44.74	**<0.001**	52.79	1.21	43.56	**<0.001**
System: Flexible system[a]	-0.10	0.42	-0.23	0.815	0.64	0.55	1.16	0.244
Gender: Male[b]	**-4.72**	2.13	-2.21	**0.028**	-1.43	2.12	-0.67	0.501
Grade level: 7[c]	1.17	0.74	1.60	0.111	1.23	0.74	1.66	0.097
Grade level: 8[c]	**3.11**	0.49	6.39	**<0.001**	**3.15**	0.49	6.44	**<0.001**
Grade level: 9[c]	**2.59**	0.34	7.53	**<0.001**	**2.62**	0.35	7.58	**<0.001**
Grade level: 11[c]	0.50	0.35	1.42	0.155	0.48	0.36	1.34	0.180
Grade level: 12[c]	**3.44**	0.55	6.21	**<0.001**	**3.37**	0.56	6.05	**<0.001**
Quarter: 2[d]	**0.82**	0.32	2.59	**0.010**	**0.82**	0.32	2.58	**0.010**
Quarter: 3[d]	0.30	0.34	0.88	0.378	0.25	0.34	0.75	0.451
Quarter: 4[d]	**2.34**	0.33	7.15	**<0.001**	**2.30**	0.33	7.00	**<0.001**
Discipline: Sciences[e]	**3.40**	0.30	11.36	**<0.001**	**5.22**	0.30	17.30	**<0.001**
Discipline: Social Sciences[e]	**3.87**	0.36	10.69	**<0.001**	**7.25**	0.37	19.33	**<0.001**
Male*Sciences	**4.74**	0.54	8.83	**<0.001**				
Male*Social Sciences	**8.03**	0.65	12.44	**<0.001**				
Flexible system*Sciences					**-1.23**	0.54	-2.29	**0.022**
Flexible system*Social Sciences					**-2.56**	0.63	-4.05	**<0.001**
Flexible system*Male					**1.32**	0.52	2.53	**0.011**
Random Effects								
σ^2	200.29				202.07			
τ_{00}	146.38 $_{\text{ID}}$				146.37 $_{\text{ID}}$			
ICC	0.42				0.42			
N	157 $_{\text{ID}}$				157 $_{\text{ID}}$			
Observations	16724				16724			
Marginal R^2 / Conditional R^2	0.033 / 0.441				0.028 / 0.436			

[a] Reference is conventional system.
[b] Reference is female.
[c] Reference is grade level 10.
[d] Reference is quarter 1.
[e] Reference is Languages.

Tab. S5 | Post hoc results of mixed models 1 and 2. Results are presented as marginal estimated means of quarterly grades scaled 0-100% (standard error), degrees of freedom. Simple contrast results are presented as estimated difference of academic grades (standard error), p-value. Degrees of freedom method: Kenward-Rogers. Results are averaged over the levels of system or gender, grade level, and quarter. Tukey method for comparison of 3 estimates.

Model 1				
Gender	Languages	Sciences	Social Sciences	Simple contrasts
female	**56.7** (1.19) 166	**60.1** (1.19) 166	**60.6** (1.21) 176	Languages-Sciences: **-3.41** (0.30), **p<.0001** Languages-Social Sciences: **-3.87** (0.36), **p<.0001** Sciences-Social Sciences: **-0.47** (0.36), p=0.3895
male	**52.0** (1.79) 164	**60.2** (1.78) 163	**63.9** (1.80) 172	Languages-Sciences: **-8.15** (0.45), **p<.0001** Languages-Social Sciences: **-11.90** (0.54), **p<.0001** Sciences-Social Sciences: **-3.76** (0.52), **p<.0001**
Simple contrasts	**4.72** (2.13) p=0.0284	-**0.00** (2.13) p=0.9915	-**3.31** (2.16) p=0.1269	
Model 2				
System	Languages	Sciences	Social Sciences	Simple contrasts
conventional	54.7 (1.08) 168	59.9 (1.08) 167	62.0 (1.10) 182	Languages-Sciences: **-5.22** (0.30), **p<0.0001** Languages-Social Sciences: **-7.25** (0.38), **p<0.0001** Sciences-Social Sciences: **-2.03** (0.37), **p<0.0001**
flexible	56.0 (1.14) 212	60.0 (1.13) 206	60.7 (1.16) 227	Languages-Sciences: **-3.99** (0.45), **p<.0001** Languages-Social Sciences: **-4.69** (0.51), **p<.0001** Sciences-Social Sciences: **-0.69** (0.49), p=0.3355
Simple contrasts	**-1.30** (0.54) **p=0.0168**	-0.07 (0.61) p=0.8849	**1.26** (0.61) **p=0.0384**	
System	Female	Male		Simple contrasts
conventional	**59.6** (1.18) 160	**58.2** (1.77) 158		female-male: 1.43 (2.12), p=0.5010
flexible	**59.0** (1.22) 186	**58.9** (1.81) 173		female-male: 0.11 (2.14), p=0.9580
Simple contrasts	0.62 (0.45) p=0.1726	0.7 (0.56) p=0.2106		

Tab. S6 | Linear mixed regression models 3a-d: Effect of changes in sleep and ≥9AM-use on grade improvements from the conventional to the flexible system. Predicted outcomes are quarterly grades (0%-100%) in 12 academic subjects from students of cohort 1 (n=63) over 4 years. "Change" refers to the absolute difference of the respective sleep variable between the conventional and the flexible system (t_1-t_0). Positive numbers mean later chronotype, longer sleep and more social jetlag in the flexible system (t_1). 9AM-use is the frequency of ≥9AM-starts at t_1 (no baseline data for calculation of change available). Abbreviations: Flex, Flexible system; b, unstandardized coefficient; se, standard error; p, p value; σ^2, variance of residuals of random effects; τ_{00}, variance of ID intercepts of random effects; ICC, intra-class correlation coefficient (describes how much variance is explained by the random effects); N, number of participants; Marginal R^2 describes the amount of variance explained by the fixed effects (predictors); Conditional R^2 describes the amount of variance explained by the full model.

	Model3a: Chronotype change			Model3b: Sleep duration change			Model3c: Social jetlag change			Model3d: 9AM-use		
Predictors	b	se	p	b	se	p	b	se	p	b	se	p
(Intercept)	53.26	1.97	**<0.001**	53.30	1.94	**<0.001**	53.29	1.95	**<0.001**	53.35	1.86	**<0.001**
System: Flexible system	-0.01	0.68	0.983	0.00	0.68	0.998	0.04	0.68	0.953	0.02	0.68	0.977
Gender: Male[a]	-2.56	3.24	0.432	-2.64	3.11	0.399	-2.64	3.15	0.405	-2.80	2.98	0.352
Grade level: 7[b]	5.96	3.33	0.074	6.31	3.35	0.060	6.46	3.34	0.053	5.89	3.33	0.077
Grade level: 8[b]	2.52	0.77	**0.001**	2.49	0.77	**0.001**	2.49	0.77	**0.001**	2.50	0.76	**0.001**
Grade level: 9[b]	2.14	0.57	**<0.001**	2.13	0.57	**<0.001**	2.11	0.57	**<0.001**	2.08	0.57	**<0.001**
Grade level: 11[b]	1.07	0.53	**0.045**	1.08	0.53	**0.042**	1.11	0.53	**0.039**	1.11	0.53	**0.038**
Grade level: 12[b]	3.42	0.82	**<0.001**	3.38	0.82	**<0.001**	3.37	0.82	**<0.001**	3.46	0.82	**<0.001**
Quarter: 2[c]	0.71	0.50	0.157	0.70	0.50	0.160	0.71	0.50	0.155	0.71	0.50	0.154
Quarter: 3[c]	0.47	0.54	0.386	0.46	0.54	0.392	0.47	0.54	0.380	0.48	0.54	0.367
Quarter: 4[c]	2.31	0.52	**<0.001**	2.30	0.52	**<0.001**	2.30	0.52	**<0.001**	2.31	0.52	**<0.001**
Discipline: Sciences[d]	6.22	0.40	**<0.001**	6.22	0.40	**<0.001**	6.22	0.40	**<0.001**	6.22	0.40	**<0.001**
Discipline: Social Sciences[d]	7.63	0.48	**<0.001**	7.63	0.48	**<0.001**	7.64	0.48	**<0.001**	7.64	0.48	**<0.001**
Chronotype change (MSF_{sc}; h)	0.40	2.18	0.855									
Flex* Chronotype change	0.10	0.53	0.845									
Sleep duration change (h)				2.32	2.94	0.434						
Flex* Sleep duration change				-0.77	0.83	0.352						
Social jetlag change (h)							0.11	2.11	0.958			
Flex* Social jetlag change							1.28	0.58	**0.027**			
9AM-use (schooldays/week)										-3.04	1.21	**0.015**
Flex*9AM-use										0.59	0.36	0.101
Random effects												
σ^2	205.07			205.05			204.92			204.99		
τ_{00}	140.27 $_{ID}$			139.10 $_{ID}$			140.24 $_{ID}$			128.08 $_{ID}$		
ICC	0.41			0.40			0.41			0.28		
N	63 $_{ID}$			63 $_{ID}$			63 $_{ID}$			63$_{ID}$		
Observations	6683			6683			6683			6683		
Marginal R^2 / Conditional R^2	0.040 / 0.430			0.043 / 0.430			0.041 / 0.430			0.071 / 0.428		

[a] Reference is female.

[b] Reference is grade level 10.

[c] Reference is quarter 1.

[d] Reference is Languages.

Tab. S7 | Linear mixed regression models 4a-e: Effect of absolute sleep characteristics on grades in the flexible system. Predicted outcomes are quarterly grades (0%-100%) in 12 academic subjects from students of cohorts 1 and 2 (n=129) over 1.5 years in the flexible system. 9AM-use is the frequency of ≥9AM-starts in the flexible system. Abbreviations: b, unstandardized coefficient; se, standard error; p, p value; σ^2, variance of residuals of random effects; τ_{00}, variance of ID intercepts of random effects; ICC, intra-class correlation coefficient (describes how much variance is explained by the random effects); N, number of participants; Marginal R^2 describes the amount of variance explained by the fixed effects (predictors); Conditional R^2 describes the amount of variance explained by the full model.

	Model 4a: Chronotype			Model 4b: Sleep duration			Model 4c: Social jetlag			Model 4d: 9AM-use			Model 4e: All		
Predictors	b	se	p	b	se	p	b	se	p	b	se	p	b	se	p
(Intercept)	59.50	5.81	**<0.001**	65.13	10.70	**<0.001**	53.68	3.40	**<0.001**	60.50	2.19	**<0.001**	64.77	14.51	**<0.001**
Gender: Male[a]	-0.31	2.61	0.907	-0.61	2.54	0.810	-0.83	2.54	0.744	-0.72	2.49	0.773	-0.15	2.63	0.956
Grade level: 9[b]	3.35	0.82	**<0.001**	3.35	0.82	**<0.001**	3.36	0.82	**<0.001**	3.27	0.82	**<0.001**	3.27	0.82	**<0.001**
Grade level: 11[b]	0.47	0.62	0.450	0.46	0.62	0.451	0.45	0.62	0.469	0.58	0.62	0.350	0.57	0.62	0.355
Grade level: 12[b]	0.02	1.00	0.984	0.02	0.99	0.984	-0.02	0.99	0.983	0.27	1.00	0.790	0.26	1.00	0.794
Quarter: 2[c]	0.38	0.65	0.559	0.38	0.65	0.559	0.38	0.65	0.559	0.38	0.65	0.562	0.38	0.65	0.561
Quarter: 3[c]	-0.05	0.65	0.933	-0.06	0.65	0.932	-0.07	0.65	0.917	0.02	0.65	0.980	0.02	0.65	0.979
Quarter: 4[c]	1.69	0.61	**0.005**	1.69	0.61	**0.005**	1.68	0.61	**0.006**	1.76	0.61	**0.004**	1.76	0.61	**0.004**
Discipline: Sciences[d]	4.27	0.43	**<0.001**	4.28	0.43	**<0.001**	4.27	0.43	**<0.001**	4.27	0.43	**<0.001**	4.28	0.43	**<0.001**
Discipline: Social Sciences[d]	4.65	0.49	**<0.001**	4.65	0.49	**<0.001**	4.65	0.49	**<0.001**	4.65	0.49	**<0.001**	4.65	0.49	**<0.001**
Chronotype (local time in h)	-0.53	1.23	0.665										-2.37	2.35	0.315
Sleep duration (h)				-1.12	1.47	0.448							-0.27	1.61	0.865
Social jetlag (h)							1.76	1.54	0.256				3.70	2.78	0.186
≥9AM-use (schooldays/week)										-2.12	0.91	**0.022**	-1.32	1.20	0.272
Random Effects															
σ^2	168.88			168.87			168.87			168.87			168.87		
τ_{00}	174.27 $_{\text{ID}}$			173.89 $_{\text{ID}}$			172.93 $_{\text{ID}}$			167.17 $_{\text{ID}}$			168.81 $_{\text{ID}}$		
ICC	0.51			0.51			0.51			0.50			0.50		
N	129 $_{\text{ID}}$			129 $_{\text{ID}}$			129 $_{\text{ID}}$			129 $_{\text{ID}}$			129 $_{\text{ID}}$		
Observations	5111			5111			5111			5111			5111		
Marginal R^2 / Conditional R^2	0.018 / 0.517			0.019 / 0.517			0.022 / 0.517			0.037 / 0.516			0.042 / 0.521		

[a]Reference is female. [b]Reference is grade level 10. [c]Reference is quarter 1. [d]Reference is Languages.

4

Project 3

"School start times and academic performance - a systematic review."

School start times and academic performance - a systematic review.

Authors:

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

*corresponding author

Affiliation:

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

[2] Graduate School of Systemic Neurosciences, LMU, Germany

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

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

Contact information:

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

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

Author contributions (CRedIT Taxonomy)

Conceptualisation: AMB, GZ, EW

Methodology: AMB, GZ, KM

Investigation: AMB, GZ

Data curation: AMB, GZ

Formal analysis: AMB, GZ

Validation: KM

Supervision: ECW, TR

Visualization: AMB, GZ

Writing – original draft: AMB, GZ

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

Abstract

Adolescents are chronically sleep deprived since their tendency to sleep late clashes with early school starting times. One obvious countermeasure is to delay school starts, and there is ample evidence that such a policy is followed by improvements in sleep and health. However, whether these improvements translate into better academic performance - given the central role of sleep in memory consolidation, learning and motivation - is difficult to ascertain given the mixed literature on this topic.

Here, we thus present a systematic review on school starting times and academic performance in middle and high-school students, considering grades and standardised test scores as performance measures. Our detailed analysis of the quality of the evidence, including risk of bias assessment, revealed that the current evidence is insufficient to support that delaying school times improves academic performance. We highlight critical methodological aspects and how to increase the quality of evidence in future studies.

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

Introduction

In many countries around the world, schools start early – often to reduce transportation costs and to adapt to organisational factors, such as after school activities. What early SSTs, however, do not accommodate is teenage sleep. Teenagers should optimally sleep on average 8 to 10 hours per night as recommended by the American Academy of Sleep Medicine[1]. Furthermore, driven my biological and behavioural changes, teenagers progressively delay their sleep window during puberty[2–7]. Early school starting times (SSTs) clash with the longer and later sleep needs of teenagers, leading to wide-spread, chronic sleep restriction in the student population[8–14]. School start changes have been at the centre of many scientific and political debates during the past decades following accumulating evidence that sleep restriction is detrimental for psychological and physical health[e.g. 15–18] and learning[e.g. 19–21], both in the short and long run. Consequently, the first schools, mainly in the USA, decided to delay their SSTs.

During the past ten years, other outcomes with regards to school start times have been investigated, such as cognitive and academic performance. This also culminated in recent scientific and media attention. In many education systems, course grades or standardised scores remain in fact one of the most important aspects of students' future, often determining pathways and possibilities for admission to higher education, jobs, and careers[22–24]. Since short sleep has been linked to detrimental effects on learning, memory, and cognitive performance[25–28], it is fair to hypothesize that delaying SSTs could result in better academic performance mediated by longer sleep duration or improved sleep quality.

However, findings from the literature on this topic are very heterogenous, likely due to methodological differences in outcome variables and study designs[e.g. 29,30]. For instance, academic performance has been operationalised in different ways (e.g. self-reported grades, single final grades, grade point averages, standardized test scores) and with different scales, making comparisons across studies, schools, or countries very difficult. In addition, study designs vary considerably across studies and performance is influenced by many student-and school-level factors[e.g. 31–35]. Therefore, study designs need to account for these confounding factors, requiring highly advanced statistical analyses.

Previous reviews have summarized the effects of delaying SSTs on various variables (e.g., sleep, tardiness, absences, motor vehicle accidents and health), of which some also considered some aspects of academic performance[e.g. 29,30,36]. We identified 11 peer-reviewed reviews that discussed SSTs in relation to academic performance in relation to SSTs but no unifying conclusion can be drawn from them[29,36–45] (only 3 out of the 11 existing reviews were systematic reviews[36,45,46]). Despite this, articles accessible to the general public often claim that later SSTs improve school performance (a Google search of "school starting times" and "grades" in December 2020 gives the following first 3 hits: "*Later school start times linked to better teen grades*"[47], "*Teens get more sleep, show improved grades and attendance with later school start times, researchers find*"[48], "*More Evidence Finds That Delaying School Start Times Improves Students' Performance, Attendance, and Sleep*"[49], while some public outreach programs, such as infographics, also convey this message[50]).

Since performance shapes future career trajectories, answering the question whether delaying SSTs improves academic performance therefore goes beyond simple and genuine scientific curiosity - a rigorous analysis of the current evidence is warranted. In our review, we therefore followed the PRISMA guidelines for systematic reviews to ascertain a high-quality reviewing process. We identified methodological concerns of previous reviews, such as missing discussion of the quality of evidence, missing detailed description of the type of outcome variable and statistical analysis, and inclusion of both high-school and college students, who differ in terms of their sleep characteristics and class schedules. The aim of this review was therefore to systematically search and review both peer and non-peer reviewed articles (to reduce publication bias) on SSTs and course grades or test scores in middle and high-school students, providing an overview of the results and an overall description of the quality of the evidence, according to the PRISMA guidelines.

Methods and Materials

Literature search

Our focused question was whether changes in school start times in middle or high-schools (or international equivalents) have any effect on academic performance as measured in (standardised) test scores or course grades (both subjectively and objectively reported). To this end, 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. All languages, article types or year of publications were allowed. The following search terms were used (in title, abstract or the 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 *standardized scores* OR *achievement*

Additionally, reference lists of previous reviews and articles were scanned to ensure complete retrieval. We included two unpublished articles that are currently under review in a peer-reviewed journal[51,52].The PRISMA flowchart (Fig. 1) was followed to adhere to preferred reporting guidelines for systematic reviews[53].

Study selection criteria

All duplicates were removed via the Endnote duplicate function or manually in case the software failed to pick it up. All titles and abstracts were subsequently screened on relevance with regards to the focus question. Full articles were only searched if the following study selection criteria were fulfilled: grade or score analyses were explained; participants were middle-school or high-school students; articles included both a change/variation in SSTs and a measure of academic performance (course grades or scores).

Data abstraction and analysis

The recommended PRISMA guidelines for data synthesis and systematic reviews were followed[53]. AMB and GZ independently and systematically extracted pre-defined study characteristics (Tab. 1).

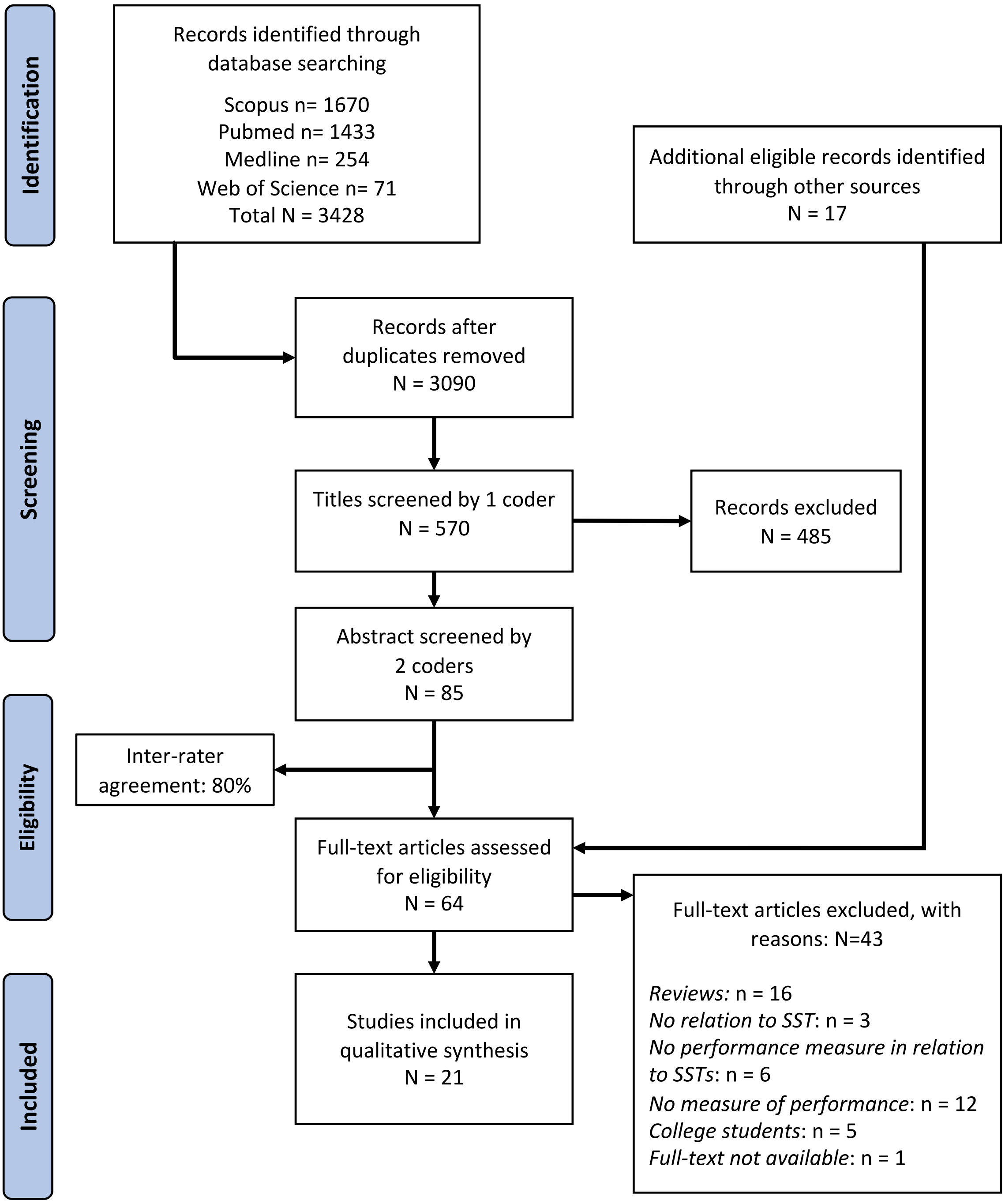
Identification
Screening
Eligibility
Included
Records identified through database searching
Scopus n= 1670
Pubmed n= 1433
Medline n= 254
Web of Science n= 71
Total N = 3428
Additional eligible records identified through other sources
N = 17
Records after duplicates removed
N = 3090
Titles screened by 1 coder
N = 570
Records excluded
N = 485
Abstract screened by 2 coders
N = 85
Inter-rater agreement: 80%
Full-text articles assessed for eligibility
N = 64
Full-text articles excluded, with reasons: N=43
Reviews: n = 16
No relation to SST: n = 3
No performance measure in relation to SSTs: n = 6
No measure of performance: n = 12
College students: n = 5
Full-text not available: n = 1
Studies included in qualitative synthesis
N = 21

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.

Studies were grouped by study design, i.e. longitudinal design with a control group, without a control group, and cross-sectional design. Note that a longitudinal design incorporates that *individual* students are followed over several time points (cohort study); a cross-sectional study compares *different* students at one time point, or over several time points. It was noticed that several cross-sectional studies described their design as longitudinal because they followed the same schools or districts over several time points (which might or might not include similar students). Authors were contacted if information was missing, not clearly defined or further analyses were available upon request. If authors responded, information was updated accordingly. If authors did not answer or failed to provide necessary information in the original article, this was marked as “NA” in Tab. 1 and flagged orange or red in the reporting bias category in Tab. 2.

Risk of bias assessment

AMB and GZ systematically conducted a pre-defined risk of bias assessment. There were no randomised controlled trials (RCT) in the final sample. Due to large methodological differences between studies bias assessment guidelines for RTC had to be adapted (given that there are no standard guidelines for non-RTCs). To this end, items from the GRADE scheme[54] and ROBINS-I tool[55], which are used for non-RTCs, were included and modified. Each study was evaluated on the following bias categories and flagged green (=low risk), orange (=intermediate risk) or red (=high risk):

Selection bias (randomization): participants were randomly assigned to the control group or the treatment group. Non-RTC are high risk by definition.

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

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

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

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

(Dis)similarity of baseline characteristics: characteristics between cross-sectional groups or between control and treatment groups were checked.

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

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

Tab. S1 contains the decision criteria for the risk of bias assessment. Mutual agreement was sought after discussion of critical points between scorers AMB and GZ. In case no agreement was possible, a third, independent scorer (KM) made the final decision. A total score was calculated as follows: green contribute 1 point, orange 0.5 points and red 0 points towards the overall score. The maximal possible score was 8 for the longitudinal studies with a control group, 6 for the longitudinal studies without a control group, and 7 points for cross-sectional studies. A good evidence score was then calculated as the percentage of the maximum score (e.g. 6 out of max 8 points = 75%). The different bias categories were not weighted.

Results

Literature search

When applying the search terms to title, abstract and keywords, a total of 3,428 articles were identified of which 3,090 remained after duplicate removal (Fig. 1). Due to this large number, studies were further screened only based on their title which resulted in 570 articles. One coder (AMB) then excluded 480 manually due to irrelevant titles. The abstracts of the remaining 85 studies were screened by both coders (AMB and GZ), who agreed on 47 studies (80% inter-rater agreement) and additionally identified 16 studies through reference lists of included studies. Two additional studies which are currently under review were also included[51,52]. Hence, 64 full-text articles were screened of which 21 were included in the qualitative synthesis. Forty-one studies were excluded based on the pre-defined exclusion criteria (Fig. 1).

Study characteristics and quality

In the following, summary information concerning all included studies are reported (see also Tab. 1).

School type and cohort characteristics

The majority of studies collected data in high-schools (>900 schools), of which 2 were also boarding-schools[56,57], 2 grammar schools and 2 vocational schools[58]. Other school types were 119 middle schools and 85 elementary schools (the latter were not considered here). In two studies school type was not specified[59,60]. Quite strikingly, the sample sizes varied drastically between 157 to >770,000 individual students and up to >1 Mio number of observations. 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 from approximately from 11-19. Most studies were conducted in the US (13)[56–69], followed by South Korea (4)[52,70–72], Germany[51], Croatia[58], England[73], and one unknown location[74] (Fig. 2a and Tab. 1). Gender ratios, ethnicity/race and a proxy for socioeconomic status (SES; free or reduced lunch eligibility) were not consistently reported.

Study designs

We mainly identified longitudinal and cross-sectional studies. The 11 longitudinal studies always included a change in SSTs and hence had an intervention group[51,55,56,59,63–65,69–72]. However, only 6 studies had an additional control group with no change[52,65,70–72] or advance of SSTs[60] (Fig. 2b). Several studies with a change in SST did not follow individual students but conducted longitudinal comparisons of schools or districts over one[69] or several years[59,68], or at one time point after the change[62]. Four studies compared several schools in various districts without an intervention but based on their different school start times[58,61,67,74]. One study also had an A-B-A design, in which the school start delay during phase B was abolished to return to baseline start time (A)[73]. This was presumably a cross-sectional design comparing the school with a national average over several years (with 2 years of overlapping students) but no clear judgment was possible, thus the study was classified as "unclear". Similarly, one other study[63] very likely cross-sectionally compared grades in schools; again the study was classified as "unclear". Authors of both studies were contacted to clarify but could not be reached.

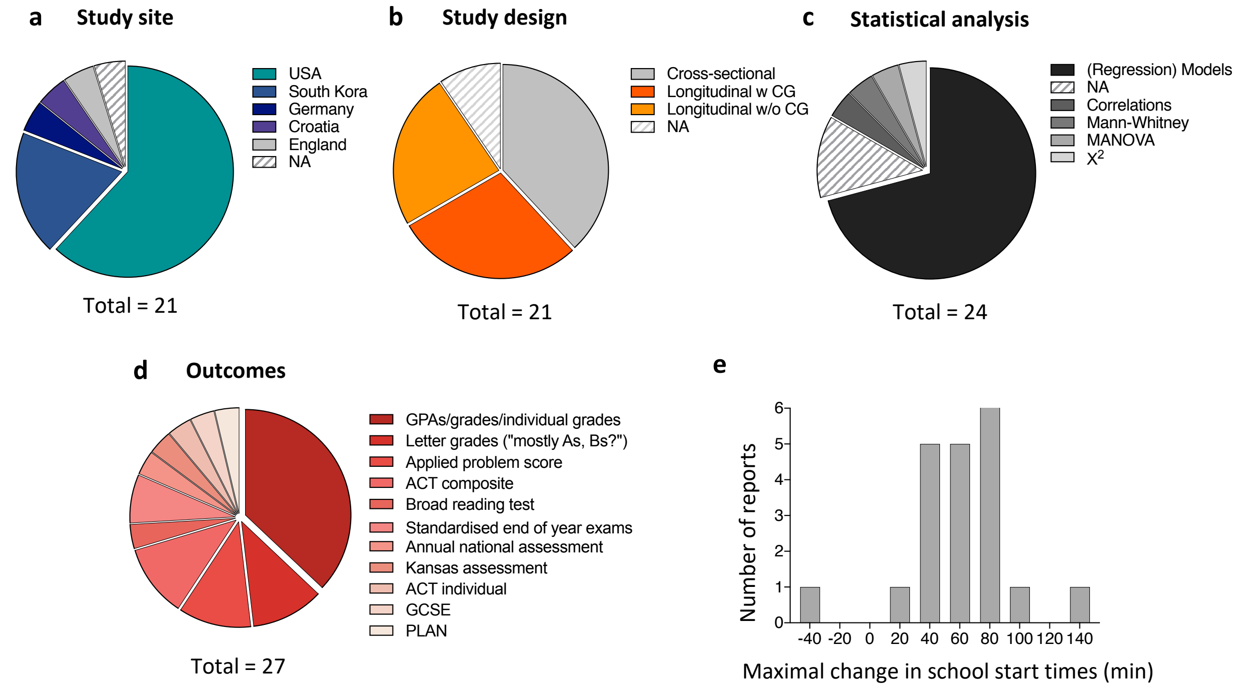

Fig. 2 | Characteristics of included studies. a-d, Pie charts depicting key characteristics of the 21 studies included in the final review. Since several studies used multiple types of analysis or assessed multiple outcomes, the total number in c,d is >21. **e,** Histogram displaying the magnitude of the school start changes reported in the 21 studies. When a study reported ranges, the maximum of the range was taken. Please note that these numbers therefore just provide a rough overview and are far from 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, preliminary ACT.

Statistical analyses

A vast range of different statistical analyses was reported (Tab. 1 and Fig. 2c). Notably, especially regressions were dominantly used, ranging from general OLS regressions[59,60,67,68,70], quantile regression[60], difference-in-difference methods[52,65,70,72], binomial regression[69,71], linear mixed models[51,66] to path analysis with probit regression[74]. One study reported Oster models with bounded effects and instrumental estimates[67]. Another[61] study used MANOVA, while more simpler analysis which did not control for covariates were also used. These were t-tests[63,66,73], X^2-tests[57], Mann-Whitney Test[58] and correlations[63]. Kelley *et al.* used t-tests and made value-added predictions about the school performance compared to the national average[73]. Several authors did not report statistical analysis[56,62,64].

Study outcome measures

Authors did not always provide complete explanations as to whether tests scores were standardised making clear distinction between course grades and scores a challenge. Clearly defined scores were ACT scores (American College Test)[59,65,68], national achievement scores or PLAN scores[63], standardised test scores from Regents Exams[66], standardised end-of-course exams[68], annual national assessment of achievement in South Korea[72], GCSE in the UK (General Certificate of Secondary Education)[73], and Woodcock-Johnson Revised Test of Basic Achievement scores[67], all of which were objectively reported (except for Groen *et al.* which was unclear[67]) (Fig. 2d). The remaining studies presumably analysed unstandardised scores or objective grades[51,52,60,61,64,68,70] (Milic *et al*. was unclear[58]) and subject grades[56,57,62,63,69,71,74]. Sampling resolution was mostly once per year, the highest reported resolution was once per academic quarter[51].

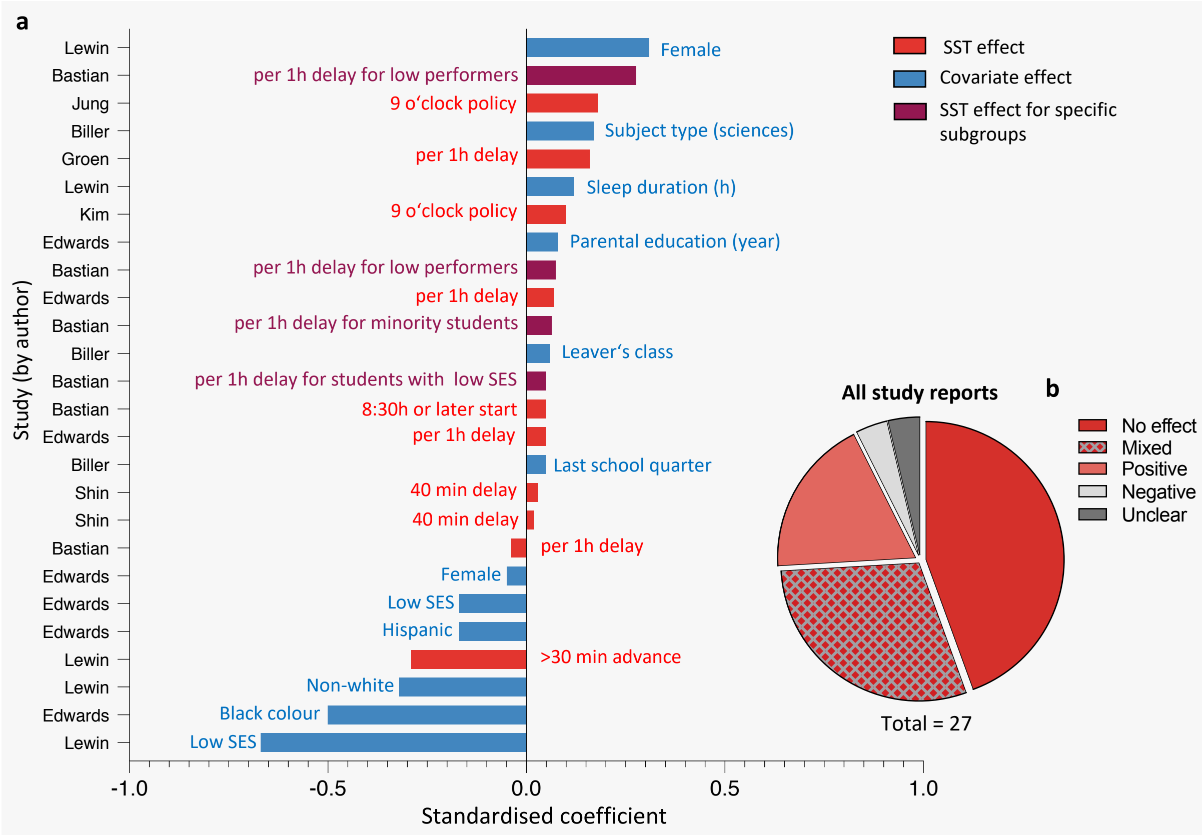

Fig. 3 | Overall study results and effects sizes. a, Standardised beta coefficients ordered by magnitude and study author from a subset of studies that reported standardised coefficients and statistically significant effects (n=8 of 21). Only these statistically significant effects are depicted, non-significant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

Amount of school start time change

The maximal delays that studies reported was on average 63.6 min (median=60, SD=26.4) with a range of 25 to 135 min (Fig. 2e). This is an approximation of the maximal possible delays reported in the studies and not the exact amount that each school per study changed. Since some studies only provided SST ranges or a minimal start delay, no precise numbers can be given here. In 2 studies, SSTs were actually advanced by 40 min and 25-45 min respectively[60,65]. One study changed to a flexible SSTs in which students could choose daily whether to attend school at 8:00h or 8:50h[51].

Magnitude of effects

In order to compare the magnitude of (statistically significant) positive results, standardised beta coefficients were compared across studies and with other covariates where these were reported (Fig. 3a). Non-white students and students with a lower SES showed performance disadvantages, while gender differences varied in both directions. One study demonstrated the harmful effects of advancing start times on ACT scores by at least 30 min[74]. Overall, Fig. 3a shows that the magnitude of the influence of school start times is smaller than students' SES or their ethnical/racial background. In line with this, studies from Edwards[60] and Bastian and Fuller[68] demonstrated that disadvantaged and minority students particularly benefitted from later starts. Note that only statistically significant and standardised coefficients are displayed in Fig. 3, which gives a biased picture towards positive results of school start times on performance (see also Fig. 3b which puts findings from 3a into the overall perspective of all included studies).

Summary of individual study results

In the following, we shortly report findings of all included studies grouped by their study design. Altogether, 5 studies found clear positive effects of delayed school start changes on academic performance[52,60,62,73,74], 5 studies reported 8 mixed effects[61,63,67,68,72], 9 studies did not detect significant effects (12 individual outcome reports)[51,56,57,59,64–66,70,71], one study reported negative effects[58], and one study's finding was unclear[69] (Fig. 3b). One study reported that an advance of at least 30 min was associated with decreased self-reported grades[74], while Lenard *et al.* did not find that an advance hampered ACT scores[65]. Four studies investigated the same 9 o'clock policy in South Korea[52,70–72]. Although they considered partly different outcomes and schools (middle vs high-schools) they likely analysed data from overlapping students, hence this cannot be entirely regarded as independent evidence. The same applies to Wahlstrom *et al.* who conducted several studies in the same district: the earliest report from 1997[62] might have been followed up longitudinally in 2004[75] but due to missing descriptions this is not entirely clear to the reader.

Longitudinal studies with control group

Edwards (2012)[60] followed several middle schools in Wake County, North Carolina (USA) over 8 years (up to $N_{observations}$>102,000) of which 9 schools delayed, 4 advanced and 11 did not change their SSTs. The authors analysed objective standardised end-of-year test scores in reading and math via regression models with pooled OLS models and accounted for various covariates both on the student and school level. They found that a 1h later school start corresponded to a 1.8-2.9 percentile increase in math (0.06-0.07 SD) and 1.0-3.6 increase in reading (0.04-0.05 SD) when adjusted for covariates, and that the effect was stronger for lower achieving students.

Jung (2018)[70] followed 85 elementary and 63 middle schools ($N_{students}$>4,000) in South Korea 3 years prior to and 2 years after a delay from 8:00h-8:20h to 9:00h. Participants were recruited as part of the Gyeonggi Education Panel Study and their objective Korean, English and math course grades were reported. The author found no effect for the longitudinal comparison with the control group (difference-in-difference estimation/OLS estimation). Similar to Kim [72] and Biller *et al.*[51], the author also found that when not controlling for covariates, test scores increased, while the effect became non-significant (statistically and biologically) when covariates were added.

Kim (2018)[72] also compared high-schools from two districts in South Korea ($N_{students}$>2,000), of which Gyeonggi adopted a 9 o'clock start time policy. Pre-change SSTs in this district ranged from 7:40-9:00h and were delayed to 9:00h post-change, while Seoul did not change (control group). The author used the difference-in-difference method and mixed within-between regression models to estimate the influence of the 9 o'clock policy on the objective Annual National Assessment of Educational Achievement for 9th and 11th graders, and the College Scholastic Ability Test (CSAT) for 12th graders (data cover 5 years pre and 2 years after the change). Only male 11th graders showed an increase of 0.06-0.08 SD for math, even after adjusting for confounders. CSAT scores did not increase significantly with the 9 o'clock policy.

Similarly, **Rhie and Chae (2018)**[71] studied South Korean districts of which Gyeonggi delayed SSTs (baseline from a range of 7:30h-8:10h) to 9:00h and Daegu, Gyeongbuk and Ulsan did not (SSTs range from 7:30h to 8:00h; control group). In their very large sample ($N_{students}$>42,000) from middle and high schools they found that self-reported GPAs increased year by year in both the intervention and the control group (data cover 2 years pre and after the change). Their logistic regression thus did not detect any significant benefit of delaying SSTs.

Shin (2018)[52] is the fourth study which investigated the South Korean 9 o'clock policy effects in Gyeonggi (change in SST from around 8:20 to 9:00 AM), compared to Seoul (control group), but the author used objective semester grades as outcome and focused on middle schools ($N_{observations}$>33,000). The data span 2 years and was analysed using the difference-in-differenced method which accounted for various individual and school-levels variables. Shin reported an 0.03 SD increase in math and 0.02 SD increase in reading grades when adjusted for time trending.

Lenard *et al.* (2020)[65] found no significant change in objective standardised American College Test (ACT) scores, neither in their longitudinal nor their cross-sectional comparison of about $N_{students}$~10,000 students in 8 cohorts in Wake Country, North Carolina, USA. The authors looked at 19 high schools of which 5 had advanced their SSTs from 8:05h to 7:25h, while the control group (14 high schools) kept their start at 7:25h. Their data spanned 4 years prior and 7 years after the change. They had also controlled for various individual and school-level variables.

Longitudinal studies without control group

Quite uniquely, **Biller *et al.* (2020)**[51] investigated the effects of *flexible* SSTs on sleep and objective, quarterly grades of senior students of a German high school for up to 2.5 prior and 1.5 years after the change. Students chose daily whether to attend school at 8:00h or 8:50h. Longitudinal linear mixed model analyses of ~17,000 grades of 12 academic subjects pooled into 3 disciplines ($N_{students}$=157) indicated that the flexible system did not have a positive effect on grades when accounted for several student and school-level factors.

Boergers *et al.* (2014)[56] studied an independent U.S. high school (boarding school) in Rhode Island that delayed its start time from 8:00h to 8:25 ($N_{students}$=197). The percentage of students who reported to obtain "mostly Bs or better" changed from 93% to 91% after 2 months, however statistics were not reported.

Owens *et al.* (2010)[57] used the same outcome variable as Boergers *et al.*[56] in their study of $N_{students}$=201 from an independent US high-school (boarding and day school) in Rhode Island (USA) over 6 months (3 time points of assessment). They found that a school start delay from 8:00h to 8:30h was associated with a non-significant increase of students reporting to mostly obtain Bs or better (82% pre vs 87.1% post, using a χ^2 test).

Thacher & Onyper (2016)[66] studied $N_{students}$~800 across 4 years from one public high school in Glen Falls, NY (USA) which delayed their SSTs from 7:45h to 8:30h. They used mixed effect analyses to analyse longitudinal effects (2 years before and after the change), adjusting for multiple covariates and including moderator effects. This analysis indicated no systematic positive effect on subjectively reported GPAs (0-100%) nor subject-specific GPAs or standardised test scores (Regents exam). They did find positive effects for 11th overall GPAs, however, only when they ran cross-sectional comparison (increase from 78.79% to 81.34%). In contrast, no systematic effects on individual academic subjects were found in this cross-sectional analysis. In fact, 2 out of 20 subjects were significantly worse after the change and also Regents exam scores decreased significantly.

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

Cross-sectional studies

Groen and Pabilonia (2019)[67] studied $N_{students}$=1200 from a sample of 790 U.S. high schools, and reported that a 1h-delay in high school start times was associated with increased reading scores (but not math scores) by 0.16 SD for females ($p \leq 0.1$), while no significant effect was found for males. The authors used OLS models, including many covariates (individual, family, high school, and community characteristics) that were added sequentially to the models. Data came from 2 years, sampled once per year.

Hinrichs (2011)[59] found no association between SSTs and ACT scores ($N_{students}$> 196,000) after a delay of 85 minutes from 7:15 to 8:40 AM in 73 schools in Minneapolis when accounting for various student-level and district level covariates and the length of the school day using OLS regression models (9 years of data). In a similar analysis, the author also found no effect on Kansas assessment scores in reading, maths, science, and social disciplines including all public high-schools in Kansas (1,666 schools; up to 5 years of data). In another sample of 75 schools in 19 districts in Virginia, again no association was found between delayed SST and test scores in standardized end-of-course exams (8 years of data).

Bastian and Fuller (2018)[68] sampled data in 410 high schools in North Carolina for 4 years of which 23 changed their start times (9 schools by ≥30 min). The authors tested both the influence of a linear SST delay per 1h and a categorised school start depending on actual start time in $N_{students}$>770,000 on overall and 1st period course grades, standardised end-of-course exams, and ACT scores. Linear regression models showed that only a start at 8:30h or later was associated with 0.05 SD improvements in 1st period course grades. Importantly, this was one of the only studies that particularly focused on specific subgroups of students: especially low-performers, students with a minority background and with low SES benefitted per 1h later starts (0.05-0.07 SD in course grades and up to 0.28 SD in ACT composite scores).

Dunster *et al.* (2018)[69] reported results from a cross-sectional comparison of one semester biology lab course grade from 2 high schools in Seattle which delayed their starts from 7:50h to 8:45h (median grade was 77.5% in year 1 and 82% in year 2; N=178). Using generalised linear models (binomial), they assessed whether *years* differed based on the predictor *grade* (as described in the methods and the analysis script). Better grades were predictive of year but this was after adjusting for sleep offset, chronotype and sleepiness. With year as the dependent variable and controlling for differences between the years, this analysis was not suited to answer whether later school starts are associated with better grades.

Milić *et al.* (2014)[58] analysed the final semester grades of 4 Croatian schools (grammar and vocational schools) with alternating morning and afternoon schedules at one time point; 2 schools followed early schedules (7:00 and 13:00), while 2 schools had later schedules (8:00 and 14:00). Based on their sample of $N_{students}$=821 they concluded that students attending the early schedules got better grades ($p<0.001$; 72.0% vs 65.6% in the later scheduled schools). A Mann-Whitney Test was used which did not consider any covariates. Additionally, their samples consisted of three times more boys in the early scheduled schools which might interfere with their results, given that gender very likely influences grades.

Wolfson *et al.* (2007)[61] compared the average fall quarter grade (0-100%) of a total of $N_{students}$=205 attending either an early middle school (starting at 7:15h) or later middle school (starting at 8:37 AM) in New England (USA). MANOVA results with school, grade and gender as predictors indicated that 8th graders in the later school obtained better objectively reported grades than their early school peers ($p<0.01$, 83.79% vs 76.85%) while no difference was found for 7th graders after half a year.

Lewin *et al.* (2017)[74] compared 26 middle schools clustered into 3 groups depending on their SSTs (earliest, early, late). The authors obtained self-reported grades ("mainly As", "mainly Bs", "mainly Cs", "mainly Ds/Fs") and sleep duration from $N_{students}$>32,000 students in 3 years (unknown location). Path-analyses with probit regression with grades as outcome, sleep duration as mediator and inclusion of several covariates showed better grade estimates in the latest SST group compared to the earliest schools) but not to the early schools. This was in a similar magnitude range as gender (females better than males) and ethnicity (non-whites worse than whites). Free lunch status as a proxy for SES clearly had the biggest impact on grades, while the influence of sleep duration as a mediator was smaller but still significant.

Wahlstrom *et al.* (1997)[62] compared high schools in three districts in Minnesota (USA) of which district A delayed its start time for high-schools to 8:30 and district B and C stayed with their earlier starts of 7:25h and 7:15h respectively. The author found that mean self-reported grades in district A were highest compared to the other 2 districts, however statistical analyses and the use of covariates were missing. Results for middle schools (7-8th graders) were comparable but again no statistics were given and differences were marginal.

Unclear study design

Kelley et al. (2017)[73] followed English high school students (N>2,000) for 1 year with a school start at 8:50h, for 2 years with a delayed start at 10:00h, and for another year in which the school changed back to the original start time (A-B-A design). As the only study in our selection, the authors measured the percentage of students making good academic progress (*i.e.,* achieving 5 or more GCSE grades of C or better in English, math and at least 3 other subjects) compared to the national average. Using t-tests they found that the delayed SST was associated with an 12% increase in the value-added number of students making good academic progress in GCSE exams taken at age 16. The value-added number describes the percentage of good academic progress above the predicted outcome from the year before. This study was classified as "unclear study design" since we were unable to judge the design with certainty given the longitudinal and cross-sectional parts with probably partly overlapping students across years.

The same applies to **Wahlstrom *et al.* (2014)**[63] who analysed self-reported grades**,** objectively reported GPAs, and standardised test scores (state-wide achievement tests or PLAN) from 9-12th graders over several years (but presumably not the same students) after a school start delay from 7:35-7:50h to 8:00-8:55h in Minnesota, Colorado, and Wyoming (USA), and found mixed and mostly not-significant effects. The author used t-test and correlations and did not consider covariates.

Risk of bias assessment

A systematic risk of bias assessment was performed to judge the evidence quality of included studies (Tab. 2). Conducting a meta-analysis was not sensible given the large differences between studies in terms of outcomes, study design, interventions, and the type of analyses. Since none of the studies were RCTs, selection bias (no randomisation or allocation concealment) was high by definition for all studies. Most longitudinal studies with a control group performed best, with 2 out of 6 studies achieving a good evidence score of at least 75%[65,70] and 3 more studies achieving a 60% score[52,60,72]. In fact, 2 studies could have improved their score to >75% simply by ensuring sufficient reporting of outcomes and statistical analyses[52,72]. Five studies followed the same students longitudinally, however did not provide any control group which unfortunately decreased their evidence quality[51,56,57,64,66]. Furthermore, these studies mostly suffered from blinding issues (all 5), inappropriate use of statistical models (3 of 5)[56,57,64], and reporting bias[56,64] (2 of 5). Within the cross-sectional studies, 3 of 8 stood out with having a good evidence score of at least 75%[59,67,68]. The remaining 5 studies lacked blinding[58,61,62,69,74], 5 reported but did not control for dissimilarities of baseline characteristics[58,61,62,69,74] and in 3 studies appropriate statistical tests were not fulfilled[58,62,69]. Two studies could not be classified; thus no overall score was given[63,73].

Taken together, in many studies basic reporting standards were only partially met, blinding was a high concern in over half of the studies (performance bias was high), and appropriate statistical models which control for confounders were not used in 8 of 21 studies (with 3 intermediate risk classifications). With over half of the studies not reaching at least 60% of our good evidence score, the quality of the evidence is only moderate. On the positive side, especially the longitudinal studies with a control group showed a high evidence quality. Furthermore, all included studies had appropriate large sample sizes (and/or high resolution) and were therefore very likely suited to detect a true effect (*i.e.,* power was presumably sufficient).

Discussion

Chronic sleep restriction in teenagers has become a serious health concern worldwide[e.g. 11,76]. The widespread sleep restriction is largely a result of the conflict between the biologically late sleep times typical of adolescence and the early SSTs imposed by society[e.g. 77–79]. Delaying bell times has the great potential of improving cognitive functioning, physical health and well-being of students by improving sleep (as reviewed elsewhere[30,38,42]) with possibly relatively little costs[80,81]. But does a delay in SSTs also translate into improved academic performance in middle and high school students? We conducted a systematic literature search and identified 21 studies that investigated whether SSTs have any systematic effect on course grades or test scores. The results showed that the current evidence is not sufficient to answer this question with any certainty, which is partially due to methodological shortcomings in the presented studies.

Methodological considerations

Our systematic risk of bias assessment showed that the evidence was mostly moderate (only 8 out of 21 studies achieved a good evidence score of at least 60%). Specifically, we did not identify any randomised controlled trial (as already lamented by a Cochrane review[36]), which is not surprising considering the given circumstances and hesitation of many schools to participate in complex and time-consuming study designs[82]. We identified longitudinal studies with a pre-post design that followed a specific cohort of students over time i) including a control group that did not change start times, and ii) without a control group. A second common design were cross-sectional studies that compared different, independent groups of students (either at one specific time point or over several years) with varying start times. Studies that performed best in the risk of bias assessment were mostly longitudinal studies with a control group, a large sample size and with appropriate and advanced statistical analyses that controlled for possible confounders.

Conclusions from good evidence studies

What do studies with low risk (*i.e.,* a >75% good evidence score) conclude about the influence of SSTs on academic performance? Lenard *et al*. (2020)[65] found that advancing SSTs by 40 minutes did not affect ACT scores, while Jung (2018)[70] showed that delaying start times by 40-60 min only improved grades when personal covariates were not controlled for. If studies with a good evidence score of 60% are also considered, the picture is more complex: two studies report small gains in math and reading[52,60], and one reported small effects on math but not on Korean nor English[72]. Three cross-sectional studies also achieved a good evidence score of over 75%[59,67,68]. The associations found between SSTs and academic performance again did not point in one direction: Groen and Pabilonia (2019) considered a range of different start times and reported small increases on the Woodcock-Test but only for females and reading[67], 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[59]. Bastian and Fuller (2018) reported that only a 8:30h or later start was necessary for positive associations with 1st period grades[68]. Furthermore, the authors showed that especially lower achievers, minority students and students with a low SES benefit from later starts. In summary, good evidence studies report either no, relatively small, or not generalisable effects of changing SSTs – a more nuance effect is thus very likely.

Do result for course grades and scores differ?

Maybe results become more consistent when we distinguish between effects on test scores and grades? 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[83] or racial segregation[84]. However, general test scores usually measure the accumulated knowledge over several schooling years making them possibly less sensitive to acute changes in SSTs[65]. They are also often scheduled at the beginning of the school day[59] and therefore confounded by time-of-day effects on attention and fluid intelligence (*e.g.* logic, reasoning, problem solving)[14,65,85,86]. Moreover, in the case of ACT or PLAN scores, tests are usually only taken by high-achieving students applying for admission to college – a specific student population which is prone to ceiling effects, making these students less likely to benefit from later SSTs compared to lower-achieving students as 2 other studies also confirmed[60,68].

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 also potentially more sensitive to acute SSTs changes and less influenced by time-of-day effects if distributed evenly across the day. On the down side, grades might be more influenced by certain student characteristics, such as conscientiousness or perseverance[87], and more subjective since they are given by teachers. Such a “teacher bias” could particularly influence the results of interventional studies if not controlled for. Altogether, both standardised test scores and course grades have their pros and cons, which might be the reason why no clear answer emerges even when results are grouped by outcomes: there was no tendency or differential effect on either objective test scores: 1 positive[73], 5 null findings[59,65,66], and 5 mixed results[63,67,68,72]; objective grades: 2 positive[52,60], 4 null findings[51,64,66,70] and 2 mixed results[61,68]); or self-reported grades: 2 positive[62,74], 3 null findings[56,57,71], and 1 mixed[63], negative[58] and unclear result respectively[69]).

Considerations of power and dose

An alternative explanation could be that studies were not large and sensitive enough to systematically detect any effect, *i.e.* that their power was not sufficient. Even though most authors did not report power specifically, this is also very unlikely: almost all studies had very large sample sizes (up to > 1 Mio) and were able to detect other influences such as gender differences, performance gaps between whites and non-whites and poorer performance in students with lower SES backgrounds. The effect sizes of these factors tended to be of larger magnitude than effect sizes for changes in school starts (Fig. 3a).

Another interesting consideration is that effects of SSTs on performance might not be linear. When exactly should schools start? How much should schools delay their bells? 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[59] tried to model this hypothesis using spline regressions but found no clear results. Furthermore, the latest start time in the studies reviewed here was 10:00h and the largest delay was 135 min. 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. Nevertheless, the American Association of Pediatrics recommends to start schools not earlier than 8:30h[88] which is supported by Bastian and Fuller[68] who found that only when school started at 8:30h or later, significant positive effects were detected on 1st period grades, although overall grades were unaffected. A second consideration about dose is how long the school has already operated in a delayed system – the longer the delay has been in place, the longer students were exposed. Several studies considered time trends for several years before and after a change (when the delay was kept) but again, no unifying evidence was reported from these studies.

Factors influencing academic performance

A very likely reason for inconclusive results is the many different variables affecting course grades and test scores which add a layer of complexity to an apparently simple question. 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[89], ethnic or racial background[68], conscientiousness[87] or prior knowledge[90]) to family-level factors (e.g. parental involvement[91], parental education[70], or SES[92]), and to classroom- and school-level factors (*e.g.* classroom size[83] and atmosphere[90], teacher quality[93]). Indeed, we also observed here that SES and race/ethnicity influence performance (Fig. 3a). Moreover, there are sleep variables, such as chronotype, sleep duration and daytime sleepiness that play an important role for health, cognition and learning and are often connected to demographic variables, such that students with difficult social backgrounds are also prone to reduced and poorer sleep than their more advantaged peers[94,95]. It is therefore likely that only a subset of students really benefit from later SSTs. Stratified analyses could answer this question, but have rarely been done (for notable exceptions see[60,68] which confirm such tendencies). In general, reflecting on confounders, their influence on academic performance and on how they might also be affected by changes in SSTs is important for designing future studies.

Suggestions for future studies

In addition to these considerations, several general methodological aspects have been identified that could be improved in future studies:

- Provide graphs that depict study designs to facilitate understanding

- Sufficient description of basic demographics of the studied population (including $N_{students}$ and $N_{observations}$, grading scales, descriptions of the outcome and a brief overview of the educational system) should be ensured
- Be aware that cross-sectional comparisons do not allow for causal interpretation
- Collecting objective grades (from a range of different academic subjects and across the year) or standardised test scores is recommended; self-reported composite grades or scores with low resolution should be avoided
- Randomisation could be achieved at the class or school level
- Control groups are feasible and important
- To control for placebo effects, it is possible to *e.g.* assess the expectations of students, teachers and parents
- For the highest evidence quality, longitudinal intra-individual assessments with a pre-post assessment and a comparison group are needed
- It is essential to use appropriate statistics for the given study design which considers the influence of covariates. The lack of covariates and accounting for trending can result in misleading conclusions even when good study designs were implemented
- Is the effect size meaningful for the individual student? Put findings into perspective

Limitations of the review

In this systematic review, the PRISMA guidelines were followed to perform a systematic research and assessment of the literature on SSTs and academic performance. Although an extensive search across different databases was carried out independently between AMB and GZ, 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 the number of studies on the same topic reported in previous reviews (2-12 studies). We also chose to report not peer-reviewed articles to reduce a possible publication bias in favour of positive results. Other previous reviews[e.g. 46] 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 not peer-reviewed articles. Finally, both AMB and GZ independently and extensively retrieved pre-defined information from all included primary studies. Where information was not available, authors were contacted to ensure correct reporting. Since the studied population was restricted to middle and high-school students, several studies which used randomisation at the class-level had to be excluded (for a review see[29]). High-school life and type of students widely differ compared to college, and adolescents are more prone to sleep restriction due to their altered physiology, thus we focused on this specific group. We nevertheless included middle schools, since sleep changes tend to start with the onset of puberty [2,96].

Conclusions

Our systematic research and analysis of the literature shows that there is no clear evidence that delaying SSTs improves or is associated with academic performance. Here we identified some methodological aspects and give suggestions that could improve the quality of future studies. We suggested to conduct more stratified analyses to identify specific groups of students (*e.g.* late chronotypes, disadvantaged students) who might benefit the most from delaying SSTs and to conduct mediation analyses or dose-response analyses. We still do not know by how much schools should delay or at what hour it would be ideal to start school for all chronotypes or other subgroups. Importantly, as much as course grades and test scores do not systematically or greatly improve - at least given the current evidence - they very likely also do not become worse with later school starts. This means that SSTs could be delayed, while academic performance is maintained at the same level or possibly achieved with less cognitive effort or time spent on studying and homework (students are likely better rested and therefore better cognitively capable and more efficient). In addition, students could benefit from other reported positive outcomes such as longer sleep, less daytime sleepiness, improved mood and motivation, decreased computer gaming, higher attendance rates and fewer tardies and suspensions[e.g. 29,36,38,45,97]. In particular, the possibility to reduce the widespread chronic teenage sleep restriction, and all the related acute and chronic health problems[15–18], is a valid and sufficient argument in favour of delaying SSTs.

References

1. Paruthi S, Brooks LJ, D'Ambrosio C, Hall WA, Kotagal S, Lloyd RM, et al. Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine. *J Clin Sleep Med*. 2016;12:785–6. doi:10.5664/jcsm.5866.

2. Roenneberg T, Kuehnle T, Pramstaller PP, Ricken J, Havel M, Guth A, et al. A marker for the end of adolescence. *Curr Biol*. 2004;14:1038–9. doi:10.1016/j.cub.2004.11.039.

3. Wolfson AR, Carskadon MA. Sleep Schedules and Daytime Functioning in Adolescents. *Child Dev*. 1998;69:875–87. doi:10.1111/j.1467-8624.1998.tb06149.x.

4. 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. doi:10.1371/journal.pone.0112199.

5. Fischer D, Lombardi DA, Marucci-Wellman H, Roenneberg T. Chronotypes in the US – Influence of age and sex. *PLoS One*. 2017;12:1–17.

6. Tonetti L, Fabbri M, Natale V. Sex difference in sleep-time preference and sleep need: A cross-sectional survey among Italian pre-adolescents, adolescents, and adults. *Chronobiol Int*. 2008;25:745–59.

7. Randler C. Age and gender differences in morningness–eveningness during adolescence. *J Genet Psychol*. 2011;172:302–8.

8. 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 Public Health*. 2006;6:1–9. doi:10.1186/1471-2458-6-116.

9. 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. doi:10.1016/j.smrv.2011.03.005.

10. Keyes KM, Maslowsky J, Hamilton A, Schulenberg J. The Great Sleep Recession: Changes in Sleep Duration Among US Adolescents, 1991-2012. *Pediatrics*. 2015;135:460–8. doi:10.1542/peds.2014-2707.

11. 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. doi:10.1016/j.sleep.2010.11.008.

12. Paksarian D, Rudolph KE, He JP, Merikangas KR. School start time and adolescent sleep patterns: Results from the US National Comorbidity Survey-adolescent supplement. *Am J Public Health*. 2015;105:1351–7. doi:10.2105/AJPH.2015.302619.

13. Eaton DK, McKnight-Eily LR, Lowry R, Perry GS, Presley-Cantrell L, Croft JB. Prevalence of Insufficient, Borderline, and Optimal Hours of Sleep Among High School Students - United States, 2007. *J Adolesc Heal*. 2010;46:399–401. doi:10.1016/j.jadohealth.2009.10.011.

14. Hansen M, Janssen I, Schiff A, Zee PC, Dubocovich ML. The impact of school daily schedule on adolescent sleep. *Pediatrics*. 2005;115:1555–61. doi:10.1542/peds.2004-1649.

15. Mullington JM, Haack M, Toth M, Serrador JM, Meier-Ewert HK. Cardiovascular, inflammatory, and metabolic consequences of sleep deprivation. *Prog Cardiovasc Dis*. 2009;51:294–302. doi:10.1016/j.pcad.2008.10.003.Cardiovascular.

16. 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;35:1308–17. doi:10.1038/ijo.2011.149.

17. 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. doi:10.1080/15402002.2015.1120198.

18. Short MA, Gradisar M, Lack LC, Wright HR. The impact of sleep on adolescent depressed mood, alertness and academic performance. *J Adolesc*. 2013;36:1025–33. doi:10.1016/j.adolescence.2013.08.007.

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

20. 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;9:517–26. doi:10.1016/j.sleep.2007.07.003.

21. 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 Public Health*. 2015. doi:10.1177/1403494814556647.

22. 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. *East Econ J*. 2015;41:370–86. doi:10.1057/eej.2014.22.

23. 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.

24. Ma J, Pender M, Welch M. Education Pays 2016. *Coll Board Trends High Educ Ser*. 2016:1–44.

25. Walker MP, Stickgold R. Sleep, memory, and plasticity. *Annu Rev Psychol*. 2006;57:139–66. doi:10.1146/annurev.psych.56.091103.070307.

26. Stickgold R. Sleep-dependent memory consolidation. *Nature*. 2005;437:1272–8. doi:10.1038/nature04286.

27. Maquet P. The role of sleep in learning and memory. *Science (80-)*. 2001;294:1048–52. doi:10.1126/science.1062856.

28. Alhola P, Polo-Kantola P. Sleep deprivation: Impact on cognitive performance. *Neuropsychiatr Dis Treat*. 2007;3:553–67.

29. Fuller SC, Bastian KC. The Relationship Between School Start Times and Educational Outcomes. *Curr Sleep Med Reports*. 2020:18–9. doi:10.1007/s40675-020-00198-4.

30. 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 Heal*. 2017;86:363–81. doi:10.1111/josh.12388.School.

31. Hofer M, Kuhnle C, Kilian B, Fries S. Cognitive ability and personality variables as predictors of school grades and test scores in adolescents. *Learn Instr*. 2012;22:368–4752.

32. Fehrmann PG, Keith TZ, Reimers TM. Home influence on school learning: Direct and indirect effects of parental involvement on high school grades. *J Educ Res*. 1987;80:330–671.

33. Keith TZ, Benson MJ. Effects of manipulable influences on high school grades across five ethnic groups. *J Educ Res*. 1992;86:85–671.

34. Lekholm AK, Cliffordson C. Discrepancies between school grades and test scores at individual and school level: effects of gender and family background. *Educ Res Eval*. 2008;14:181–3611.

35. Caprara GV, Barbaranelli C, Steca P, Malone PS. Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *J Sch Psychol*. 2006;44:473–90. doi:10.1016/j.jsp.2006.09.001.

36. 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. *Cochrane Database Syst Rev*. 2017;2017. doi:10.1002/14651858.CD009467.pub2.

37. Wolfson AR, Ziporyn T. Adolescent sleep and later school start times. Sleep, Heal. Soc. From Aetiol. to Public Heal., 2018, p. 215–23. doi:10.1093/oso/9780198778240.003.0024.

38. 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 Public Health*. 2020;17. doi:10.3390/ijerph17072574.

39. Berger AT, Widome R, Troxel WM. Delayed school start times and adolescent health. Elsevier Inc.; 2019. doi:10.1016/B978-0-12-815373-4.00033-2.

40. Gomez Fonseca A, Genzel L. Sleep and academic performance: considering amount, quality and timing. *Curr Opin Behav Sci*. 2020;33:65–71. doi:10.1016/j.cobeha.2019.12.008.

41. Hershner S. Sleep and academic performance: measuring the impact of sleep. *Curr Opin Behav Sci*. 2020;33:51–6. doi:10.1016/j.cobeha.2019.11.009.

42. 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. doi:10.1016/j.smrv.2015.06.002.

43. 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;12:1681–9. doi:10.5664/jcsm.6358.

44. Wahlstrom KL, Owens JA. School start time effects on adolescent learning and academic performance, emotional health and behaviour. *Curr Opin Psychiatry*. 2017;30:485–90. doi:10.1097/YCO.0000000000000368.

45. 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 Heal*. 2016;86:363–81. doi:10.1111/josh.12388.School.

46. 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. doi:10.1016/j.smrv.2015.06.002.

47. Schmidt S. Later school starts linked to better teen grades 2019. https://www.sciencenewsforstudents.org/article/later-school-starts-linked-better-teen-grades (accessed December 21, 2020).

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

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

50. Ackerman X, Phan T, Gee A, Kim A, Imani S, Welkie D, et al. School Start Times 2019. https://ccb.ucsd.edu/_files/bioclock/Infographic PDF, School Start Times 2019, Ackerman, Phan, Gee, Kim, Imani, Welkie, Golden.pdf (accessed December 21, 2020).

51. Biller AM, Molenda C, Zerbini G, Obster F, Förtsch C, Roenneberg T, et al. One year later: longitudinal effects of flexible school start times on teenage sleep, psychological benefits, and academic grades. 2020.

52. Shin J. Sleep More, Study Less ? The Impact of Delayed School Start Time on Sleep and Academic Performance. 2018.

53. Moher D, Liberati A, Tetzlaff J, Altman D. Preferred Reporting Items for Systematic Reviews and MetaAnalyses: The PRISMA Statement. *PLoS Med*. 2009;6:e1000097. doi:10.1371/journal.pmed1000097.

54. 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;64:407–15. doi:10.1016/j.jclinepi.2010.07.017.

55. 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;355:i4919. doi:10.1136/bmj.i4919.

56. 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;35:11–7. doi:10.1097/DBP.0000000000000018.

57. Owens JA, Belon K, Moss P. Impact of delaying school start time on adolescent sleep, mood, and behavior. *Arch Pediatr Adolesc Med*. 2010;164:608–14. doi:10.1001/archpediatrics.2010.96.

58. 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;38:889–94.

59. Hinrichs P. When the bell Tolls: The effects of school starting times on academic achievement. *Educ Financ Policy*. 2011;6:486–507. doi:10.1162/EDFP_a_00045.

60. Edwards F. Early to rise? The effect of daily start times on academic performance. *Econ Educ Rev*. 2012;31:970–83. doi:10.1016/j.econedurev.2012.07.006.

61. 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;5:194–209. doi:10.1080/15402000701263809.

62. Wahlstrom KL, Frederickson J, Wrobel G. School Start Time Study : Technical Report, Volume II Analysis of Student Survey Data. 1997.

63. 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.

64. Wahlstrom K. Changing Times: Findings From the First Longitudinal Study of Later High School Start Times. *NASSP Bull*. 2002;86:3–21. doi:10.1177/019263650208663302.

65. Lenard M, Morrill MS, Westall J. High school start times and student achievement: Looking beyond test scores. *Econ Educ Rev*. 2020;76. doi:10.1016/j.econedurev.2020.101975.

66. Thacher P V, Onyper S V. Longitudinal Outcomes of start time delay on sleep, behavior, and achievement in high school. *Sleep*. 2016;39:271–81. doi:10.5665/sleep.5426.

67. Groen JA, Pabilonia SW. Snooze or lose: High school start times and academic achievement. *Econ Educ Rev*. 2019;72:204–18. doi:10.1016/j.econedurev.2019.05.011.

68. Bastian KC, Fuller SC. Answering the Bell: High School Start Times and Student Academic

Outcomes. *AERA Open*. 2018;4:233285841881242. doi:10.1177/2332858418812424.

69. 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:eaau6200. doi:10.1126/sciadv.aau6200.

70. Jung H. A late bird or a good bird? The effect of 9 o'clock attendance policy on student's achievement. *Asia Pacific Educ Rev*. 2018;19:511–29. doi:10.1007/s12564-018-9558-1.

71. Rhie S, Chae KY. Effects of school time on sleep duration and sleepiness in adolescents. *PLoS One*. 2018;13:e0203318. doi:10.1371/journal.pone.0203318.

72. 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. doi:10.2139/ssrn.3160037.

73. 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. doi:10.3389/fnhum.2017.00588.

74. 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 Heal*. 2017;61:205–11. doi:10.1016/j.jadohealth.2017.02.017.

75. Carskadon MA. Adolescent Sleep Patterns. vol. 37. Cambridge University Press; 2001. doi:10.1017/CBO9780511499999.

76. 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. doi:10.3390/healthcare7010001.

77. Carskadon MA. Factors influencing sleep patterns of adolescents. Adolesc. Sleep Patterns Biol. Soc. Psychol. Influ., New York: Cambridge University Press; 2002.

78. Wittmann M, Dinich J, Merrow M, Roenneberg T. Social Jetlag: Misalignment of Biological and Social Time. *Chronobiol Int*. 2006;23:497–509. doi:10.1080/07420520500545979.

79. 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. doi:10.1016/j.adolescence.2018.06.001.

80. Hafner M, Stepanek M, Troxel WM. The economic implications of later school start times in the United States. *Sleep Heal*. 2017;3:451–7. doi:10.1016/j.sleh.2017.08.007.

81. Jacob BA, Rockoff JE. Organizing schools to improve student achievement: Start times, grade configurations, and teacher assignments. *Hamilt Proj*. 2011:24.

82. 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. doi:10.1016/j.sleep.2018.10.021.

83. 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;111:1–28.

84. Card D, Rothstein J. Racial segregation and the black–white test score gap. *J Public Econ*. 2007;91:2158–84.

85. Fimm B, Brand T, Spijkers W. Time-of-day variation of visuo-spatial attention. *Br J Psychol*. 2016;107:299–321.

86. 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;7:4385. doi:10.1038/s41598-017-04076-y.

87. Rimfeld K, Kovas Y, Dale PS, Plomin R. True grit and genetics: Predicting academic achievement from personality. *J Pers Soc Psychol*. 2016;111:780–9. doi:10.1037/pspp0000089.

88. American Academy of Pediatrics. School Start Times for Adolescents. *Pediatrics*. 2014;134:642–9. doi:10.1542/peds.2014-1697.

89. Zerbini G, Merrow M. Time to learn: How chronotype impacts education. *Psych J*. 2017;6:263–76. doi:10.1002/pchj.178.

90. Neumann K, Kauertz A, Fischer HE. Quality of instruction in science education. In: Fraser BJ, Tobin KG, McRobbie CJ, editors. Second Int. Handb. Sci. Educ., Berlin: Springer; 2012, p. 247–58. doi:10.1007/978-1-4020-9041-7.

91. Juang LP, Silbereisen RK. The relationship between adolescent academic capability beliefs, parenting and school grades. *J Adolesc*. 2002;25:3–18.

92. Pokropek A, Borgonovi F, Jakubowski M. Socio-economic disparities in academic achievement: A comparative analysis of mechanisms and pathways. *Learn Individ Differ*. 2015;42:10–8.

93. Rockoff JE. The impact of individual teachers on student achievement: Evidence from panel data. *Am Econ Rev*. 2004;94:247–52.

94. Jarrin DC, McGrath JJ, Quon EC. Objective and subjective socioeconomic gradients exist for sleep in children and adolescents. *Health Psychol*. 2014;33:301–5. doi:10.1037/a0032924.

95. 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;81:870–83.

96. Dahl RE, Allen NB, Wilbrecht L, Suleiman AB. Importance of investing in adolescence from a developmental science perspective. *Nature*. 2018;554:441–50. doi:10.1038/nature25770.

97. Bowers JM, Moyer A. Effects of school start time on students' sleep duration, daytime sleepiness, and attendance: a meta-analysis. *Sleep Heal*. 2017;3:423–31. doi:10.1016/j.sleh.2017.08.004.

Tab. 1 | Detailed descriptions of included studies. Studies are ordered by their type of study design (longitudinal with control group, without control group and cross-sectional). Study designs of two studies could not clearly be clearly defined. Abbreviations: SST, school start time; OLS, ordinary least square; SD, standard deviation; b, unstandardised beta coefficient; β, standardised beta coefficient; µ, average; CG, control group; IG, intervention group; CSAT, College Scholastic Ability Test; GPA, grade point average; ACT, American College Test; OR, odds ratio; NA, not available.

Author(s) (Year)	Study Design	Sample characteristics	Measure of school performance	Type of analysis	Change in SSTs results in	Key findings
	LONGITUDINAL with control group					
Edwards *et al.* (2012) [60]	22-28 or 15-17 middle schools? 9 schools delayed, 4 advanced, 11 did not change SST pre-change: 07:30 - 08:45 SST post-change: 7:30-8:25 Assessment: 1999-2006 Sampling resolution: once per year	$N_{students}$=20,530 or 10,544 + 6,082? (1999-2000) $N_{students}$=27,686 or 7,191 + 7,675? (2005-2006) $N_{observations}$: up to 102,506 Grade levels: 6th – 8th Age: 11-14.5 Gender ratio: ~ 51% males Ethnicity/race: Caucasian, Black, Hispanic Location: Wake County, North Carolina, USA	End of year standardised test scores in reading and math Provided by: Wake Country administration Scale: 0%-100% (inferred); converted to percentile scores for each student within their grade and current year	Pooled OLS models, quantile regression model predicting scores Covariates: several on student-level and school-level Fixed-effects: student and school	Per1h delay in SST: 1.8-2.9 percent points (0.06-0.07 SD) increase in maths and 1.0-3.4 percent points (0.04-0.05 SD) increase in reading when using within schools variation or both within and between school variation (both p<0.01) Some covariate results for maths and school fixed effect: Black colour: β= -0.50 (p<0.01) Hispanic: β= -0.17 (p<0.01) Female: β= -0.054 (p<0.01) Free lunch status: β= -0.17 (p<0.01) Parent education (years): β =0.08 (p<0.01)	Up to 0.07 SD gains in maths and 0.05 SD increase in reading end-of-year standardised scores even when adjusted for covariates The effect was stronger for lower achieving students
Jung (2018) [70]	Sample from 85 elementary and 63 middle schools; Cohorts from the Gyeonggi Education Panel Study SST pre-change: 8:00-8:20 SST post-change: 9:00 Assessment: 2012-2017; SST delay in 2014; i.e. data cover 3 years prior and 2 years after the change Sampling resolution: once per year	Group 1: longitudinal cohort with change in SST (IG) compared to 220 students (CG) who did not delay: N_{total}=2,562 Grade levels: 4th – 9th Age: 11 – 16 Gender ratio: 50.5% (IG) -57.7% (CG) male Ethnicity/race: NA Group 2: cross-sectional cohorts N_{IG}=4,026 (2015) N_{CG}=2,562 (2012) Grade levels: 7th Age: 14 Gender ratio: ~51 % male Ethnicity/race: NA	Korean, English and math grades at the end of the spring semester Provided by: governmental agency Scale: NA	Difference-in-difference estimation / OLS estimation Covariates: various student-level and school-level characteristics, Fixed effects: year, individual Cross-sectional comparison as robustness check	At first sight, math and English test scores increased (also Korean also but p>0.05) when controlling for personal covariates the effect becomes smaller and not significant for math; when applying individual-fixed effect estimation the result becomes negative (significantly for Korean) Model specification 4 with Year fixed effects and all personal covariates: Korean: β= 0.048 (p>0.05) Math: β= 0.16 (p>0.05) English: β= 0.18 (p<0.01) Robustness check confirms longitudinal results	No effect on Korean, English or math test scores when controlling for personal covariates or unobserved individual heterogeneities Caveat: Sleep duration did not differ between CG and IG!

		Location: Gyeonggi, South Korea				
Kim (2018) [72]	High schools from 2 districts (Gyeonggi and Seoul) SST pre-change in Gyeonggi (IG): varying between before 7:40 and 9:00 SST post-change in Gyeonggi (IG): 9:00 SST in Seoul (CG): varying between before 8:00 and 9:00 Assessment: 2009 to 2016; policy change to 9:00 starts in Sept 1,2014 in Gyeonggi; i.e. data cover 5 years prior and 2 years after the change Sampling resolution: once per year	Group A: Schools in Gyeonggi district (IG) Group B: Schools in Seoul (CG) $N_{observations}$= up to 1,479,131 Grade levels: 9th-12th Age:15-18 Gender ratio: 52% males Ethnicity/race: NA Location: Gyeonggi and Seoul, South Korea	Annual National Assessment of Educational Achievement (Korean, math, English) for 9th and 11th graders College Scholastic Ability Test (CSAT) for 12th graders Provided by: EduDataService System Scale: NA	Difference-in-differences method Covariates: regional time trends Fixed effects: individual and school Several robustness checks	Results 11th graders: Math scores especially in male students increased by 0.06-0.1 SD (p<0.01). Results are robust when adding covariates to the model. Korean and English scores become non-significant when control variables are added to the model. For 12th graders: For CSAT no statistically significant benefit from the 9am policy	Small effects on math, but not on Korean or English standardised scores No effect on CSAT scores (possible time-of-day interference: the CSAT is scheduled before 9 am)
Rhie & Chae (2018) [71]	Several middle and high schools Gyeonggi district delayed SST (intervention groups; IG), three other districts did not (control group; CG) SST IG pre-change: 7:30 – 8:10 SST IG post-change: 9:00 SST CG: 7:30 – 8:00 MS delayed by 30-60min HS delayed by 1-1.5h Assessment: 2 years of data before and after the change (2012–2016); 2014 as the year of change was excluded Sampling resolution: once per year	N_{IG}=42,517 N_{CG}=28,287 Grade levels: 7th – 11th or 12th Age: NA Gender ratio: ~52% male Ethnicity/race: NA Location IG: Gyeonggi district, South Korea Location CG: Daegu/Gyeongbuk/Ulsan district, South Korea	Self-reported GPAs Provided by: participants Scale: Percentage of students having" high and moderate GPAs"	Logistic regression analysis using complex samples Covariates: NA	Percentage of students reporting "high and mid high GPAs": Years 2012,2013,2015,2016: IG= 34.3%, 33.9%, 38.4%*, 37.8%* CG= 39.8%*, 36.7%, 40.6%*, 39.4%* *: different from 2013 on p<0.05	No 9AM policy effect on self-reported GPAs

Shin (2018) [52]	All middle schools in 2 districts (599 schools in Gyeonggi and 383 schools in Seoul) SST pre-change in Gyeonggi (IG): around 8:20 SST post-change in Gyeonggi (IG): 9:00 SST in Seoul (CG): varying between before 8:00 and 9:00 Assessment: 2013-2015 with the policy change to 9:00 starts in Sept 1,2014 in Gyeonggi Sampling resolution: once per semester	$N_{observations}$= up to 33,282 Grade levels: 7th -9th Age: NA Gender ratio: ~50% male (direct contact with author) Ethnicity/race: mostly Asian (private correspondence with author) Location: Gyeonggi and Seoul, South Korea	Semester grades (standardised) for math and reading Provided by: Korean Education & Research Information Service Scale: numeric; 0-100; normalized by population distribution	Difference-in-difference methods Covariates: various individual and school-level variables Fixed effects: year, month	Increase in math (0.03 SD) and reading grades (0.02 SD; (both p<0.001)	Up to 0.03 SD increase in math and 0.02 SD increase in reading semester grades when adjusted for time-trends and other covariates
Lenard *et al.* (2020) [65]	19 high schools, of which 5 schools advanced SST (=intervention group; IG) and 14 did not (=control group; CG) SST IG pre-change: 8:05 SST IG post-change: 7:25 SST CG: 7:25 Assessment: data span 2008-2019 with SST advance in 2012-2013 Sampling resolution: once per year	$N_{students}$~10,000 per each 8 cohorts $N_{observations}$= up to 52,854 (ACT scores) Grade levels: 8th – 12th (inferred) Age: NA Gender ratio: ~ 50% males Ethnicity/race: White, African American, Hispanic, other Location: Wake County, North Carolina, USA	ACT scores in 11th grade (composite and individual scores for English, reading, math and science) Provided by: Wake County administration Scale: 1-36 (=best)	Difference-in-difference estimation approach Comparative interrupted time series Covariates: various individual and school-level variables	No effect of earlier start times on ACT composite or individual subject ACT scores (independent of length of exposure) ACT composite scores: Partial exposure: β= 0.023, p>0.05 Early start all years: β= -0.167, p>0.05 Treated schools all: β= 0.273, p>0.05 Scores were trending in all schools with math scores dropping over subsequent cohort groups, while English, reading and science were rising	No effect on individual or composite ACT scores when start times are advanced
Author(s) (Year)	**Study Design**	**Sample characteristics**	**Measure of school performance**	**Type of analysis**	**Change in SSTs results in**	**Key findings**
	LONGITUDINAL without control group					
Biller *et al.*[51]	1 high school SST pre-change: Mostly 8:00 SST post-change: 8:00 or 8:50 (flexible choice on a daily basis) 4 years (2.5 before and 1.5 years after the change	$N_{students}$= 63-157 $N_{observations}$= up to 16,724 Grade levels: 7th – 12th Age: 14-21 Gender ratio: 30-40% males Ethnicity/race: NA Location: Alsdorf, Aachen region, Germany	Quarterly grades of 12 academic subjects of 3 disciplines (sciences, social sciences, languages) Provided by: school registry Scale: numeric, 0%-100% (=best)	Linear mixed models predicting quarterly grades Covariates (in all models): gender, grade level, academic discipline, academic quarter	No effect of flexible system on grades: β = 0.00 (p>0.05) No absolute sleep effects on grades No effects of change in sleep duration or chronotype on grades except for social jetlag (post β=0.03, p=0.027)	No effect of the flexible system nor sleep duration or chronotype on quarterly objective grades when controlled for covariates

	Sampling resolution: 4 times per year			Predictors (in some models): chronotype (+change from t_0-t_1), social jetlag (+change from t_0-t_1), sleep duration (+change from t_0-t_1), amount of 8:50AM-use	Covariates (from models 3a-d): Male: β= <0.07(p>0.05) Grade level 12: β=0.06 (p<0.001) Quarter 4: β=0.05 (p<0.001) Social Sciences: β=0.17 (p<0.001)	
Boergers *et al.* (2014) [56]	1 independent high school (boarding school) SST pre-change: 08:00 SST post-change: 08:25 Assessment: Nov 2010 (SSTs: 8:00), Mar 2011 (SSTs: 8:25), and May 2011 (SSTs: 8:00) Sampling resolution: once per time point	$N_{students}$=197 Grade levels: 9th – 12th Age: µ = 15.6 Gender ratio: 41% males Ethnicity/race: White, Black, Hispanic Asian, multiracial or other Location: Rhode Island, USA	Self-reported grades Provided by: students Scale: categorical; “mostly Bs or better”	No type of analyses stated Covariates: NA	After the delay in SST, the percentage of self-reported “mostly Bs or better” changed from 93% to 91 %	Unclear as statistics are not reported; authors report no effect
Owens *et al.* (2010) [57]	1 independent high school (boarding and day school) SST pre-change: 8:00 SST post-change: 8:30 (for 2 months, January-March 2009) Assessment: Dec 2008, Mar 2009 Sampling resolution: once per time point	$N_{students}$=201 Grade levels: 9th – 12th Age: µ~16.5 Gender ratio: ~ 43% males Ethnicity/race: NA Location: Rhode Island, USA	Self-reported grades Provided by: participants Scale: categorical; “mostly B’s or better”	χ^2 analysis Covariates: NA	After the delay in SST, the percentage of self-reported “mostly B’s or better” changed from 82.2% to 87.1%; OR= 0.70; 95% CI= 0.41-1.20, X^2=1.71, p = 0.22	No effect on self-reported grades
Thacher & Onyper (2016) [66]	1 public high school SST pre-change: 7:45 SST post-change: 8:30 Assessment: data span 2010-2014; 2 years of data before and after the change in 2012 Sampling resolution: once per year	$N_{students}$ ~ 650 – 800 across 4 years (but t-test for cross-sectional comparisons seems to be less) Grade levels: 9th – 12th Age: µ~16.5 Gender ratio: NA Ethnicity/race: NA Location: Glen Falls, NY, USA	Weighted average GPAs and subject-specific GPAs (English, science, math, social studies, art, music, foreign language, and health studies) Standardised test scores from Regents Exams for cross-sectional comparison Provided by: school	**Longitudinal comparisons:** mixed effect analyses including within-subject effect control Moderator: gender and free/reduced lunch status	**Longitudinal comparison:** no statistically significant evidence for improvement/decline in school performance (overall and by subject) Effect of grade level (higher grade levels better grade) and gender (males worse) No actual statistical results reported	No (systematic) effect on overall GPAs standardised test scores

			Scale: numeric; 100 point scale for GPAs Scale for Regents Exam: NA	**Cross-sectional comparisons:** independent-samples t-tests for grades and standardised test scores	**Cross-sectional by grade level:** Only 11th graders GPAs increased by 2.55 percent points after the change: t_{295}=2.20, p = 0.028 Regents exams: 2 of 20 subject test scores (10th grade Earth Sciences and 11th grade Algebra) were significant better before the change (p<0.007)	GPA test scores of 11th graders improved cross-sectionally with later SSTs No systematic effects for individual subjects of standardised test scores cross-sectionally
Wahl-strom (2002) [64]	7 comprehensive high schools SST pre-change: 7:15 SST post-change: 8:40 Assessment: 6 years; data cover 3 years before and after the change in the 1997-1998 school year Sampling resolution: NA	$N_{students}$=1200? $N_{observations}$= over 1 million Grades levels: 9th – 12th? Age: NA Gender ratio: NA Ethnicity/race: NA Location: Minneapolis, Minnesota, USA	All letter grades (semester and trimester grades) Provided by: school district administration Scale: categorical letter grading	Statistical analysis not reported Covariates: NA	“A small improvement in grades earned overall but not statistically significant” No actual numbers are reported	No effect on letter grades
Author(s) (Year)	**Study Design**	**Sample characteristics**	**Measure of school performance**	**Type of analysis**	**Change in SSTs results in**	**Key findings**
	CROSS-SECTIONAL					
Groen & Pabilonia (2019) [67]	790 high-schools Data from the Child Development Supplement to the Panel Study of Income Dynamics (PSID-CDS) Range of start times from 07:00 to 09:15 (average start time of 7:53) Assessment: data from years 2002/03 and 2007/08 Sampling resolution: once per year	$N_{students}$= 1200? (national representative sample of students) Grade levels: 9th – 12th Age: 13-18 years Gender ratio: 50% males Ethnicity/race: White, Black, Asian, Hispanic Location: USA	Broad-reading test score and applied-problems (math) test score Both age-adjusted and from the Woodcock-Johnson Revised Tests of Basic Achievement Provided by: NA (probably by research assistant) Scale: normalised by survey year	Linear OLS model predicting test scores; Oster model (bounded effects); instrumental-variable estimates Predictors: SST, school day length Covariates: several on student-level and school-level	1h delay in high school start times increases females’ reading scores by 0.16 SD (p<0.1) No sign. effect for females’ math scores and for both males’ applied problems and reading scores From Oster model (bounded effects): 0.16-0.28 increase in reading for females 0.05-0.12 increase in applied-problems for males → probably mediated by an increase of 36 min in sleep duration for every 1 h of later SSTs for females but not for males	Woodcock-Johnsons Test scores increased by up to 0.28 SD in reading for females, no significant effect for males’ scores No significant effect on applied-problems scores for either females or males

Hinrichs 2011[59]	48 districts (73 schools) Minneapolis and some suburbs delayed; St. Paul and other suburbs maintained schedules Minnesota delayed SSTs in 1997/1998: Pre-SST: 7:15 Post-SST: 8:40 St.Paul: SST: 7:30 Assessment: data span 1993-2002 Sampling resolution: once per year	$N_{observations}$=196,617 Number of students not exactly known, but slightly less than the number of observations according to author (private correspondence) Grade levels: 10th-12th Age: NA Gender ratio: 44% males Ethnicity/race: White, Black, Asian, Hispanic, Other Location: Twin Cities metropolitan area, Minneapolis, USA	Individual composite ACT test scores Provided by: ACT test company with permission from schools Scale: numeric; 0-36 (=best)	OLS regression predicting ACT scores Predictors: SST, school day length Covariates: several on student-level and school/district-level	No effect of SSTs on ACT scores (from full specification): 1h later SSTs: 0.02 SD, p>0.05 Covariates: Males: b=0.25 SD, p<0.01 Black: b=-2.47 SD, p<0.01 Low income: b=-0.92 SD, p<0.01	No effect on ACT scores
	Every public high school in Kansas state SSTs: NA Assessment: 2000-2006 (11th grade reading and 10th grade maths between 2001-2006; 11th grade social science and 10th grade science between 2000-2006) Sampling resolution: once per year	$N_{schools}$=1,666 Grades: 10th-11th Age: NA Gender ratio: 40% white females, 9% non-white females, 9% non-white males Ethnicity/race: white and non-white Location: Kansas, USA	School-level test score data on state-wide Kansas Assessments in math, reading, science and social studies Provided by: Kansas Department of Education Scale: 0-100% (inferred, not stated)	OLS regression predicting Kansas assessment test scores Covariates: several on school-level variables	No effect on any of the test scores (maths, reading, science, or social studies) For reading: 1h later SSTs (from full specification): b=0.95, p>0.05	No effect on Kansas Assessment scores in math, reading, science or social studies
	75 schools in 19 districts in Virginia Some schools delayed SSTs (e.g. Arlington Public Schools in 2001/2002 Assessment: data span 2000-2007 Sampling resolution: once per year	$N_{observations}$=171 (number of district-by-year pairs) Grade level/Age/Gender ratio/Ethnicity/race/: NA Location: Virginia suburbs of Washington, DC, USA	End of course exams Provided by: Virginia Department of Education Scale: 0-100%	Analysis: OLS regression predicting end of course exams Covariates: several on school-level	"The results, which are not reported here but are available upon request, are somewhat imprecise, but they do not give evidence for an effect of the timing of the school day on test scores." → requested and confirmed	No effect on end of course exams
Bastian & Fuller (2018)[68]	410 high schools 1,591 schools by year (includes all public school students in North Carolina) 23 schools changed start times, 9 changed start times by ≥30 min 44 districts (incl. 278 schools) had across-school variation in SSTs (average time difference between	$N_{students}$=770,623 Grade level:9th-12th (inferred) Age: 14-18 years (inferred) Gender ratio: NA Ethnicity/race: White, Black, Hispanic, American Indian, Asian, multiracial Location: North Carolina, USA	Average course grades and course grades in 1st period classes in math, English, science and social studies Scale: 4-point scale Conversation of numeric course grades into unweighted grade points	Linear regression models Covariates: several on student-level and school-level; incl. year-fixed effects Specification checks with school fixed effect and with school district fixed	**Course grades:** Per 1h delay: No effect of SSTs on overall course grades β =0.012, p>0.05 SSTs effect on course grades in 1st period: For ≥8:30h starts (compared to <7:30h start): β =0.050, p<0.05 Economically disadvantaged students, minority and low-performing students	No significant relationship between SSTs and average course grades Grades in 1st period class were improved by 0.05 quality points associated with a ≥8:30h start

	earliest and latest was 33min (min 5min to max 2hours); remaining 69 districts (132 schools) had no variation SSTs: ranged from 7:00 to 9:30 Assessment: data span school years 2011-2012 through 2014-2015		Test scores from state wise standardised end-of-course exams (EOC) in algebra, biology, English (normalised) ACT composite scores ACT scale: 0-36 Provided by: North Carolina Department of Public Instruction	effects (=robustness check of main results)	benefited more in course grades overall and in 1st period (per 1h later): β= from 0.049 to 0.074, p>0.05 or 0.01 **EOC scores:** Mixed results for EOC Algebra (improvements, p>0.05.), EOC Biology (reductions, p<0.05), and EOC English (reductions, p>0.05.) **ACT scores:** Per 1h delay: overall SSTs effect on ACT composite: β=0.107, p>0.05; low-performing students: β=0.277, p<0.05	Later SSTs did not systematically predict EOC or ACT but low-performing students did better on the ACT Later SSTs were associated with better course grades (overall and 1st period) for disadvantaged students
Dunster *et al.* (2018)[69]	2 public high-schools (RHS and FHS) SST pre-change: 07:50 SST post-change: 08:45 Assessment: spring 2016 (pre) and spring 2017 (post) Sampling resolution: once per year (second semester)	$N_{students}$=178 (pooled from both schools during each year) 2 independent samples: Sample 2016: n=51 (RHS) + n=41 (FHS) Sample 2017: n=41 (RHS) + n= 41(FHS) Grade level: 10th Age: µ~16 Gender ratio: ~47% male Ethnicity/race: White, Black, Asian, African American, unknown/other Location: Seattle, USA	One 2nd semester grade from a science lab class Provided by: teacher Scale: NA (probably 0%-100%)	As described in the methods: Generalized linear models (binomial) "with year as the dependent variable, testing the hypothesis that years differed based on the basis of the other variables" Other variables: school, sleep offset, grade, mood, chronotype, sleepiness	Years differed after controlling for differences, including grade and sleep variables Median grade 2016: 77.5% Mean grade 2016: 74.6% Median grade 2017: 82% Mean grade 2017: 76.6%	Statistics are not meaningful (not controlled for higher percentage of whites in 2017 sample)
Milić *et al.* (2014) [58]	4 schools (2 grammar schools and 2 vocational schools), all with weekly alternating morning and afternoon schedules 2 schools with early schedule: 07:00 (morning schedule) and 13:00 (afternoon schedule) 2 schools with late schedule: 08:00 (morning schedule) and 14:00 (afternoon schedule)	$N_{students}$= 821 Sample Early schedule: n=452 Sample Late schedule: n=369 Grade levels: NA Age: 15-19 years Gender ratio: across entire sample: 54% males; sample early schedule): 73% males: sample late schedule: 30 % males Ethnicity/race: NA Location: Osijek, Croatia	Final grade in last semester Provided by: NA Scale: numeric; 1-5 (=best)	Mann-Whitney Test Covariates: no	Students attending the early schedule obtained better grades (p<0.001) SST at 07:00: Mean: 3.60 (SD 1.08) = 72.0% SST at 8:00: Mean: 3.28 (SD 1.19) = 65.6%	Final semester grades were better in earlier schools

	Assessment: May and June 2011 Sampling resolution: once					
Wolfson *et al.* (2007) [61]	2 middle schools School E's SST: 7:15 School L's SST: 8:37 Assessment: fall 2003, spring 2004 Sampling resolution: once	$N_{students}$= 205 School E: n=79 School L: n=126 Grade levels: 7th (n=99) – 8th (n=106) Age: NA Gender ratio: 40% males Ethnicity/race: White, African American, Hispanic, other Location: New England, USA	Average fall quarter grade based on mean of 4 subjects (English, science, math and social studies) Provided by: schools Scale: numeric; 0-100% (= best)	MANOVA, Bonferroni correction for group comparisons Variables: school, grade, gender Other covariates: no, schools were similar in SES, size, and ethnic distribution of students (except for a higher percentage of Whites in School L (60% vs. 46%)	Significant School x Grade interaction: F(1,208)=17.06, p<0.001; i.e. there were no school differences for 7th graders but 8th graders Students at School L had higher average grades than students at School E: F(1,104)=10.60, p<0.01; SST at 7:15: μ= 83.16% (SD 7.16) for 7th graders μ= 76.85% (SD 9.45) for 8th graders SST at 8:37: μ= 80.46% (SD 10.11) for 7th graders μ= 83.79% (SD 8.80) for 8th graders No gender differences were found	Increased averaged fall quarter grades for 8th graders in later school No difference for 7th graders
Lewin *et al.* (2017) [74]	26 middle schools with variable SSTs (country wide surveillance data) "Earliest" SSTs: 7:20 – 7:30 "Early" SSTs: 7:40 – 7:55 "Late" SSTs: 8:00 – 8:10 Assessment: surveys in 2008, 2010, and 2012 Sampling resolution: once per time point	$N_{students}$ ~ 32,000 Pooled from all sample years Sample 2008: n=6,936 Sample 2010: n=11,991 Sample 2012: n=10,768 Sample "Earliest" SSTs: n=7,206 Sample "Early" SSTs: n=13,161 Sample "Late" SSTs: n=12,613 Grade levels: 8th Age: 13-14 years Gender ratio: 49.8% males Ethnicity/race: White, non-white Location: NA but most likely USA	Self-reported grades Provided by: participant Scale: 4-point categorical; "Do you mainly get A's, B's, C's, or D's/F's?"	Path analysis with probit regression predicting grades: Predictor: SSTs Mediator: sleep duration (Sobel test) Covariates student-level: survey year, gender, race Covariates school-level: free lunch status Hierarchical structure: students nested within schools	Self-reported grades of students attending the "earliest schools" were significant lower (β=-0.286, p=0.012), no sign. effect for "earlier schools" (β=-0.114, p=0.126) The negative effect of SST on grades was overall mediated by sleep duration: β=0.115, p<0.001 Covariates: Female: β=0.312, p<0.001 Non-white: β=-0.321, p<0.001 Free lunch status: up to β=-0.668, p<0.001	An advance of at least 30 min was associated with worse self-reported grades Longer sleep duration *per se* was also associated with increased grades
Wahl-strom 1997[62]	3 districts;1 delayed SSTs, 2 did not delay High-schools and middle schools High schools (10-12th grades): SST at district A: 8:30 (changed SST) SST at district B: 7:25 SST at district C: 7:15 Middle schools (7-8th grade): SST at district A: 7:35	$N_{students}$= a not further defined sample was drawn from 7,168 students of 17 districts Grade levels: 10th – 12th and 7th – 8th Age: NA Gender ratio: NA Ethnicity/race: NA Location: Minnesota, USA	Self-reported grades Provided by: participants Scale: NA	Statistical analysis: NA Covariates: NA	Mean self-reported grades in district A were highest (p<0.05) compared to district B and C for 10-12th graders: District A: 7.08 District B: 6.50 District C: 6.37 For 7-8th graders: District A: 6.66 District B: 6.91 District C: 6.60	Students in a district with later start time reported getting higher grades than in two districts with earlier SSTs (high schools) For middle schools, students who started later were either

	SST at district B: 8:00 SST at district C: 8:00 Assessment: NA Sampling resolution: NA					better or similar to students from a school which started earlier
Author(s) (Year)	Study Design	Sample characteristics	Measure of school performance	Type of analysis	Change in SSTs results in	Key findings
	STUDY DESIGN NOT CLEAR					
Kelley *et al.* (2017) [73]	English state-funded high school Year 0: SST pre-change (A): 08:50 Year 1-2: SST post-change (B): 10:00 Year 3: SST back-change (A): 08:50 Assessment: 4 years Sampling resolution: once per year	Year 0: $n_{students}$=169 Year 1: $n_{tudents}$=166 Year 2: $n_{tudents}$=164 Year 3: $n_{tudents}$=179 Grade levels: NA Age: 14-16 Gender ratio: NA Ethnicity/race: NA Location: urban-area of 0.7 million in a region where achievement was lower than national average, England	Standard National Examination (GCSE) Provided by: UK Office of National Statistics Scale: G-A* (=best)	T-test; Cohen's d and h for effect size Value-added analysis (predictions) Percentage of students achieving "good academic progress" (=achieving 5 or more GCSE grades of C or better in English, math and at least 3 other subjects) Covariates: NA	Change in value-added as % of national (compared to national average): Year 1 vs 0: +15%, p<0.0005 Year 2 vs 0: +20%, p<0.0005, Year 3 vs 2: -7%, p<0.0005 Percentage of students making good academic progress compared to national average: Year 0: -40%, p<0.005 Year 1: -9%, p=0.182 Year 2: -11%, p=0.081 Year 3: -15%, p=0.014	Delay is associated with higher % of students making good academic progress and higher value added number compared to national average
Wahl-strom (2014) [63]	8 public high schools in 5 school districts in 3 states changed SSTs and participated in a survey on sleep habits Grades were retrieved from 6 schools in 3 districts. SST pre-change: 7:35-7:50 SST post-change: 8:00-8:55 Assessment: 2010-2011 (Minnesota), 2011-2012 (Colorado); 2011-2012 (pre-change in Wyoming) vs 2012-2013 (post change in Wyoming) Sampling resolution: once per time point	$N_{students}$= 9,089 (sleep habits survey) $N_{students}$: NA (grade analyses) Grade levels: 9th-12th Age: 13-19 Gender ratio: 50.6% Ethnicity/race: White, Black/African American, Hispanic/Latino, Asian/Asian American, Other Location: Minnesota/Colorado/Wyoming, USA	Grades in English, maths, social studies, science in 1st and 3rd period-classes or all GPAs Standardised test scores (state-wide achievement tests or PLAN) Provided by: GPAs from districts; categorical grades by students Scale: categorical; "mostly A's=9" to "mostly F's"=1	t-tests, correlations Covariates: NA	Longitudinal standardised test scores: mainly non-significant results and some mixed results for both composite scores (ACT or PLAN) and individual subjects	Mixed and often non-significant effects on GPAs and standardised test scores

Tab. 2 | Risk of bias assessment. Included studies are ordered by study design 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 performance, a right arrow indicates mixed findings. NA, not applicable.

	Longitudinal studies with control group						Longitudinal studies without control group					Cross-sectional studies								NA	
	Edwards 2012	Jung 2018	Kim 2018	Rhie 2018	Lenard 2020	Shin 2018	Biller 2020	Boergers 2014	Owens 2010	Thacher 2016	Wahlstrom 2002	Groen 2019	Hinrichs 2011	Bastain 2018	Dunster 2018	Milić 2014	Wolfson 2007	Wahlstrom 1997	Lewin 2017	Wahlstrom 2014	Kelley 2017
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[1]	🟩 +	🟩 +	🟩 +	🟥 –	🟩 +	🟩 +	🟩 +	🟥 –	🟥 –	🟩 +	🟩 +	🟩 +	🟩 +	🟩 +	🟩 +	🟧 ?	🟩 +	🟥 –	🟥 –	🟩 +	🟩 +
Blinding of participants/personnel (performance bias)[2]	🟩 +	🟩 +	🟩 +	🟥 –	🟩 +	🟩 +	🟥 –	🟥 –	🟥 –	🟥 –	🟧 ?	🟩 +	🟩 +	🟩 +	🟥 –	🟥 –	🟥 –	🟥 –	🟥 –	🟥 –	🟩 +
(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.	N.A.	N.A.	N.A.	N.A.	N.A.	🟧
Total score[3]	5.5/ 8	6/ 8	5.5/ 8	3/ 8	6/ 8	5.5/ 8	3/ 6	0/ 6	1/ 6	2.5/ 6	1.5/ 6	5/ 7	5/ 7	5/ 7	1.5/ 7	2/ 7	3/ 7	0.5/ 7	3/ 7	NA	NA
At least 75% good evidence score		✓			✓							✓	✓	✓							
At least 60% good evidence score	✓	✓	✓		✓	✓						✓	✓	✓							
Results	⇧	⇨	⇧⇨		⇨	⇧						⇧⇨	⇨	⇧⇨							

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

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

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

APPENDIX - Project 3

Supplementary information for

"School start times and academic performance - a systematic review."

Authors:

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

Corresponding authors:

anna.biller@med.uni-muenchen.de
giulia.zerbini@med.uni-augsburg.de

Tab. S1 | Protocol detailing reasons for risk of bias assessment decisions of Tab. 2.

LONGITUDINAL STUDIES WITH CONTROL GROUP			
Study	**Bias**	**Decision**	**Reasons**
Edwards 2012[60]	Reporting bias on author level	Green	Sample characteristics are well described; actual sample size is unclear (the number of observations are always reported in the tables). Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Orange	State-wide data are used to construct percentile scores for each student within their grade and current year; no direct comparison with the national average
Jung 2018[70]	Reporting bias on author level	Green	Sample characteristics are well described. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Green	Yes
Kim 2018[72]	Reporting bias on author level	Orange	Sample characteristics are not well described. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported.
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs. The EDSS provided a 70% randomly extracted sample from the population (includes 3 exam scores, gender and school ID)
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Green	Yes
Rhie 2018[71]	Reporting bias on author level	Orange	Sample characteristics are well described but some information is contradictory (cfr Table 1 and Table S1 grade 7th to 11th or 12th). Statistical analyses are not fully reported. P values are reported but statistical tests are not reported. Overall there is no effect of the 9AM policy, but the third sentence of the discussion says: “Self-reported school performance of the intervention group was more improved than the control”. This analysis is not reported and it is contradictory with the rest of the results
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Data retrieved from surveys
	(Dis)similarity of baseline characteristics	Green	Considered (in terms of sleep onset, sleep offset and sleep duration; not for gender)
	Appropriate statistical models which control for confounders	Orange	No. The authors used a logistic regression with complex samples comparing each year to a baseline year, and for intervention and control group but no covariates were included
	Control group present	Green	Yes

Lenard 2020[65]	Reporting bias on author level	Green	Sample characteristics are well described. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported
	Responder bias on student level	Green	Data objectively reported (ACT scores)
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Green	Yes
Shin 2018[52]	Reporting bias on author level	Orange	Sample characteristics are not well described. Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported.
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
	Control group present	Green	Yes
LONGITUDINAL STUDIES WITHOUT CONTROL GROUP			
Study	**Bias**	**Decision**	**Reasons**
Biller 2020[51]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Green	Data objectively reported
	Blinding	Red	Even though grades were obtained objectively from the school registry, students were aware they were participating in the experiment (they had to fill in questionnaires)
	Appropriate statistical models which control for confounders	Green	Yes
Boergers 2014[56]	Reporting bias on author level	Red	Statistical analyses regarding grades are not reported
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	Appropriate statistical models which control for confounders	Red	Unable to judge due to missing information
Owens 2010[57]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses). Ethnicity is missing
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	Appropriate statistical models which control for confounders	Red	Simple Chi-Square Test not controlling for confounders
Thacher 2016[66]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Green	Data objectively reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	Appropriate statistical models which control for confounders	Orange	The longitudinal analysis is appropriate (mixed linear model with some moderators/covariates); cross-sectional analyses are only simple t-tests; it is not clear why the authors do not run a mixed within-between model and combine longitudinal and cross-sectional analyses. Nevertheless, several analyses are reported which supports the notion that the data were extensively explored to reach the conclusions of the paper

Wahl-strom 2002[64]	Reporting bias on author level	Red	Sample size (n students) for the grade analyses not reported, statistical analyses not reported, demographic characteristics of the sample not fully reported; author(s) were contacted but did not respond
	Responder bias on student level	Green	Data objectively reported
	Blinding	Orange	Students were aware they were participating in the experiment (they had to fill in questionnaires) but grades were collected also in students who did not fill in questionnaires
	Appropriate statistical models which control for confounders	Red	Incomplete information to judge
CROSS-SECTIONAL STUDIES			
Study	**Bias**	**Decision**	**Reasons**
Groen 2019[67]	Reporting bias on author level	Green	Statistical analyses regarding results in Table 1 are not reported; not always specified that the statistically significant results were at the 0.1 level
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
Hinrichs 2011[59]	Reporting bias on author level	Green	Originally orange but author provided all details after contacting
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
Bastian 2018[68]	Reporting bias on author level	Green	Statistical analyses and models are well described; the model equations, p-values and beta coefficients are reported.
	Responder bias on student level	Green	Data objectively reported
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Green	Many variables considered and analysed in association with SSTs
	Appropriate statistical models which control for confounders	Green	Yes
Dunster 2018[69]	Reporting bias on author level	Red	Statistical analyses are contradictory reported in the methods and in the results. Only p-values are reported and not the statistical tests performed. The conclusions about academic performance are not supported by the statistical analyses performed
	Responder bias on student level	Green	Data objectively reported
	Blinding	Red	Students were aware they were participating in the experiment, and teachers who provided the grades as well (the authors also recognize a potential teacher-level bias)
	(Dis)similarity of baseline characteristics	Orange	Variables such as ethnicity and other possible confounders were assessed but not considered in the analyses; higher percentage of whites were found in 2017 sample
	Appropriate statistical models which control for confounders	Red	Better grades were predictive of school year but this was after accounting for other predictive variables. Authors thus tested year as dependent variable; the conclusions are wrong based on the test
Milic 2014[58]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Orange	Grades are possibly subjectively reported from questionnaires but this is unclear; author(s) were contacted but did not respond
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)

	(Dis)similarity of baseline characteristics	Orange	The different gender composition of the early schedule and late schedule samples is reported and discussed but not controlled for in the analyses
	Appropriate statistical models which control for confounders	Red	Simple Mann-Whitney Test without controlling for confounders
Wolfson 2007[61]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Green	Data objectively reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	(Dis)similarity of baseline characteristics	Orange	Schools are very similar, except for a higher percentage of white students in the school with later SSTs
	Appropriate statistical models which control for confounders	Orange	MANOVA with Bonferroni correction for multiple comparisons. One problem is that ethnicity was not controlled for although there were more white students in the school with later SSTs.
Wahl-strom 1997[62]	Reporting bias on author level	Red	All information necessary to critically read the results are missing (sample information, statistical analyses etc.)
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	(Dis)similarity of baseline characteristics	Orange	The author reports that the schools are similar but data are not reported. One caveat is that students from district B and C also spent more time on homework.
	Appropriate statistical models which control for confounders	Red	Insufficient information to be judged
Lewin 2017[74]	Reporting bias on author level	Green	All information necessary to critically read the results are reported (sample information, school schedule, statistical analyses)
	Responder bias on student level	Red	Grades are self-reported
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires)
	(Dis)similarity of baseline characteristics	Green	Reported and controlled for
	Appropriate statistical models which control for confounders	Green	Yes
STUDY DESIGN NOT CLEAR			
Wahl-strom 2014[63]	Reporting bias on author level	Red	Sample size (n students) for the grade analyses are not reported; statistical analyses are not reported (t-test, p-values); the average GPAs before and after the change are only reported in the appendix but without specifying the SD and range and therefore it is difficult to judge the effect size
	Responder bias on student level	Green	Grades are probably objectively reported (GPAs and test scores); author(s) were contacted but did not respond
	Blinding	Red	Students were aware they were participating in the experiment (they had to fill in questionnaires). The authors do not report whether grades were collected also in other students that did not fill in questionnaires.
	Appropriate statistical models which control for confounders	Red	One can only infer that independent t-tests were used as the authors do not clearly report it. T-tests without controlling for co-variates would be problematic
Kelley 2017[73]	Reporting bias on author level	Orange	Sample characteristics are not well described. Statistical analyses and models are described, p-values are reported but t-tests and also Cohens-coefficients are not always reported
	Responder bias on student level	Green	Grades are objective
	Blinding	Green	Data retrieved from archive
	(Dis)similarity of baseline characteristics	Red	Not reported

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

5

General discussion

5.1. Research summary

Teenager worldwide suffer from inadequate sleep, partly because their biological does not tick in synchrony with the social clock around them. Physiology and social influence drastically change during puberty and this also affects students' chronobiology and sleep. Mostly cross-sectional studies[*e.g.* 128–130] but also longitudinal studies which tracked individuals over several years[133,134] have provided evidence that teenagers phase delay across adolescents (see also Fig. 8). The consequence: late sleep that is drastically cut short by early school starts (SSTs). Germany is no exception to this common problem. In most schools in this country, students have to be ready at 8:00h sharp or earlier. But the reality is that students accumulate substantial amounts of sleep loss across the week, which they try to reduce by oversleeping on the weekends. Fig. 7 displays the amount of sleep German teenagers get across the school week - clearly less than the recommended, healthy amounts of 8-10 hours, which would be crucial for optimal development and functioning[192,196,197]. So how could we address this public health concern? We can only marginally influence our biology but we change social schedules. The aim of this thesis was to shed light on longitudinal effects of *flexible* school start times on teenagers' sleep, their subjective psychological functioning and wellbeing, and their academic grades. In cooperation with a high-school in Alsdorf, Germany, we investigated how these factors are influenced when students could choose daily whether to start their school day at 8:00h or 8:50h. This school had won the German School Award in 2013 due to its innovative teaching concept before it introduced this new start time scheduling in 2016. The Alsdorf Gymnasium is part of the Dalton School Network that encourage self-study and self-responsibility among their senior students to prepare them for their adult life and university-type studying. This school thus serves as a special role model, not only for other Dalton schools around the world but also for schools within Germany. Educational reforms in Germany have got a bad reputation for being particularly slow[198] and it seems that changing school start times is no exception[199]. With Germany losing ground in international comparison, such as in PISA (Programme for International Student Assessment[200]) major educational changes have repeatedly been requested for the upcoming years and decades[201].

My scientific contributions to this possible change are the results from two research studies and one systematic review that are presented in this thesis. In our first study, we showed that students at the Gymnasium Alsdorf slept longer during nights before school days with later starts (**Manuscript 1**) and that this still holds true after 1 year in such a flexible system (**Manuscript 2**). Our main outcome measure was sleep duration, which we hypothesised to increase in the new system to counteract the acute and chronic sleep restrictions observed in teenagers worldwide[129] . The sleep gain (difference in sleep duration during nights before an 8:00h school start vs an 8:50h start) of about 60 minutes straight after the introduction of the flexi system in 2016 was maintained after one year. Interestingly, girls were slightly more successful in keeping this gain while boys slightly delayed their sleep onset in 2017 (for more details see Discussion of **Manuscript 2**). Still, in our studies all students - independent of chronotype, age or gender - significantly (biologically and statistically) benefitted from this new start system both in the short and in the long-term, hinting towards the severe sleep restriction all students accumulate over the course of an average school week. Indeed, as reported in **Manuscript 2**, only 3%-15% of students reached at least 8 hours of sleep on conventional school start days, while 70%-85% manage to achieve this on weekends. On the positive side, when school started later, 45%-59% of students achieved the recommended sleep minimum. These findings are in line with several other longitudinal studies, which also showed maintained sleep gains at several follow-up times (0.25 months to 2 years)[193,202,203] after a delay in SSTs, even though others reported that students merely shifted and the gain was consequently lost[204,205]. Importantly though, as also stated in the Discussions of **Manuscript 1 and 2**, the evidence is relatively mixed and the majority of studies were conducted cross-sectionally thus often preventing any causal interpretation (see also chapter 5.4).

Importantly, the objective success with regards to increased sleep duration in our studies also resonated in subjective ratings of psychological functioning and wellbeing. Compared to the old system, students rated their sleep quality and duration, quality of study at home, and concentration higher, while their alarm driven waking and their tiredness significantly reduced in the flexible system (**Manuscript 1**). In **Manuscript 2** we compared ratings for early (8:00h) vs later school starts (≥8:50h) and again found that students liked and benefited from the late starts: ratings were higher for concentration and wellbeing, and students also believed that their quality of studying at home increased. Importantly, students were also more motivated to go to school and their general attitude towards school was better. Other studies have also found that students reported decreased depression symptoms, while truancy rates reduced and attendance rates and alertness levels increased (as reviewed in [193–195,206].

Finally, we also investigated academic performance in our students. Since grades open doors for higher education in Germany and are thus a determinant for future success (in contrast to the US[j]), we decided to investigate if the flexible system had any positive effect on students' grades even though we were aware of all the possible confounding variables that influence them. With over 16,000 observations our data were numerous enough to apply linear mixed model analyses to test the influence of several other factors on grades, such as chronotype, sleep duration, social jetlag or how often the students chose to start later ("9AM-use") while simultaneously allowing for inter-individual differences. We did not find any systematic evidence in any of our models that the flexible system allowed students to perform better academically, although we observed that a simple t-test (which only tests for differences of aggregated grades before and after the change and does not control for covariates) would have suggested so (see **Manuscript 2**). This is a major concern that we further discussed in the systematic review (**Manuscript 3**) and warrants attention as interpretation of study results is heavily influenced by the study design and the type of statistical analyses one carries out (see also chapter 5.4). Indeed, as we also found in the systematic review, the literature is full of very mixed

[j] In the U.S., students' performance in the ACT (American College Test) rather than GPA determines access to higher education.

findings as to whether delayed start times really improve grades or test scores (**Manuscript 3**) and we were unable to find any tendency in relation to study design and outcome, sample size, or dose (*i.e.* amount of delay or duration of exposure). However, as Jung systematically demonstrated, adding different families of covariates (on the student or school-level) greatly decreases the size of a positive effect, eventually resulting in a relatively small, although still statistically significant, influence[207]. While some authors acknowledge that such a small effect might not really be biologically meaningful for the individual student, others failed to take this into consideration.

It might be helpful to compare effects sizes of positive reports of later starts with those of covariates or other predictors to put them into perspective (see also **Manuscript 3**). Edwards and colleagues for example reported up to 0.07 SD increases in maths and 0.05 SD reading respectively, which is only roughly 14% of the black-white performance gap, 85% of the gain of one additional year of parental education, but 40% of the gap between students with low or high socio-economic status in their study[208]. Lewin *et al*. also calculated the indirect effect of sleep duration on grades and found that 1h more sleep improved grades by 0.12 SD, while we did not observe such an effect in our study (**Manuscript 2**). We found instead that subject type had the highest impact on grades (0.17 SD), which was also supported by Goldin *et al*. (2020)[209]. It should thus be considered that influences of start times, even if significant, are often smaller compared to other effects.

In summary, the presented findings from my three manuscripts contribute to the current knowledge in several ways. Firstly, longitudinal assessments of sleep changes after a delay or change of SSTs were long warranted. Even though investigations into this topic already started in the late 1990s[122] and several other "middle-term" studies had investigated this before (as reviewed in [193]), only 3 tracked students for at least one year [202,210,211]. Long-term studies are important in this context as teenagers still undergo major physical and cognitive changes during adolescence[*e.g.* 112] and studies investigating the effects of altered SSTs on sleep need to take this into account: even if a delay in SSTs is helpful in the short-term, administrative effort to change bell times is considerable so it needs to pay off long-term for students, parents and staff. Secondly, to the best of my knowledge, this is the first study to investigate *flexible* school start times worldwide and the first to look for effects of changed SSTs in Germany. A flexible system offers several advantages and possible disadvantages for students, which will be further discussed in chapter 5.2. Additionally, most of the evidence concerning this policy change comes from the U.S. or South Korea, however the educational systems in these countries are highly different compared to Germany. While biological changes might be relatively universal across ethnic or racial background, socio-cultural differences (which includes the *social clock*) could prevent findings to universally hold true, thus a German perspective is important to inform policy making. Thirdly, we conducted a systematic review including a rigorous risk of bias assessment and showed that there is no clear evidence that supports that delayed SSTs lead to better academic performance even though the scientific community somewhat established such a consensus (**Manuscript 3**). While others reviewed the evidence on altered SSTs and sleep and/or performance in general[193–195,206,212–217], only 2 of these were systematic reviews[195,206], and no review existed that specifically looked at academic performance in detail. We concluded that there seems to be no harmful effect on academic performance when schools start later – but very likely it also does not bump performance substantially, at least at the dose that students received in the presented studies. Importantly, sleep is often improved when school delays, which is accompanied by increased well-being, better psychology functioning and a preference of students to start later – all prerequisites for healthy learning.

5.2. The flexible system: innovation or the devil in disguise?

One of the most interesting questions to discuss further is the flexible system itself. Usually, schools either advance or delay their start times by a fixed amount of time, mostly 30 minutes to 1 hour (see **Manuscript 3**), which also permanently delays or advances the entire school schedule. This was different in the present high school as the flexibility was provided on a daily basis and neatly nested into the students' schedules. In this way the last scheduled class still finished at 16:15h latest, just as in the conventional system before the change (see **Manuscript 1** and **2** for the exact scheduling).

One big caveat in the flexible system was that students did not choose to attend school at 8:50h as often as we had expected (see also **Manuscript 1 and 2**). This was mostly due to timetable and transportation conflicts but students also indicated that they used the first hours for studying. This low uptake of later starts on about 22-39% of their school days very likely explains why we did not obverse an overall increase in sleep duration across the school week in the flexible system. Thus, the chronic sleep loss students suffered from in the old system was not ameliorated with the flexible starts, only acutely when they opted for a later start the next day. Given the choice, students thus might not necessarily take the option which would be good for their biology and health, they also consider social obligations and other reasons – an argument which would point towards a permanent delay in SSTs.

On the other hand, depending on the exact scheduling, location and time of year, students might actually phase-delay in a permanently delayed system. Morning light advances the circadian phase, while evening light (especially blue-light) delays it (see also Fig. 3 and chapter 1.1.3). Thus, permanently delaying the school starts day could exaggerate this shifting: missing important phase-advancing morning light negatively adds to the low-intensity, artificial light levels throughout the day (an inappropriate zeitgeber for the circadian clock[42], see also chapter 1.1.3) and the extensive use of blue-light emitting devices (*e.g* from phones, tablets, computers) at night before bedtime. The difficulty lies in the specific location within the given time zone. In our case, students woke up before sunrise in February when classes started at 8:00h but could potentially wake after sunrise for classes at 08:50h exposing them to *more* light in the morning when they *woke up later* and thus potentially preventing a delay (Fig. 10a). This is somewhat in line with a mathematical modelling paper, which predicted that delays in social schedules only prevent phase-shifts if the social day had previously started before sunrise[218]. Therefore, a flexible system would allow students to decide when to start school in accordance with the photoperiod and its changes throughout the year while simultaneously not getting home later. This is also psychologically relevant, as a simple option to choose[k] provides students with a sense of control and self-efficacy[112], thus potentially easing ruminations about sleep, which in turn might improve sleep latency and quality. The flexible system is an innovative and promising alternative to a permanent delay in SSTs but given its novelty, more studies are warranted to follow up our hypothesised advantages and problems.

[k] Note that too many options on the other hand might be counterproductive (as reviewed in[234]).

5.3. Owl-friendly school systems: some suggestions

Several other or additional possibilities to match students' internal time more closely with the external (sun) time also exist or are conceivable. For example, start times could be (permanently or flexibly) delayed only during the winter months to align seasonal changes in photoperiod and consequently sleep with school times. If we consider sunrise times in Alsdorf (West Germany, study location of **Manuscripts 1+2**), we see that the sun does not rise before 8:00h between mid-November and mid-February (standard time; Fig. 10a). In these months, allowing flexible starting hours or a fixed start at 09:00h would give students the chance to receive an additional 30-60 min of light prior to their first class. Implementing constant daylight-saving time (DST) as discussed in the EU, on the other hand, would expand the "dark period" before 8:00h to last 2 months longer, ranging an entire half year from mid-October to mid-March (Fig. 10b). This is quite alarming: even though implementing DST throughout the year is a current public debate following an EU survey, from a chronobiological viewpoint it is clearly disadvantageous since it further aggravates circadian misalignment, which can lead to mental and physical health problems (see chapter 1.3.3).

Instead, another healthy possibility is morning exercise, which has been shown to advance circadian rhythms, independent of the phase-advancing and acute alerting effects of morning light[219]. Physical education could thus be moved forward to the first morning class in schools. This could be complemented by blocking or reducing blue-light exposure in the evening or at night, for example by means of blue light filters, which many electronical devices now feature, by smart home lighting, which gradually filters out blue light over the course of the evening, or by wearing blue-light blocking glasses[220]. While the latter has been shown to successfully phase advance students, it is questionable how teenagers during puberty would like to adhere to such drastic measures. This holds also true for short term sleep restrictions on the weekend: Misiunaite and colleagues investigated whether cutting students' sleep short on a Saturday morning by several hours would advance students dim light melatonin onset and thus facilitate earlier bedtimes in the subsequent nights[221]. While this strategy phase-advanced DLMO the consecutive day and thus might be sensible once in a while to give the system a sort of phase advancing "kick-start", it is likely that such effects do not hold long-term (the study has not followed this up more than one day yet).

In contrast, one relatively simple change would involve teaching main subjects (*e.g.* maths, natural sciences and languages) during core hours when both early and late chronotypes are cognitively awake, for example between 9:30h and 12:30h, or in the afternoon (see also [209]) as it is already implemented in some schools in the Netherlands[151]. Lastly, time- and subject-specific examination could also equalize performance differences between chronotypes as shown in several previous studies[151,209,222]. An optimal solution to cater for circadian changes during adolescence and inter-individual chronotype differences would probably contain adjusting several screws - all of which involve time scheduling amendments on the school side and some behavioural changes on the student side.

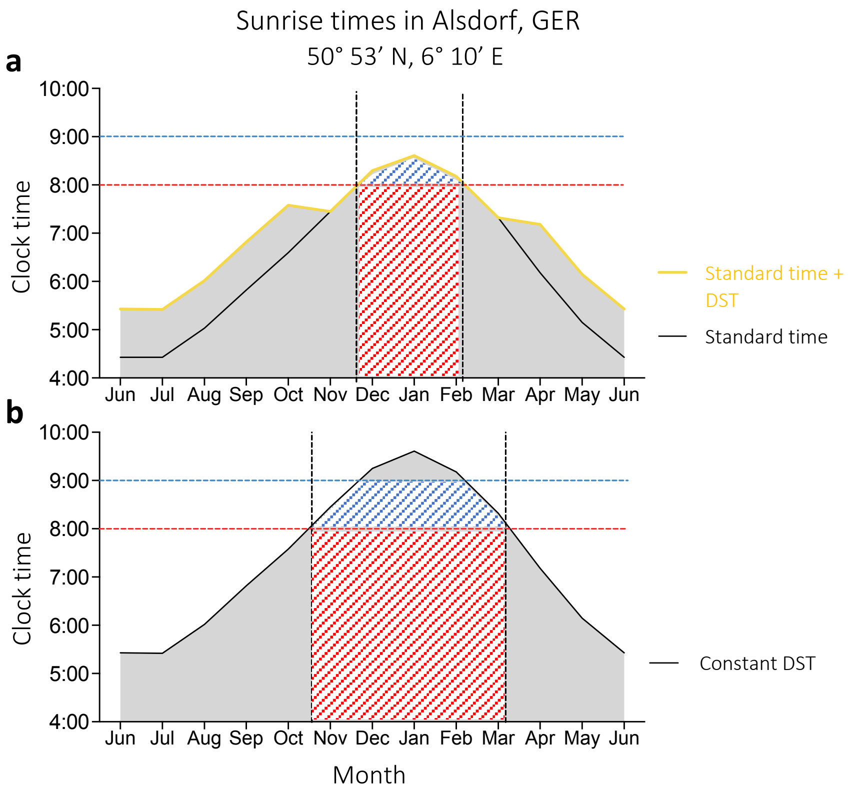

Fig. 10 | Sunrise times in our study location in Germany. Depicted are sunrise times across the year according to **a,** standard clock time with daylight saving time between end of March and end of October (DST, yellow), constant standard time (black) and **b,** constant DST throughout the year. Red-striped areas under the curve describe the dark period prior to sunrise at 8:00h; blue-striped areas describe the additional dark period when sunrise is later than 8:00h. Sunrise times taken from *www.timeanddate.de*.

5.4. Limitations of the current evidence

Specific limitations of the presented studies are detailed in the respective sections of each manuscript. Here, I want to review the general or overarching obstacles encountered with field studies. While the ecological validity is greatly increased in field- compared to lab-based studies, there are numerous factors that researchers cannot control and are sometimes not even aware of. Careful study designs and appropriate statistics are thus needed to draw sound conclusions. In general, double-blind randomised controlled trials (RCTs) are the gold standard for establishing causal effects but RCTs are inherently unfeasible to determine the effects of altered school start times on sleep, wellbeing and performance, even though a research group at Oxford made a bold attempt with regards to randomisation[223] (see also **Discussion of Manuscript 3**). Interestingly, there are several existing settings that lend themselves beautifully to the study of these effects as students are randomly allocated to schools (South Korea[207,224,225]), or to either alternating (Croatia[226]), or fixed school shifts taking place in the morning, afternoon or evening (*e.g.* in Uruguay[138]). Thus, time-of-day effects of examination, chronotype differences, and effects of school timing on sleep can be studied.

Especially for effects on academic performance, it is crucial that confounding variables are taken care of or accounted for (see **Manuscript 3**). Kim, Jung, and Shin provide some good examples[207,224,225] in this regard: in a series of papers the authors independently examined a sudden introduction of a 9:00h start ("9 o'clock policy") in the Gyeonggi province in South Korea. Students there have similar racial backgrounds, are allocated randomly to schools within a district, and curriculums and exams are standardised and accessible through a national archive. Thus, many variables were already controlled for in these studies and data from many students was accessible. The authors were also able to use data from provinces in which the schools did not change to a later start times, thus they made use of a natural comparison group.

Despite all this, these studies also showed small, mixed and null-effects on grades and scores. It should thus be considered that a change in SSTs does not necessarily have a direct effect on academic performance, but rather an indirect effect mediated by other variables such as sleep (duration or daytime sleepiness) as shown in previous path-analyses[227]. While the evidence is strong that more sleep leads to better cognitive performance and psychological functioning, and that sleep restriction severely hampers both[92,183,184,187,228], the sleep improvements following a short-term or small delay in SSTs might not be influential enough to translate into better performance. Additionally, there might be a good time for learning and a good time for testing, both of which are currently not sufficiently considered in school schedules and research studies.

5.5. Future research avenues

Based on the limitations of the presented studies and the ones identified in **Manuscript 3** and chapter 5.3, future longitudinal (cohort/panel) studies which include comparisons groups are still warranted. Most importantly, appropriate and advanced statistics are needed to investigate the influences of changes in SSTs on academic performance to avoid false effects. Studies that do not consider covariates (at least partially) are very likely to report inappropriate results. We still need to answer many basic questions, *e.g.* should school start flexibly or with a fixed delay? When should school start exactly and how long is a good school day? When is the ideal time for learning and when for testing - for all students, not just early chronotypes or high-achievers?

Furthermore, it would be fantastic to test some of the ideas presented in chapter 5.3 in field studies. While some of these measures to counteract sleep deprivation in students are already in place in some schools (*e.g*. core learning hours), not much research has been conducted to verify if these are indeed helpful for students. As Dahl points out, "the importance of student-driven learning (autonomy), collaborative learning (social engagement in learning) and school and classroom climates that honour adolescent sensitivities to status, respect and purpose, are likely to have powerful positive effects on learning"[112]. A flexible start system that is coupled to a student-driven learning (as in our studied school, the "Dalton concept") might thus be a positive way forward in education.

5.6. Conclusion

Since the first studies on teenage sleep and delaying SSTs in the 1990s have been conducted mainly by Mary Carskadon and colleagues[e.g. 122,229] we have come a long way, at least in theory. Fig. 11 summarises the current evidence in an infographic showing that the biological and social challenges that teenagers face during adolescence can lead to acute and chronic sleep restriction. There are possibly several ways in how we could tackle this problem, one of which is to (permanently or flexibly) delay school start times. Even though the jury is still out with regards to better grades or scores, this has the great potential of improving sleep and wellbeing and possibly to decrease accidents, tardiness, and absences at potentially relatively low costs[230]. However, schools need to take up suggestions from science to address the severe sleep restriction of their students. The current Covid-19 pandemic has forced schools all over the world to implement home schooling and distant learning at a fast pace which might have lowered the objections of many policy makers and school administrations to try out new scheduling and teaching formats. This situation might thus offer a unique possibility to implement innovations and educational changes much faster than anybody would have imagined a year ago.

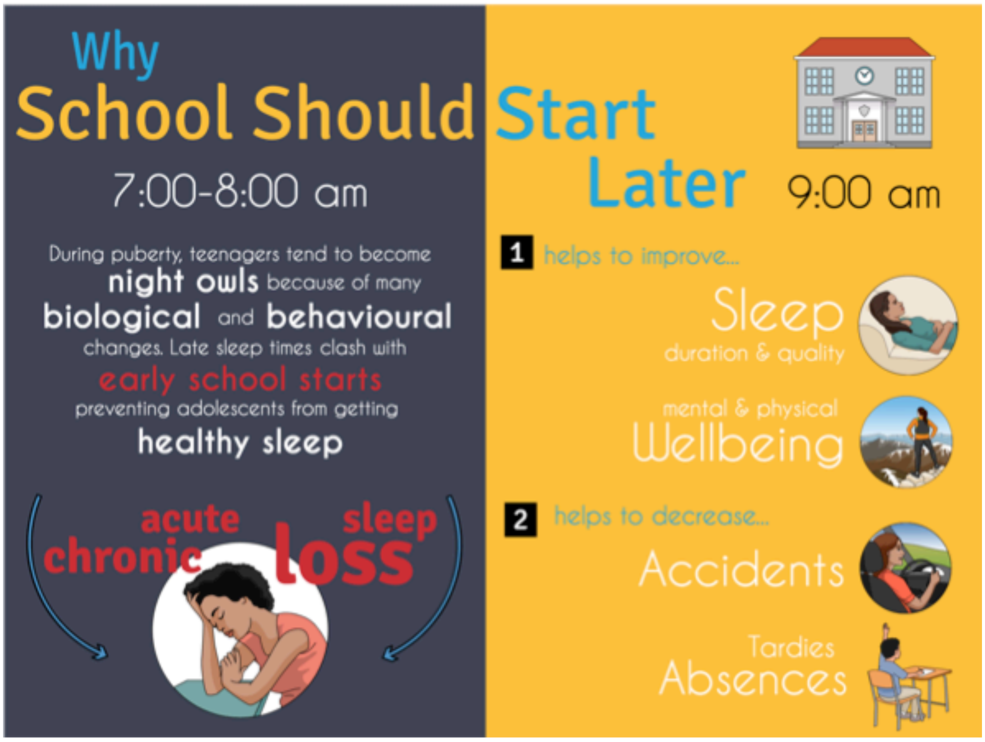
Why
School Should Start
Later
7:00-8:00 am
9:00 am
During puberty, teenagers tend to become night owls because of many biological and behavioural changes. Late sleep times clash with early school starts preventing adolescents from getting healthy sleep
acute
chronic
sleep
loss
1 helps to improve...
Sleep
duration & quality
mental & physical
Wellbeing
2 helps to decrease...
Accidents
Tardies
Absences

Fig. 11 | Infographic describing positive effects of later school start times. Figure created in the Mind the Graph platform (*www.mindthegraph.com*).

General References

1. MacPherson, H. The cosmological ideas among the Greeks. *Pop. Astron.* **24**, 358 (1916).
2. Aristotle. Aristotle. Met. 1.983b. in *Metaphysics* (Harvard University Press, 1989).
3. Ashliman, D. L. The Blind Men and the Elephant. (2014).
4. Aristotle. Aristotle. Met. 1.986b. in *Metaphysics* (Harvard University Press, 1989).
5. Kreitzman, L. & Foster, R. *The rhythms of life: The biological clocks that control the daily lives of every living thing*. (Profile books, 2011).
6. Williams, G. E. Geological constraints on the Precambrian history of Earth's rotation and the Moon's orbit. *Rev. Geophys.* **38**, 37–59 (2000).
7. Cameron, A. G. W. & Benz, W. The origin of the moon and the single impact hypothesis IV. *Icarus* **92**, 204–216 (1991).
8. Foster, R. G. & Roenneberg, T. Human Responses to the Geophysical Daily, Annual and Lunar Cycles. *Curr. Biol.* **18**, 784–794 (2008).
9. Dvornyk, V., Vinogradova, O. & Nevo, E. Origin and evolution of circadian clock genes in prokaryotes. *Proc. Natl. Acad. Sci.* **100**, 2495–2500 (2003).
10. Palmer, J. An introduction to biological rhythms. (1976).
11. Halberg, F. Physiologic 24-hour periodicity in human beings and mice, the lighting regimen and daily routine. in *Photoperiodism and related phenomena in plants and animals.* (ed. Withrow, E.) 803–878 (AAAS, 1959).
12. Golden, S. S., Ishiura, M., Johnson, C. H. & Kondo, T. Cyanobacterial circadian rhythms. *Annu. Rev. Plant Biol.* **48**, 327–354 (1997).
13. de Mairan, J. Observation botanique. *Hist. l'Academie R. des Sci. Paris* (1729).
14. Du Pré, B. C. *et al.* Circadian rhythms in cell maturation. *Physiology* **29**, 72–83 (2014).
15. Duhamel du Monceau, H. L. La physique des arbres. *Guerin & Delatour, Párizs* (1758).
16. Sweeney, B. M. *Rhythmic phenomena in plants*. (Academic Press, 2013).
17. Bunning, E. Die endonome Tagesrhythmik als Grundlage der photoperiodischen Reaktion. *Ber. Deut. Bot. Ges.* **54**, 590–607 (1937).
18. Pittendrigh, C. S. The circadian oscillation in Drosophila pseudoobscura pupae: a model for the photoperiodic clock. *Zeitschrift für Pflanzenphsysiologie* 275–307 (1966).
19. Aschoff, J. & Background, B. Cireadian Rhythms. *Science (80-.).* **148**, 1427–1432 (1965).
20. Aschoff, J. On the Relationship between Motor Activity and the Sleep-Wake Cycle in Humans during Temporal Isolation. *J. Biol. Rhythms* **8**, 33–46 (1993).
21. Kleitman, N. Biological rhythms and cycles. *Physiol. Rev.* **29**, 1–30 (1949).
22. Mundim, K. C., Baraldi, S., Machado, H. G. & Vieira, F. M. C. Temperature coefficient (Q10) and its applications in biological systems: Beyond the Arrhenius theory. *Ecol. Modell.* **431**, 109127 (2020).
23. Moore, R. Y. Organization and function of a central nervous system circadian oscillator: the suprachiasmatic hypothalamic nucleus. in *Federation proceedings* **42**, 2783–2789 (1983).
24. Ralph, M. R. & Menaker, M. A mutation of the circadian system in golden hamsters. *Science*

(80-.). **241**, 1225–1227 (1988).

25. Ralph, M. R., Foster, R. G., Davis, F. C. & Menaker, M. Transplanted suprachiasmatic nucleus determines circadian period. *Science (80-.).* **247**, 975–978 (1990).

26. Pett, J. P., Kondoff, M., Bordyugov, G., Kramer, A. & Herzel, H. Co-existing feedback loops generate tissue-specific circadian rhythms. *Life Sci. Alliance* **1**, 1–11 (2018).

27. Hardin, P. E., Hall, J. C. & Rosbash, M. Feedback of the Drosophila period gene product on circadian cycling of its messenger RNA levels. *Nature* **343**, 536–540 (1990).

28. Stratmann, M. & Schibler, U. Properties, entrainment, and physiological functions of mammalian peripheral oscillators. *J. Biol. Rhythms* **21**, 494–506 (2006).

29. Yamazaki, S. *et al.* Resetting central and peripheral circadian oscillators in transgenic rats. *Science (80-.).* **288**, 682–685 (2000).

30. Moore, R. Y. & Lenn, N. J. A retinohypothalamic projection in the rat. *J. Comp. Neurol.* **146**, 1–14 (1972).

31. Johnson, R. F., Moore, R. Y. & Morin, L. P. Loss of entrainment and anatomical plasticity after lesions of the hamster retinohypothalamic tract. *Brain Res.* **460**, 297–313 (1988).

32. Nelson, R. J. & Zucker, I. Absence of extraocular photoreception in diurnal and nocturnal rodents exposed to direct sunlight. *Comp. Biochem. Physiol. Part A Physiol.* **69**, 145–148 (1981).

33. Wright, K. P. & Czeisler, C. A. Absence of circadian phase resetting in response to bright light behind the knees. *Science (80-.).* **297**, 571 (2002).

34. Lucas, R. J., Freedman, M. S., Munoz, M., Garcia-Fernández, J.-M. & Foster, R. G. Regulation of the mammalian pineal by non-rod, non-cone, ocular photoreceptors. *Science (80-.).* **284**, 505–507 (1999).

35. Do, M. T. H. & Yau, K. W. Intrinsically photosensitive retinal ganglion cells. *Physiol. Rev.* **90**, 1547–1581 (2010).

36. Provencio, I. *et al.* A novel human opsin in the inner retina. *J. Neurosci.* **20**, 600–605 (2000).

37. Berson, D. M., Dunn, F. A. & Takao, M. Phototransduction by retinal ganglion cells that set the circadian clock. *Science (80-.).* **295**, 1070–1073 (2002).

38. Ruby, N. F. *et al.* Role of melanopsin in circadian responses to light. *Science (80-.).* **298**, 2211–2213 (2002).

39. Hattar, S., Liao, H.-W., Takao, M., Berson, D. M. & Yau, K.-W. Melanopsin-Containing Retinal Ganglion Cells: Architecture, Projections, and Intrinsic Photosensitivity. *Science (80-.).* **295**, 1065 LP – 1070 (2002).

40. Panda, S. *et al.* Melanopsin is required for non-image-forming photic responses in blind mice. *Science (80-.).* **301**, 525–527 (2003).

41. Roenneberg, T., Daan, S. & Merrow, M. The art of entrainment. *J. Biol. Rhythms* **18**, 183–194 (2003).

42. Roenneberg, T., Pilz, L. K., Zerbini, G. & Winnebeck, E. C. Chronotype and social jetlag: A (self-) critical review. *Biology (Basel).* **8**, 1–19 (2019).

43. Pittendrigh, C. S. Circadian rhythms and the circadian organization of living systems. in *Cold Spring Harbor symposia on quantitative biology* **25**, 159–184 (Cold Spring Harbor Laboratory Press, 1960).

44. Zimmerman, W. F., Pittendrigh, C. S. & Pavlidis, T. Temperature compensation of the circadian

oscillation in Drosophila pseudoobscura and its entrainment by temperature cycles. *J. Insect Physiol.* **14**, 669–684 (1968).

45. Pittendrigh, C., Bruce, V. & Kaus, P. On the significance of transients in daily rhythms. *Proc. Natl. Acad. Sci. U. S. A.* **44**, 965 (1958).

46. Aschoff, J. Die Tagesperiodik licht-und dunkelaktiver Tiere. *Rev. suisse zool* **71**, 528–558 (1964).

47. Roenneberg, T., Hut, R., Daan, S. & Merrow, M. Entrainment concepts revisited. *J. Biol. Rhythms* **25**, 329–339 (2010).

48. Stephan, F. K., Swann, J. M. & Sisk, C. L. Entrainment of circadian rhythms by feeding schedules in rats with suprachiasmatic lesions. *Behav. Neural Biol.* **25**, 545–554 (1979).

49. Marchant, E. G. & Mistlberger, R. E. Entrainment and phase shifting of circadian rhythms in mice by forced treadmill running. *Physiol. Behav.* **60**, 657–663 (1996).

50. Mistlberger, R. E. & Skene, D. J. Social influences on mammalian circadian rhythms: animal and human studies. *Biol. Rev.* **79**, 533–556 (2004).

51. Mistlberger, R. E. & Skene, D. J. Nonphotic entrainment in humans? *J. Biol. Rhythms* **20**, 339–352 (2005).

52. Kräuchi, K. The thermophysiological cascade leading to sleep initiation in relation to phase of entrainment. *Sleep Med. Rev.* **11**, 439–451 (2007).

53. Roenneberg, T. What is chronotype? *Sleep Biol. Rhythms* **10**, 75–76 (2012).

54. Klerman, E. B., Gershengorn, H. B., Duffy, J. F. & Kronauer, R. E. Comparisons of the variability of three markers of the human circadian pacemaker. *J. Biol. Rhythms* **17**, 181–193 (2002).

55. Lewy, A. J. & Sack, R. L. The dim light melatonin onset as a marker for Orcadian phase position. *Chronobiol. Int.* **6**, 93–102 (1989).

56. Lewy, A. J., Cutler, N. L. & Sack, R. L. The Endogenous Melatonin Profile as a Marker for Circadian Phase Position. *J. Biol. Rhythm. Endog. Melatonin Phase J. Biol. Rhythm.* **14**, 227–236 (1999).

57. Randler, C. CSM. Composite Scale of Morningness. in *Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)* (Elektronisches Testarchiv (ZPID), 2014). doi:https://doi.org/10.23668/psycharchives.440

58. Horne, J. A. & Östberg, O. A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int. J. Chronobiol.* (1976).

59. Wirz-Justice, A., Roenneberg, T. & Merrow, M. Life between Clocks: Daily Temporal Patterns of Human Chronotypes. *J Biol Rhythm.* **18**, 80–90 (2003).

60. Martin, S. K. & Eastman, C. I. Sleep logs of young adults with self-selected sleep times predict the dim light melatonin onset. *Chronobiol. Int.* **19**, 695–707 (2002).

61. Roenneberg, T. *et al.* Human activity and rest in situ. *Methods in Enzymology* **552**, (2015).

62. Sadeh, A., Acebo, C., Seifer, R., Aytur, S. & Carskadon, M. A. Activity-based assessment of sleep-wake patterns during the 1st year of life. *Infant Behav. Dev.* **18**, 329–337 (1995).

63. Kripke, D. F. *et al.* Wrist actigraphic scoring for sleep laboratory patients: Algorithm development. *J. Sleep Res.* **19**, 612–619 (2010).

64. Oakley, N. R. Validation with polysomnography of the Sleepwatch sleep/wake scoring algorithm used by the Actiwatch activity monitoring system. *Bend Mini Mitter, Cambridge Neurotechnology* (1997).

65. Loock, A.-S. *et al. Validation of the Munich Actimetry Sleep Detection Algorithm for estimating sleep-wake patterns from activity recordings.*

66. Rechtschaffen, A. The control of sleep. in *Human behaviour and its control.* (ed. Hynt, W.) (Shenkman Publishing Company, Inc., 1971).

67. Rattenborg, N. C. Sleeping on the wing. *Interface Focus* **7**, 0–2 (2017).

68. Joiner, W. J. Unraveling the evolutionary determinants of sleep. *Curr. Biol.* **26**, R1073–R1087 (2016).

69. Hobson, J. A. Sleep is of the brain, by the brain and for the brain. *Nature* **437**, 1254–1256 (2005).

70. Siegel, J. M. Clues to the functions of mammalian sleep. *Nature* **437**, 1264–1271 (2005).

71. Flanigan Jr, W. F., Wilcox, R. H. & Rechtschaffen, A. The EEG and behavioral continuum of the crocodilian, Caiman sclerops. *Electroencephalogr. Clin. Neurophysiol.* **34**, 521–538 (1973).

72. Aserinsky, E. & Kleitman, N. Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science (80-.).* **118**, 273–274 (1953).

73. Kety, S. S., Landau, W. M., Freygang, W. H., Rowland, L. P. & Sokoloff, L. The local circulation of the living brain; values in the unanesthetized and anesthetized cat. *Trans. Am. Neurol. Assoc* 125–129 (1955).

74. Hobson, J. A. *Dreaming:An Introduction to the Science of Sleep*. (Oxford University Press, 2002).

75. Everson, C. A., Smith, C. B. & Sokoloff, L. Effects of prolonged sleep deprivation on local rates of cerebral energy metabolism in freely moving rats. *J. Neurosci.* **14**, 6769–6778 (1994).

76. Nofzinger, E. A. *et al.* Functional neuroimaging evidence for hyperarousal in insomnia. *Am. J. Psychiatry* **161**, 2126–2128 (2004).

77. Maquet, P. *et al.* Regional organisation of brain activity during paradoxical sleep (PS). *Arch. Ital. Biol.* **142**, 413–419 (2004).

78. Borbely, A. A. Two-Process Model of Sleep Regulation. *Encycl. Neurosci.* 195–204 (1982). doi:10.1007/978-3-540-29678-2_6166

79. Daan, S., Beersma, D. G. & Borbély, A. A. Timing of human sleep: recovery process gated by a circadian pacemaker. *Am. J. Physiol. Integr. Comp. Physiol.* **246**, R161–R183 (1984).

80. Borbély, A. A., Daan, S., Wirz-Justice, A. & Deboer, T. The two-process model of sleep regulation: A reappraisal. *J. Sleep Res.* **25**, (2016).

81. Liu, C. *et al.* Molecular dissection of two distinct actions of melatonin on the suprachiasmatic circadian clock. *Neuron* **19**, 91–102 (1997).

82. Franziska Reichert, C., Cajochen, C., Schmidt, C. & Cajochen, C. Sleep-wake regulation and its impact on working memory performance: The role of adenosine. *Biology (Basel).* **5**, 1–25 (2016).

83. Carley, D. W. & Farabi, S. S. Physiology of sleep. *Diabetes Spectr.* **29**, 5–9 (2016).

84. Silber, M. H. *et al.* The visual scoring of sleep in adults. *J. Clin. Sleep Med.* **3**, 121–131 (2007).

85. Peraita-Adrados, R. Electroencephalography, polysomnography, and other sleep recording systems. *Physiol. Nat. sleep* **103**, (2005).

86. Ackermann, S. & Rasch, B. Differential effects of non-REM and REM sleep on memory consolidation? *Curr. Neurol. Neurosci. Rep.* **14**, (2014).

87. Hobson, J. A., McCarley, R. W. & Wyzinski, P. W. Sleep cycle oscillation: reciprocal discharge by two brainstem neuronal groups. *Science (80-.).* **189**, 55–58 (1975).

88. Dunmyre, J. R., Mashour, G. A. & Booth, V. Coupled flip-flop model for REM sleep regulation in the rat. *PLoS One* **9**, e94481 (2014).

89. Schmidt, M. H. The energy allocation function of sleep: A unifying theory of sleep, torpor, and continuous wakefulness. *Neurosci. Biobehav. Rev.* **47**, 122–153 (2014).

90. Aristotle. On Sleep And Sleeplessness. Available at: http://infomotions.com/etexts/philosophy/400BC-301BC/aristotle-on-267.htm. (Accessed: 27th December 2020)

91. Assefa, S. Z., Diaz-Abad, M., Wickwire, E. M. & Scharf, S. M. The functions of sleep. *AIMS Neurosci.* **2**, 155–171 (2015).

92. Horne, J. A. Sleep loss and 'divergent' thinking ability. *Sleep* **11**, 528–536 (1988).

93. Rechtschaffen, A. Current perspectives on the function of sleep. *Perspect. Biol. Med.* **41**, 359–390 (1998).

94. Zepelin, H. & Rechtschaffen, A. *Relationships between mammalian sleep parameters and other constitutional variables*. (1973).

95. Tononi, G. & Cirelli, C. Sleep function and synaptic homeostasis. *Sleep Med. Rev.* **10**, 49–62 (2006).

96. Song, C. & Tagliazucchi, E. Linking the nature and functions of sleep: insights from multimodal imaging of the sleeping brain. *Curr. Opin. Physiol.* **15**, 29–36 (2020).

97. Vyazovskiy, V. V, Cirelli, C., Pfister-Genskow, M., Faraguna, U. & Tononi, G. Molecular and electrophysiological evidence for net synaptic potentiation in wake and depression in sleep. *Nat. Neurosci.* **11**, 200–208 (2008).

98. Liu, Z.-W., Faraguna, U., Cirelli, C., Tononi, G. & Gao, X.-B. Direct evidence for wake-related increases and sleep-related decreases in synaptic strength in rodent cortex. *J. Neurosci.* **30**, 8671–8675 (2010).

99. Huber, R. *et al.* Human cortical excitability increases with time awake. *Cereb. cortex* **23**, 1–7 (2013).

100. Diering, G. H. *et al.* Homer1a drives homeostatic scaling-down of excitatory synapses during sleep. *Science (80-.).* **355**, 511–515 (2017).

101. De Vivo, L. *et al.* Ultrastructural evidence for synaptic scaling across the wake/sleep cycle. *Science (80-.).* **355**, 507–510 (2017).

102. Fogel, S. M. & Smith, C. T. The function of the sleep spindle: a physiological index of intelligence and a mechanism for sleep-dependent memory consolidation. *Neurosci. Biobehav. Rev.* **35**, 1154–1165 (2011).

103. Gais, S. & Born, J. Declarative memory consolidation: mechanisms acting during human sleep. *Learn. Mem.* **11**, 679–685 (2004).

104. Stickgold, R. & Walker, M. P. Sleep-dependent memory consolidation and reconsolidation. *Sleep Med.* **8**, 331–343 (2007).

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

106. Kopasz, M. *et al.* Sleep and memory in healthy children and adolescents - A critical review. *Sleep Med. Rev.* **14**, 167–177 (2010).

107. Smith, C. Sleep states and memory processes in humans: Procedural versus declarative memory systems. *Sleep Med. Rev.* **5**, 491–506 (2001).

108. Diekelmann, S. & Born, J. The memory function of sleep. *Nat. Rev. Neurosci.* **11**, 114–126 (2010).

109. Plihal, W. & Born, J. Effects of Early and Late Nocturnal Sleep on Declarative and Procedural Memory. *J. Cogn. Neurosci.* **9**, 534–547 (1997).

110. Wagner, U., Gais, S. & Born, J. Emotional memory formation is enhanced across sleep intervals with high amounts of rapid eye movement sleep. *Learn. Mem.* **8**, 112–119 (2001).

111. Stickgold, R. Parsing the role of sleep in memory processing. *Curr. Opin. Neurobiol.* **23**, 847–853 (2013).

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

113. Crone, E. A. & Dahl, R. E. Understanding adolescence as a period of social–affective engagement and goal flexibility. *Nat. Rev. Neurosci.* **13**, 636–650 (2012).

114. Nelson, E. E., Jarcho, J. M. & Guyer, A. E. Social re-orientation and brain development: An expanded and updated view. *Dev. Cogn. Neurosci.* **17**, 118–127 (2016).

115. Piekarski, D. J. *et al.* Does puberty mark a transition in sensitive periods for plasticity in the associative neocortex? *Brain Res.* **1654**, 123–144 (2017).

116. Petanjek, Z. *et al.* Extraordinary neoteny of synaptic spines in the human prefrontal cortex. *Proc. Natl. Acad. Sci.* **108**, 13281–13286 (2011).

117. Rakic, P., Bourgeois, J.-P. & Goldman-Rakic, P. S. Synaptic development of the cerebral cortex: implications for learning, memory, and mental illness. in *Progress in brain research* **102**, 227–243 (Elsevier, 1994).

118. Drzewiecki, C. M., Willing, J. & Juraska, J. M. Synaptic number changes in the medial prefrontal cortex across adolescence in male and female rats: a role for pubertal onset. *Synapse* **70**, 361–368 (2016).

119. Jenni, O. G. & Carskadon, M. A. Spectral analysis of the sleep electroencephalogram during adolescence. *Sleep* **27**, 774–783 (2004).

120. Huber, R. High-density sleep EEG recordings during adolescence. *J Sleep Res* 1–39 (2008).

121. Giedd, J. N. Structural magnetic resonance imaging of the adolescent brain. *Ann. N. Y. Acad. Sci.* **1021**, 77–85 (2004).

122. Carskadon, M. A., Wolfson, A. R., Acebo, C., Tzischinsky, O. & Seifer, R. Adolescent sleep patterns, circadian timing, and sleepiness at a transition to early school days. *Sleep* **21**, 871–881 (1998).

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

124. Tarokh, L. & Carskadon, M. A. Sleep in Adolescents. *Encycl. Neurosci.* 1005–1012 (2009). doi:10.1016/B978-008045046-9.00066-8

125. Jenni, O. G., Achermann, P. & Carskadon, M. A. Homeostatic sleep regulation in adolescents. *Sleep* **28**, 1446–1454 (2005).

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

127. Winnebeck, E. C. *et al.* Later school start times in a flexible system improve teenage sleep.

Sleep **43**, (2020).

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

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

130. Borisenkov, M. F., Perminova, E. V. & Kosova, A. L. Chronotype, sleep length, and school achievement of 11- to 23-year-old students in Northern European Russia. *Chronobiol. Int.* **27**, 1259–1270 (2010).

131. Fischer, D., Lombardi, D. A., Marucci-Wellman, H. & Roenneberg, T. Chronotypes in the US – Influence of age and sex. *PLoS One* **12**, 1–17 (2017).

132. Masal, E. *et al.* Effects of longitude, latitude and social factors on chronotype in Turkish students. *Pers. Individ. Dif.* **86**, 73–81 (2015).

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

134. Thorleifsdottir, B., Björnsson, J. K., Benediktsdottir, B., Gislason, T. & Kristbjarnarson, H. Sleep and sleep habits from childhood to young adulthood over a 10-year period. *J. Psychosom. Res.* **53**, 529–537 (2002).

135. Kuula, L. *et al.* Using big data to explore worldwide trends in objective sleep in the transition to adulthood. *Sleep Med.* **62**, 69–76 (2019).

136. Hagenauer, M. H., Perryman, J. I., Lee, T. M. & Carskadon, M. A. Adolescent changes in the homeostatic and circadian regulation of sleep. *Dev. Neurosci.* **31**, 276–284 (2009).

137. Rhie, S. K., Lee, S. H. & Chae, K. Y. Sleep patterns and school performance of Korean adolescents assessed using a Korean version of the pediatric daytime sleepiness scale. *Korean J. Pediatr.* **54**, 29–35 (2011).

138. Estevan, I., Silva, A., Vetter, C. & Tassino, B. Short Sleep Duration and Extremely Delayed Chronotypes in Uruguayan Youth: The Role of School Start Times and Social Constraints. *J. Biol. Rhythms* 0748730420927601 (2020).

139. Randler, C. Differences in sleep and circadian preference between Eastern and Western German adolescents. *Chronobiol. Int.* **25**, 565–575 (2008).

140. Estevan, I., Silva, A. & Tassino, B. School start times matter, eveningness does not. *Chronobiol. Int.* **35**, 1753–1757 (2018).

141. Borisenkov, M. F. The pattern of entrainment of the human sleep-wake rhythm by the natural photoperiod in the north. *Chronobiol. Int.* **28**, 921–929 (2011).

142. Roenneberg, T. *et al.* Epidemiology of the human circadian clock. *Sleep Med. Rev.* **11**, 429–438 (2007).

143. Roenneberg, T. The human circadian clock entrains to sum time. **17**, 44–45

144. Carter, B., Rees, P., Hale, L., Bhattacharjee, D. & Paradkar, M. A meta-analysis of the effect of media devices on sleep outcomes. *JAMA Pediatr.* **170**, 1202 (2016).

145. Hysing, M. *et al.* Sleep and use of electronic devices in adolescence: results from a large population-based study. *BMJ Open* **5**, e006748–e006748 (2015).

146. Carskadon, M. A. Evolution of sleep and daytime sleepiness in adolescents. *Sleep/wake Disord. Nat. Hist. Epidemiol. long-term Evol.* (1983).

147. Hirshkowitz, M. *et al.* National Sleep Foundation's updated sleep duration recommendations:

Final report. *Sleep Heal.* **1**, 233–243 (2015).

148. 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.* **12**, 110–118 (2011).

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

150. Komada, Y., Okajima, I., Kitamura, S. & Inoue, Y. A survey on social jetlag in Japan: a nationwide, cross-sectional internet survey. *Sleep Biol. Rhythms* **17**, 417–422 (2019).

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

152. Roenneberg, T., Kantermann, T., Juda, M., Vetter, C. & Allebrandt, K. V. Light and the human circadian clock. in *Circadian clocks* 311–331 (Springer, 2013).

153. Åkerstedt, T. *et al.* Sleep duration and mortality – Does weekend sleep matter? *J. Sleep Res.* **28**, 1–11 (2019).

154. Bonnet, M. H. & Arand, D. L. We are chronically sleep deprived. *Sleep* **18**, 908–911 (1995).

155. Roenneberg, T. The human sleep project. *Nature* **498**, 427–428 (2013).

156. Riemann, D., Berger, M. & Voderholzer, U. Sleep and depression—results from psychobiological studies: an overview. *Biol. Psychol.* **57**, 67–103 (2001).

157. Kronfeld-Schor, N. & Einat, H. Circadian rhythms and depression: human psychopathology and animal models. *Neuropharmacology* **62**, 101–114 (2012).

158. Beebe, D. W. Cognitive, behavioral, and functional consequences of inadequate sleep in children and adolescents. *Pediatr. Clin.* **58**, 649–665 (2011).

159. Fitzgerald, C. T., Messias, E. & Buysse, D. J. Teen sleep and suicidality: results from the youth risk behavior surveys of 2007 and 2009. *J. Clin. sleep Med.* (2011).

160. Fredriksen, K., Rhodes, J., Reddy, R. & Way, N. Sleepless in Chicago: Tracking the Effects of Adolescent Sleep Loss During the Middle School Years. *Child Dev.* **75**, 84–95 (2004).

161. Kaplan, H. L. & Sadock, B. J. Schizophrenia and other psychotic disorders. *Synopsis psychiatry, 8th ed. Balt. Williams Wilkins* 456–500 (2002).

162. Levandovski, R. *et al.* Depression Scores Associate With Chronotype and Social Jetlag in a Rural Population. *Chronobiol. Int.* **28**, 771–778 (2011).

163. Wiebe, S. T., Cassoff, J. & Gruber, R. Sleep patterns and the risk for unipolar depression: A review. *Nat. Sci. Sleep* **4**, 63–71 (2012).

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

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

166. Roenneberg, T., Allebrandt, K. V., Merrow, M. & Vetter, C. Social jetlag and obesity. *Curr. Biol.* **22**, 939–943 (2012).

167. Kantermann, T. *et al.* Atherosclerotic risk and social jetlag in rotating shift-workers: first evidence from a pilot study. *Work* **46**, 273–282 (2013).

168. McKnight-Eily, L. R. *et al.* Relationships between hours of sleep and health-risk behaviors in US adolescent students. *Prev. Med. (Baltim).* **53**, 271–273 (2011).

169. Dahl, R. E. & Lewin, D. S. Pathways to adolescent health: Sleep regulation and behavior. *J. Adolesc. Heal.* (2002). doi:10.1016/S1054-139X(02)00506-2

170. Gau, S. S.-F. *et al.* Association between morningness-eveningness and behavioral/emotional problems among adolescents. *J. Biol. Rhythms* **22**, 268–274 (2007).

171. Bolton, J. M., Robinson, J. & Sareen, J. Self-medication of mood disorders with alcohol and drugs in the National Epidemiologic Survey on Alcohol and Related Conditions. *J. Affect. Disord.* **115**, 367–375 (2009).

172. Taylor, D. J. & Bramoweth, A. D. Patterns and Consequences of Inadequate Sleep in College Students: Substance Use and Motor Vehicle Accidents. *J. Adolesc. Heal.* **46**, 610–612 (2010).

173. Pizza, F. *et al.* Sleep quality and motor vehicle crashes in adolescents. *J. Clin. Sleep Med.* **6**, 41–45 (2010).

174. Lufi, D., Tzischinsky, O. & Hadar, S. Delaying school starting time by one hour: Some effects on attention levels in adolescents. *J. Clin. Sleep Med.* **7**, 137–143 (2011).

175. Vedaa, Ø., Saxvig, I. W., Wilhelmsen-Langeland, A., Bjorvatn, B. & Pallesen, S. School start time, sleepiness and functioning in Norwegian adolescents. *Scand. J. Educ. Res.* **56**, 55–67 (2012).

176. Owens, D. S. *et al.* Diurnal trends in mood and performance do not all parallel alertness. *Scand. J. Work. Environ. Health* 109–114 (1998).

177. Casagrande, M., Violani, C., Curcio, G. & Bertini, M. Assessing vigilance through a brief pencil and paper letter cancellation task (LCT): effects of one night of sleep deprivation and of the time of day. *Ergonomics* **40**, 613–630 (1997).

178. Rosekind, M. R., Gander, P. H. & Dinges, D. F. *Alertness management in flight operations: Strategic napping*. (SAE Technical Paper, 1991).

179. Fairclough, S. H. & Graham, R. Impairment of driving performance caused by sleep deprivation or alcohol: a comparative study. *Hum. Factors* **41**, 118–128 (1999).

180. Gillen-O'Neel, C., Huynh, V. W. & Fuligni, A. J. To Study or to Sleep? The Academic Costs of Extra Studying at the Expense of Sleep. *Child Dev.* **84**, 133–142 (2013).

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

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

183. Killgore, W. D. S. *et al.* Sleep deprivation reduces perceived emotional intelligence and constructive thinking skills. *Sleep Med.* **9**, 517–526 (2008).

184. Randazzo, A. C., Muehlbach, M. J., Schweitzer, P. K. & Walsh, J. K. Cognitive Function Following Acute Sleep Restriction in Children Ages 10–14. *Sleep* **21**, (1998).

185. Sadeh, A., Gruber, R. & Raviv, A. The effects of sleep restriction and extension on school-age children: What a difference an hour makes. *Child Dev.* **74**, 444–455 (2003).

186. Harrison, Y. & Horne, J. A. Sleep deprivation affects speech. *Sleep* **20**, 871–877 (1997).

187. Harrison, Y. & Horne, J. A. Sleep loss impairs short and novel language tasks having a prefrontal focus. *J. Sleep Res.* **7**, 95–100 (1998).

188. Curcio, G., Ferrara, M. & De Gennaro, L. Sleep loss, learning capacity and academic performance. *Sleep Medicine Reviews* **10**, 323–337 (2006).

189. Durmer, J. S. & Dinges, D. F. Neurocognitive consequences of sleep deprivation. in *Seminars in neurology* **25**, 117–129 (Copyright© 2005 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New …, 2005).

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

191. Murphy, S. L., Xu, J., Kochanek, K. D. & Arias, E. Mortality in the United States, 2017. *NCHS Data Brief* 1–8 (2018).

192. Paruthi, S. *et al.* Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine. *J. Clin. Sleep Med.* **12**, 785–786 (2016).

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

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

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

196. Kuula, L. *et al.* Development of Late Circadian Preference: Sleep Timing From Childhood to Late Adolescence. *J. Pediatr.* **194**, 182-189.e1 (2018).

197. Short, M. A. *et al.* A Cross-Cultural Comparison of Sleep Duration Between U.S. and Australian Adolescents: The Effect of School Start Time, Parent-Set Bedtimes, and Extracurricular Load. *Heal. Educ. Behav.* **40**, 323–330 (2013).

198. WELT.de. PISA-Studie: Deutsche Schulreform kommt nur langsam voran. (2005).

199. Dieckmann, C. Das Grauen um 8 Uhr morgens. (2019). Available at: https://www.bayerische-staatszeitung.de/staatszeitung/leben-in-bayern/detailansicht-leben-in-bayern/artikel/das-grauen-um-8-uhr-morgens.html#topPosition. (Accessed: 30th December 2020)

200. PISA - Internationale Schulleistungsstudie der OECD. (2018). Available at: http://www.oecd.org/berlin/themen/pisa-studie/. (Accessed: 30th December 2020)

201. Kuhn, A. Der größte Fehler ist die Rückkehr von G8 auf G9. (2020). Available at: https://deutsches-schulportal.de/bildungswesen/schulreformen-der-groesste-fehler-ist-die-rueckkehr-von-g8-zu-g9/. (Accessed: 30th December 2020)

202. Widome, R. *et al.* Association of Delaying School Start Time with Sleep Duration, Timing, and Quality among Adolescents. *JAMA Pediatr.* **174**, 697–704 (2020).

203. Lo, J. C. *et al.* Sustained benefits of delaying school start time on adolescent sleep and well-being. *Sleep* **41**, (2018).

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

205. Das-Friebel, A., Gkiouleka, A., Grob, A. & Lemola, S. Effects of a 20 minutes delay in school start time on bed and wake up times, daytime tiredness, behavioral persistence, and positive attitude towards life in adolescents. *Sleep Med* **66**, 103–109 (2020).

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

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

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

209. Goldin, A. P., Sigman, M., Braier, G., Golombek, D. A. & Leone, M. J. Interplay of chronotype and school timing predicts school performance. *Nat. Hum. Behav.* **4**, 387–396 (2020).

210. Biller, A. M. *et al. One year later: longitudinal effects of flexible school start times on teenage sleep, psychological benefits, and academic grades*. (2020).

211. Thacher, P. V & Onyper, S. V. Longitudinal Outcomes of Start Time Delay on Sleep, Behavior, and Achievement in High School. *Sleep* **39**, 271–281 (2016).

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

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

214. Sharman, R. & Illingworth, G. Adolescent sleep and school performance — the problem of sleepy teenagers. *Curr. Opin. Physiol.* **15**, 23–28 (2020).

215. Berger, A. T., Widome, R. & Troxel, W. M. Delayed school start times and adolescent health. in *Sleep and Health* 447–454 (2019). doi:10.1016/B978-0-12-815373-4.00033-2

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

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

218. Skeldon, A. C., Phillips, A. J. K. & Dijk, D.-J. The effects of self-selected light-dark cycles and social constraints on human sleep and circadian timing: a modeling approach. *Sci. Rep.* **7**, 45158 (2017).

219. Kalak, N. *et al.* Daily morning running for 3 weeks improved sleep and psychological functioning in healthy adolescents compared with controls. *J. Adolesc. Heal.* **51**, 615–622 (2012).

220. Van der Lely, S. *et al.* Blue blocker glasses as a countermeasure for alerting effects of evening light-emitting diode screen exposure in male teenagers. *J. Adolesc. Heal.* **56**, 113–119 (2015).

221. Misiunaite, I., Eastman, C. I. & Crowley, S. J. Circadian phase advances in response to weekend morning light in adolescents with short sleep and late bedtimes on school nights. *Front. Neurosci.* **14**, 99 (2020).

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

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

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

225. Shin, J. *Sleep More, Study Less ? The Impact of Delayed School Start Time on Sleep and Academic Performance*. (2018).

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

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

228. Harrison, Y. & Horne, J. The impact of sleep deprivation on decision making: a review. *J. Exp. Psychol. Appl.* **6 3**, 236–249 (2000).

229. Carskadon, M. A., Vieira, C. & Acebo, C. Association between puberty and delayed phase preference. *Sleep* **16**, 258–262 (1993).

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

231. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13**, 407–420 (2012).

232. Hillman, E. M. C. Coupling mechanism and significance of the BOLD signal: a stfile:///Users/anna/LRZ Sync+Share/Anna (personal)/1 %7C PhD Unterlagen/1_THESIS/y_Background and literature/Silber.pdfatus report. *Annu. Rev. Neurosci.* **37**, 161–181 (2014).

233. Rétey, J. V. *et al.* A genetic variation in the adenosine A2A receptor gene (ADORA2A) contributes to individual sensitivity to caffeine effects on sleep. *Clin. Pharmacol. Ther.* **81**, 692–698 (2007).

234. Scheibehenne, B., Greifeneder, R. & Todd, P. M. Can there ever be too many options? A meta-analytic review of choice overload. *J. Consum. Res.* **37**, 409–425 (2010).

Acknowledgements

First and foremost, I would like to wholeheartedly thank my parents: Mama, Papa, ihr seids da wenn i eich brauch, ihr glaubts an mi, ihr habts ma zoagt, nach vorn zu schaun und ned z‘bereun, ihr habts ma glernt, dankbar zu sein im Lebn. Immer a Gaudi mit eich – so solls bleibn!

I would like to thank my supervisor, Dr. Eva Winnebeck, with whom I share a great and colourful journey throughout the last four years. It was a tough school, but a high quality one! I’ve learned so many things from you I can’t even name them all; from English grammar, concise writing, coding, being critical, learning to accept feedback, … but maybe the most important thing was to persevere, not to give up – to strive to yet a better version than I would have imagined (even after version number 3875XYZ of each manuscript). Sometimes, it was a high mountain to climb but it was always worth it! I’m incredible thankful for all your support, Eva.

Nothing would have been possible without Prof. Till Roenneberg, le grand maître. You inspired me and very often totally surprised me but the one thing that I’m most grateful for is that you let me sleep as long as I wanted! Due to this *laisser-faire* attitude I managed to have a negative social jetlag – lie-ins throughout the week but early wake-ups for mountaineering on the weekend! I’m entirely certain that your trust in a good night’s sleep kept me mentally as healthy as possible throughout my PhD.

I would like to thank all my friends with whom I share mountain adventures and my passion for nature. You all kept me healthy, happy, thankful and deeply humble. Your friendship means everything to me. I can’t wait to go out and celebrate with you all, each one of you was part of this journey. Thank you very much Carmen, for sharing the pain, graphing problems, and the sweets!

I’m also deeply happy to be part of the circadian biology and sleep field – I’ve met so many wonderful, inspiring, friendly and lovely people, friends with whom I share a common passion. I hope I will continue working with you or start new collaborations, you all inspired me and kept me going.

Thank you, Tom, for being the most critical person I know and yet the warm and loving person you are; for forcing me to go outdoors and get some sunlight and fresh air, for cooking so I wouldn’t starve. Simply, for your love.

Last but definitely not least, I want to deeply thank the GSN for great structural and uncomplicated financial support – without you, I certainly couldn’t have done it. And finally, thanks to the reader who made it until here! Thanks for your perseverance and motivation and hopefully it was - at best - a fun read.