LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

The Long-Term Proficiency of Early, Middle, and Late Starters Learning English as a Foreign Language at School: A Narrative Review and Empirical Study

Jürgen Baumert,[a] Johanna Fleckenstein,[b] Michael Leucht,[b] Olaf Köller,[b] and Jens Möller[c]

[a]Max Planck Institute for Human Development, Berlin, [b]Leibniz Institute for Science and Mathematics Education at Kiel University, and [c]Kiel University

Throughout Europe, there is a growing trend for students to start learning foreign languages at elementary school. Although policymakers expect early-start programs to boost second language skills, empirical findings are mixed; recent studies have raised many questions. In this large-scale study, we aimed to close some of these gaps. We examined the effects of early-start English on receptive language proficiency in a random sample of 19,858 students from 1,431 Year 9 classes in Germany, comparing the reading and listening comprehension of early starters (English from Year 1), a middle group (Year 3), and late starters (Year 5), and analyzing to what extent foreign language instruction at secondary level builds on students' existing knowledge. By Year 9, the proficiency levels of the three groups differed only slightly. We provide evidence that this lack of long-term impact may be attributable to English teaching at secondary level being insufficiently adaptive to students' prior knowledge.

Keywords early foreign language learning; receptive language skills; learning rate; age of onset; amount of exposure; proficiency

We would like to thank the reviewers and the editor for extremely careful reviews, comments, and suggestions that helped to strengthen this article. The responsibility for any errors or weaknesses lies with the authors. We thank Susannah Goss for translation and editing.

Correspondence concerning this article should be addressed to Jürgen Baumert, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. E-mail: sekbaumert@mpib-berlin.mpg.de

The handling editor for this article was Emma Marsden.

DOI: 10.1111/lang.12414

Introduction

In quantitative terms, the provision of early language learning at elementary level can be described as a European success story (European Commission, 2017a). The European Union's "1 + 2" language policy states that every European should learn to speak two languages in addition to their first language (L1) during compulsory education (notwithstanding that many children in European countries grow up with more than one L1), and the European Commission (2017b) assumes that "the best way to achieve this would be to introduce children to 2 foreign languages from an early age." This positive view of early language learning is echoed in political and administrative progress reports. In Germany, for example, the Standing Conference of Ministers of Education and Cultural Affairs (KMK) has concluded that "foreign language teaching at elementary level … has been widely accepted and is being successfully implemented" (KMK, 2013, p. 10, our translation). Teachers and teacher educators are generally positive about the benefits of foreign language learning at elementary level, but see room for improvement in its implementation (Hempel, Kötter, & Rymarczyk, 2018).

Empirical findings seem to tell another story, however. Review studies consistently report that later starters learn faster, and conclusions on the long-term effects of an early start are mixed to negative (Huang, 2016; Lambelet & Berthele, 2015; Muñoz & Singleton, 2011; Pfenninger & Singleton, 2017, 2019). Pfenninger and Singleton are perhaps the most forceful critics, arguing that early-start programs are built on a myth: "There is no real dispute about the scientific facts, which are that primary school instruction in L2 fails to equip learners with a level of L2 proficiency which by the end of secondary schooling is superior to that of those whose instruction begins later" (Singleton & Pfenninger, 2019, p. 30). Given these contrasting conclusions, a careful review of the relevant research seemed warranted. Our review will show that findings are not conclusive and that recent studies on the long-term effects of an early start raise more questions than they answer.

Following the review, we present a large-scale study in which we aimed to answer some of these questions. Our analysis examined whether an early start has positive effects on the long-term development of language proficiency as measured at the end of compulsory education and investigated whether secondary language instruction succeeds in adapting to and building on students' prior knowledge. Specifically, we compared the effects of starting English at different times in a large nonselective random sample of Year 9 students (age 15–16 years) representative for Germany. The starting times we compared, with contrasts in the age of onset (AO), were as follows: at the beginning of elementary school (AO: 6–7 years), somewhat later in elementary school (AO: 8–9 years), or at secondary level (AO: 10 years). The study's quasi-experimental design allowed us to carefully control for individual and institutional covariates that may be confounded with enrollment in early-start programs.

Background Literature

Theoretical Underpinnings for Early-Start Programs, and Research Findings

Early-start programs are widely expected to foster swift and successful language learning and to positively impact factors such as motivation, intercultural understanding, and willingness to communicate. These expectations are largely informed by two ideas that have influenced the theory and practice of language teaching and learning for decades. On the one hand, the critical period hypothesis (for a summary, see DeKeyser & Larson-Hall, 2005) states that there is a sensitive time window during which learners pick up certain aspects of language knowledge and skills (primarily related to phonology and grammar) in ways that are thought to be superior in terms of the nature and durability of the knowledge. Supportive evidence derives mainly from studies on the language acquisition of early and late bilinguals, including immigrants. However, applying these ideas to early foreign language instruction may not be appropriate given that naturalistic contexts and immersion programs provide very different amounts and types of input relative to the low-input environment of elementary schooling (for a review, see Singleton & Muñoz, 2011). The second idea associated with introducing languages early is that the more learning opportunities that are available, the better the result (aligning with perspectives on learning that foreground the importance of frequency of input and practice). Yet the value of increased input and practice may be moderated by characteristics such as age or aptitude, instructional approach, and context (DeKeyser, 2020).

In sum, these two basic ideas can be used to argue that starting to learn a language at the beginning of elementary education (i.e., providing early *and* increased exposure) will lead to better language proficiency. However, recent reviews and studies have challenged such arguments. In the following sections, we highlight key findings to date in four main areas relevant to the current study: proficiency at the end of elementary school, age-dependent learning rates, longer-term outcomes of an early start (at college and at secondary school), and individual and institutional influences on the effects of an early start.

Note that we did not aim in this study to address theoretical issues relating to the precise mechanisms involved in earlier versus later language learning; rather, we aimed to provide a large-scale evaluation of policies of starting to teach foreign languages before secondary school by focusing on the long-term effects of an early start.

Foreign Language Proficiency at the End of Elementary School

Longitudinal studies of early-start programs have reported substantial gains in proficiency by the end of elementary level, dependent on the amount of exposure to the L2 (Goorhuis-Brouwer & de Bot, 2010; Graham, Courtney, Marinis, & Tonkyn, 2017; Heinzmann, Müller, Oliveira, Haenni Hoti, & Wicki, 2009; Hopp, Vogelbacher, Kieseier, & Thoma, 2019; Szpotowicz & Lindgren, 2011; Unsworth, Persson, Prins, & de Bot, 2015), with effect sizes of annual learning gains ranging from Cohen's $d \approx 0.30$ to $d \approx 1.0$. Likewise, cross-sectional findings have shown that most elementary students reach curriculum attainment targets equivalent to level A1 of the Common European Framework of Reference for Languages (CEFR) after 2 to 3 years of foreign language instruction (Council of Europe, 2001, 2018).[1] The Evaluation of English in Primary School (EVENING) study drew on large random samples ($N = 1{,}748$ and $N = 1{,}344$) to investigate the English proficiency of two cohorts of Year 4 students (age 9–10 years) in North Rhine-Westphalia, Germany's most populous state. After 2 years of English instruction, most Year 4 students had reached or surpassed the curriculum attainment targets for reading and listening comprehension and speaking, equivalent to level A1 (listening comprehension: A1/A2.1)[2] of the CEFR (Groot-Wilken & Husfeldt, 2013). The findings of the EVENING study have been broadly reproduced in several large-scale studies on learning English in Germany, Switzerland, and Liechtenstein (the BIG study; Barucki et al., 2015; Von Ow, Husfeldt, & Bader-Lehmann, 2012). Similar findings for an early start in French have been reported from Switzerland (Peyer, Andexlinger, Kofler, & Lenz, 2016). In a large-scale study in Switzerland, Heinzmann et al. (2009) found that after 4 years of early-start English, the majority of students had reached or surpassed the target level of A2 (CEFR).

To summarize, the learning gains observed for early-start programs are substantial. After 2 to 3 years of two lessons per week in English as a foreign language, most students reach a level corresponding to CEFR level A1.1 to A1.2 for receptive skills; in some countries (e.g., the Netherlands and Sweden), they probably far exceed it (see Szpotowicz & Lindgren, 2011, pp. 132–133). There is thus no doubting the effectiveness of early-start programs at the point of leaving elementary school. This finding is crucial when it comes to interpreting results on the long-term effects of early-start programs, which we report later.

Age-Dependent Learning Rates at the Start of Second Language Education

To what extent do late starters benefit from a “catch-up” effect due to age-dependent rates of learning, with older students (aged 10 years and above) learning faster than younger students (aged 5–8 years)? The reviews by Lambelet and Berthele (2015) and Huang (2016) summarized research investigating age-dependent rates of foreign language learning in the school context. The studies reviewed by these authors typically used a design in which AO was varied systematically and the length or amount of exposure was held constant by testing students of different ages. Both reviews reported significant learning rate advantages for older students (see Appendix S1, Table S1.1, in the Supporting Information online).

The Barcelona Age Factor Project (BAF) is arguably one of the most important studies on age-dependent rates of foreign language learning to date (Muñoz, 2006a). The quasi-experimental study took advantage of a change in the Catalan education system, when students began learning English in Year 3 (AO: 8 years) of elementary school rather than in Year 6 (AO: 11 years). During a transitional period, early and late starters belonging to different cohorts were taught in the same schools. The study ran from 1996 to 2002 with two cohorts of bilingual students (Catalan–Spanish) who began learning English as a L3 at age 8 ($N = 164$) or 11 ($N = 107$) years. Both cohorts were tested after 200, 416, and 726 hours of instruction. Late starters showed considerably faster rates of learning than early starters after 200 and 416 hours of instruction. Their mean proficiency scores after the same amount of exposure were higher, and the size of the achievement gap increased from the first to the second point of measurement. By the third point of measurement, after 726 hours of instruction (and about 7 years), however, there was no longer a significant difference in the groups’ learning rates: The differences in favor of the late starters observed after 416 hours of instruction had decreased or remained stable, depending on the language dimension under consideration (Muñoz, 2006b).

The BAF study also offered the opportunity to investigate whether the learning rate advantage of late starters offset the exposure advantage of early starters. For a study to answer this type of question, both cohorts have to be tested at least once at the same age after differing amounts of exposure. In the BAF study, this was the case at one point in the study: at the early starters’ second point of measurement and the late starters’ first point of measurement.

Here, both cohorts were tested at the end of Year 7, that is, at the same age, but after 416 versus 200 hours or 5 versus 2 years of English, respectively. A comparison of the means at the end of Year 7 published by Muñoz (2006b, p. 26) revealed that the late starters did not catch up fully in any dimension. The early starters' advantage ranged from Cohen's $d = 0.33$ (cloze test) to $d = 1.5$ (listening comprehension). This was the net effect of 216 hours' additional exposure. At the third point of measurement, after 726 hours of instruction for both cohorts, the two cohorts then differed in age: The early starters were in Year 9, the late starters in Year 12. At that point, therefore, an estimate of whether the learning rate of later starters offset the greater exposure of early starters was not possible (as this would have required the participants to be the same age at the time of testing). Kalberer (2007) found a similar pattern of results for a small convenience sample of students at academic-track secondary schools in Switzerland. These findings are crucial as they provide a point of comparison for results from Switzerland reported below (Pfenninger & Singleton, 2017).

Longer-Term Effects of Early-Start Programs

Effects at College/University Level

Various strands of research have investigated the long-term benefits for foreign language proficiency of an early start in elementary school. One cluster of studies has investigated the language proficiency of young adults at college and examined the impact of their different AOs. We identified five such studies (summarized in Appendix S1, Table S1.2, in the Supporting Information online). Their findings either were mixed (Larson-Hall, 2008; Lin, Chang, & Cheung, 2004) or suggested no relationship between AO and language proficiency (Al Thubaiti, 2010; Muñoz, 2011, 2014). However, all five studies share two serious limitations. First, they were based on small, highly selective convenience samples of college students who needed a good standard of English to qualify for tertiary education or were even majoring in English. Second, the reliability of their findings is dependent on the equivalence of learning opportunities after transfer to secondary school, which is almost impossible to control retrospectively. Thus, these studies are not a reliable basis for gauging the long-term effects of early-start programs.

We therefore now move on to describe studies comparing the proficiency of early and late starters when at the same age during the secondary school years, as these are more useful for evaluating the general effects of an early start. We distinguish studies that found support for a sustained impact of an early start from studies that found no such support. Given its relevance to the present study's design and aims, we present the Beyond Age Effects study (BAE; Pfenninger & Singleton, 2017) in some detail.

Effects at Secondary School: Studies Finding a Sustained Positive Impact of an Early Start

We highlight here three studies that provide some evidence of positive, although limited, benefits of an early start. One study, conducted by Oller and Nagato (1974), drew on a convenience sample of 223 students at private girls' secondary schools in Japan. Oller and Nagato tested English proficiency in three cohorts of Year 7, 9, and 11 students. In each cohort, early starters (beginning in Year 1 of elementary school; AO: 6 years) were compared with late starters (beginning in Year 7 of secondary school; AO: 13 years). From Year 8 onward, early and late starters were taught together. The late starters lagged behind by Cohen's $d = 1.1$ at the end of Year 7 after 1 year of English, by $d = 0.56$ at the end of Year 9, and by $d = 0.36$ at the end of Year 11. In other words, the gap between early and late starters decreased over time but did not disappear entirely even after 4 years of joint English lessons. On the one hand, these findings testify to a learning rate advantage for late starters, as discussed in the previous section, and/or they may indicate that being taught together has a leveling-out effect. On the other hand, they also show that an early start had a sustained positive effect on students' language proficiency. This finding is consistent with the results of Muñoz (2006b), according to which late starters' learning rate advantage did not fully offset early starters' 216-hour exposure advantage for learning English.

A second study, by Mihaljevic Djigunovic, Nikolov, and Otto (2008), used large convenience samples to examine the English proficiency of Year 8 students (age 14 years) in Croatia and Hungary who started to learn English in Year 4 (AO: 10 years) or earlier. They found moderate to strong significant correlations between AO and test scores. The correlations with the total score (covering listening, reading, and writing skills) indicated mean differences of $d \approx 1.8$ in Hungary and $d \approx 0.52$ in Croatia in favor of the early starters.

Finally, Boyson, Semmer, Thompson, and Rosenbusch (2013) reported significant effects of a similar size for an early-start program in Spanish. They investigated the switch from a short-sequence Spanish program (Years 5–8; AO: 10 years) to a long-sequence program (K–8; AO: 5 years) in Connecticut, United States. At the end of Year 8, the early starters significantly outperformed the late starters, with effect sizes of between $d = 0.67$ for grammar and $d = 1.12$ for listening comprehension.

Effects at Secondary School: Studies Finding No Positive Impact of an Early Start

We highlight here three key studies that have shown negligible or no positive impact of an early start, before providing a more detailed account of one highly relevant study. Burstall, Jamieson, Cohen, and Hargreaves (1974) studied the effects of early French lessons in England and Wales, examining three cohorts of students who began to learn French in Year 4 (AO: 8 years), with sample sizes of between 5,000 and 6,000 students. Burstall (1975, p. 195) summarized the findings as follows: "By the age of 16, the only area in which the pupils taught French from the age of 8 consistently showed any superiority was that of listening comprehension." No information was given on the quality of transition arrangements between primary and secondary level, that is, on whether or not secondary schools responded adaptively to incoming students' knowledge levels. Furthermore, Bennett (1975) cast doubt on the Burstall et al. findings, citing insufficient control for group differences that may have impacted proficiency levels.

Genelot (1997) reported similar findings for early-start English in France. Genelot compared the total scores of 1,000 students participating in an early-start program in Dijon, who began English lessons in Year 4 or 5 of elementary school (AO: 8 or 9 years), with a control group of 500 late starters, who began in the first year of secondary school (Year 6; AO: 10 years). She found that the early starters were at a statistically significant advantage at the end of Year 6, but that the late starters had closed the gap by the end of Year 7 after 2 years of English instruction (Genelot, 1997, p. 39).

In a recent study, Jaekel, Schurig, Florian, and Ritter (2017) used a shift in the onset of English instruction from Year 3 (AO: 8 years) to Year 1 (AO: 6 years) of elementary schooling in North Rhine-Westphalia as a natural experiment to examine the medium-term effects of an early start. The study used a highly selective convenience sample from 31 *Gymnasium* schools (the academic secondary track in Germany; see Appendix S2, Table S2.1, in the Supporting Information online). It compared 2,632 control group students (AO: 8 years) with 2,468 early starters of the same age (AO: 6 years). As expected, at the beginning of Year 5 (age 10 years), the AO-6 starters significantly outperformed the AO-8 starters in reading and listening comprehension, with effect sizes of $d = 0.28$ and $d = 0.34$, respectively. Two years later, however, the AO-6 starters lagged significantly behind the AO-8 starters, with effect sizes of $d = -0.35$ in reading and $d = -0.17$ in listening comprehension. This pattern of results is inconsistent with all other findings reported thus far, as it suggests a disadvantage of an early start. Although the authors discussed potential moderating factors that may have contributed to these effects, this main finding is still surprising given that both groups began to learn English in elementary school—meaning that both were affected by discontinuity at the transition to secondary level—and both groups were taught in the same *Gymnasium* schools. Notably, however, a publication by Ritter, Jaekel, Meister, and Lewandowska (2015, p. 315) noted that 295 of the students in the AO-8 cohort received considerably *more* English instruction at secondary level than the AO-6 cohort. The Jaekel et al. (2017) study did not control for such exposure differences between the two cohorts after transfer to secondary school.

The Beyond Age Effects Study. Of particular relevance to the context of the present study is the longitudinal Beyond Age Effects (BAE) study conducted in the German-speaking Swiss canton of Zurich (Pfenninger & Singleton, 2017). The BAE study used the shift from late- to early-start English initiated in the 2004–2005 school year in Zurich[3]—from Year 7 of secondary school (AO: 13 years) to Year 2 of elementary school (AO: 8 years)—as a natural experiment: During a transitional period, some secondary students came from elementary schools that had already transitioned to an early start, whereas others in the same year did not. During this period, it was possible to compare, within the same secondary school, the English proficiency of students of the same age and cohort but a different AO (see also Kalberer, 2007). The BAE study drew on a convenience sample of five *Gymnasium* schools: highly selective academic-track schools attended by (at Year 7) the best-performing 15% or so of students in a year. The students ($N = 325$ in the treatment group and $N = 311$ in the control group) came from a total of 12 classes, with early and late starters being taught in separate classes. The English proficiency and learning gains of AO-8 and AO-13 students were gauged at two points of measurement: the middle of Year 7 and the end of Year 12. Proficiency in standard German was assessed by means of an argumentative essay.

The descriptive findings provide valuable insights into the English proficiency of early versus late starters (Pfenninger, 2016, p. 225; Pfenninger & Singleton, 2016, p. 323; 2017, p. 61). Even at the first point of measurement, when the late starters had had 6 months of English lessons (= 50 hours) and the early starters 5.5 years (= 440 hours), the late starters were already well ahead in terms of linguistic accuracy. The error rates differed by $d = 0.50$ for written language production and $d = 0.40$ for oral language production. In contrast, the early starters performed significantly better in more semantic dimensions, with effect sizes ranging from $d = 0.70$ (lexical complexity) to $d = 1.0$ (receptive vocabulary). Given their considerable exposure advantage, however, the early starters' lead in the semantic dimensions was small. To put these findings in context, the BAF study (Muñoz, 2006b), in contrast, reported similar or larger performance gaps in *early* starters' favor in *all* language dimensions after 2 years of English and with a much smaller exposure advantage (416 vs. 200 hours instead of 440 vs. 50 hours in the BAE study). By the second point of measurement, at the end of Year 12, there was no difference in the performance of the two groups. All students made good progress over their 6 years of English instruction at *Gymnasium*. Among those who had the same level of proficiency in Year 7, early and late starters showed the same learning gains. Among those who had different levels of proficiency in Year 7, the gaps were closed by higher gains in the group that initially lagged behind—be it the early or the late starters.

If we are to draw valid conclusions from these findings about the medium- and long-term effectiveness of early English learning, two conditions must be met. First, the samples of early and late starters must be equivalent. In other words, the switch to an early start must be independent of student (and family) characteristics. Second, the instruction received by both early and late starters at secondary level must be developmentally appropriate, taking into account prior knowledge and proceeding at an appropriate pace.

Pfenninger and Singleton assumed that the equivalence condition was met in their study, without demonstrating this to be the case by reference to achievement-related sample parameters (Pfenninger & Singleton, 2016, p. 316; 2017, pp. 25–26). There are reasons to question this assumption, however. Pfenninger (2016, 2017) mentioned substantial differences in written German (L1) proficiency at the first point of measurement, amounting to between $d = 0.50$ and $d = 1.0$ in favor of late starters depending on the language dimension.[4] Pfenninger and Singleton's (2017) interpretation was that the early starters lagged behind in their L1 German proficiency because they had learned to read and write in both German and English at the same time. This is not the only possible interpretation, however, and it is inconsistent with the bulk of findings about simultaneous alphabetization (i.e., literacy learning), which found that proficiency in the L1 was not significantly disadvantaged by learning literacy in another language at the same time (for Switzerland: Haenni Hoti & Werlen, 2007; Haenni Hoti, Müller, Heinzmann, Wicki, & Werlen, 2009; Heinzmann et al., 2009; for Germany: Baumert et al., 2017; Gebauer, Zaunbauer, & Möller, 2012; for the Netherlands: Goorhuis-Brouwer & de Bot, 2010). These other findings speak in favor of an alternative interpretation of the lack of parity between early and late starters at the first measurement point in the BAE study, namely, selective intake to the control group of late starters. For example, when elementary schools can decide on the timing of introducing an early start in English (often in consultation with parents and local authorities), more conservative and/or performance-oriented schools might hesitate to introduce English due to concerns about negative effects on students' achievement in other subjects, such as German and French. The differences observed by Pfenninger and Singleton may therefore reflect achievement-influenced (conflated) selection of late-starter students from schools that had taken longer to introduce early-start English—a difference that would need to be controlled for in all analyses.[5]

Doubts are also warranted as to whether the second condition was met, that is, whether instruction was properly tailored to students' knowledge level. First, recalling the results of Heinzmann et al. (2009), it can be assumed that the highly selective sample of *Gymnasium* students in the BAE study reached at least CEFR level A2.2 for English at the beginning of Year 7, with a notable proportion already reaching level B1.1. In other words, to be tailored to students' prior knowledge, the instruction provided for early and late starters would have to differ fundamentally. In the BAE schools, early and late starters were taught English in separate classes, but by the same teachers, following the same curriculum, using the same textbook, and with the same targets. Pfenninger and Singleton (2017) also drew attention to this problem. The processes of convergence observed in the descriptive data were therefore perhaps preprogrammed.

Against this background, we conclude that the findings of Pfenninger and Singleton (2017) do not—contrary to what the authors suggested—provide compelling evidence that elementary-level English instruction fails to achieve its aims. Allocation of students to groups must be properly controlled in order to accurately gauge the size of any performance gaps observed between early and late starters.

Individual and Institutional Factors Moderating the Impact of Early-Start Programs

Various individual and institutional factors have been proposed as moderating the relationship between AO and language proficiency. These discussions are generally driven by concerns that early-start language programs can, under certain conditions, impair development in other subjects, especially the language of instruction (usually the dominant language in society). It has, for example, been suggested that learning an additional language might overstretch less cognitively able students (see Haenni Hoti et al., 2009), or disadvantage students whose language development in the L1 is delayed or whose L1 is not the majority language. These assumptions are based on the finding that a sound command of the dominant language is an important asset for learning a L2, even when the L2 is taught in the target language (Sparks, Patton, Ganschow, Humbach, & Javorsky, 2008).

Testing the hypothesis that weaker-performing students are overstretched by early-start programs would require large, unselected samples of early and late starters, such as that used in the longitudinal Swiss study by Haenni Hoti et al. (2009). Their findings showed that students needing remedial support had weaknesses in all subjects examined, and felt particularly out of their depth at the end of Year 5. However, whether they started to learn English early or late made no difference to these students' (weak) reading proficiency in German (p. 23 and p. 25). To the best of our knowledge, no data from nonselective samples are available that would allow us to examine the extent to which the level of a student's command of the language of instruction (whether their L1 is or is not the majority language) may constitute an advantage or a risk when they begin to learn a new language in an early-start program. Pfenninger (2016) reported a strong positive correlation between L1 German and L2 English scores for Swiss *Gymnasium* students, but no interactions were observed between AO and level of German achievement.

Bilingualism is sometimes regarded as a resource for learning a L3, although findings are mixed (see reviews by Dyssegaard, de Hemmer Egeberg, Sommersel, Steenberg, & Vestergaard, 2015, and Fleckenstein, Möller, & Baumert, 2018). In particular, biliteracy—meaning that a critical competence threshold has been crossed in both languages—seems to have positive effects on L3 learning (Fleckenstein et al., 2018; Sanz, 2008). But to what extent can bilingualism serve as a resource for students from immigrant families who learn to read and write in their L2 and in a L3 simultaneously, or might bilingualism negatively impact the development of one or both new languages being learnt in school? In their random sample from Switzerland, Haenni Hoti and Werlen (2007) found that an early start in English as a foreign language (Year 3; AO: 9 years) did not have any negative effects on the proficiency in German (the language of instruction) of students with a L1 other than Swiss German after 1 year of instruction. Moreover, Haenni Hoti et al. (2009) found no difference in the English proficiency of bilinguals and that of students who grew up speaking Swiss German only. Similar results have been reported for academically selective samples (Pfenninger & Singleton 2019; Wilden & Porsch, 2016).

Summary of Research Findings and Need for Further Research

Despite considerable progress in research, many questions on the impact of early-start programs remain unanswered, due to several (often unavoidable) design problems with previous studies. Studies of long-term effects comparing the English proficiency of college students of the same age but different AO have been unsuitable for evaluating the long-term impact of early-start programs, as there were too many uncontrolled differences during the secondary school years. The findings of studies investigating language proficiency at secondary level as a function of AO are contradictory. Most of these studies (the only exceptions being Genelot, 1997, and Jaekel et al., 2017) do not meet the basic requirement of controlling for potentially conflatory factors that might have determined assignment to the late- or early-starter conditions. Most studies (the only exception being Burstall et al., 1974) were based on convenience and/or highly selective samples, meaning that it remains unclear to which population it is possible to generalize findings; and *none* of the studies established the extent to which secondary-level instruction recognized and built on the knowledge of early-start students—a necessary condition for conclusions to be drawn about the value of elementary-level English.

At the same time, findings from studies at elementary level are largely consistent in showing that most students in early-start programs achieve the curriculum attainment targets of A1 (CEFR) after 2 years and A2 (CEFR) after 4 years of instruction, even under low-input conditions. Thus, there is no empirical reason to question the effectiveness of early-start programs per se in the short term. However, findings are also consistent in showing that late starters learn at a faster pace, although these higher learning rates do not always entirely offset the exposure advantage of early starters, in the short or middle term. Nevertheless, full "catch-up" effects have been reported in the longer term, with both early and late starters closing initial gaps—findings that may suggest leveling-off effects that are potentially caused by a lack of adaptivity in secondary language instruction. However, as noted, no previous study on long-term effects has explicitly tested whether or not secondary-level instruction recognizes and builds on the prior knowledge of early-start students. Likewise, there has been little investigation of the extent to which long-term effects of early-start programs are moderated by individual and institutional factors.

It would thus be premature to conclude that "more or less everything important has been clarified" (Singleton & Pfenninger, 2019, p. 38). What is needed are studies that draw on nonselective random samples representative of larger political units, that carefully control for selective assignment to treatment (early-starter) and control (late-starter) groups, that address the problem that secondary-level language teaching may not be sufficiently adaptive, and that test for moderating effects of individual and institutional factors. The present study has gone some way toward closing these gaps.

The Present Study

This study examined the long-term effects of early English instruction on receptive language proficiency at the end of full-time compulsory schooling in Germany. Two sets of findings from the review above determined key elements of our study.

First, given the findings reported above, it seems reasonable to assume that most students in early-start programs receiving at least two English lessons per week will reach the equivalent of level A1.1 to A1.2 of the CEFR in key dimensions of language proficiency after 2 years of instruction, and potentially level A2.1 to A2.2 after 4 years. Importantly, however, language development does not seem to be linear. Gradual, rudimentary foreign language learning at preschool level (De Bot, 2014) is followed by an accelerating learning curve at elementary level that seems to become even steeper at secondary level (Graham et al., 2017) before flattening off in the longer term. It thus seems appropriate to use a logistic function as a model of second language development. The parameters of the function can be varied depending on learning opportunities, individual characteristics, and their interaction (Van Geert, 1991).

A second robust finding highlighted in the review above is that second language learners show age-dependent rates of learning at school: Older beginners initially learn faster and thus start with a learning rate advantage. Using a logistic function as a formal model of development, one would expect to observe different slopes of developmental curves as a function of AO (see Appendix S2, Figure S2.1, in the Supporting Information online), with the initial gap between the curves of early and late starters quickly becoming smaller, and the curves beginning to run parallel (perhaps once the late starters have reached the level that the early starters reached at elementary school). Under the assumption that instruction is developmentally appropriate and takes into account the prior knowledge of each group, we conceptualize the overall impact of an early start as a function of the amount of exposure, weighted by the age-dependent rate of learning. In the following, we formulate our specific research questions and hypotheses.

The Core Hypothesis and Research Aim: Testing the Adaptivity of Foreign Language Instruction at Secondary Level

Our core hypothesis was that early-start English has a positive effect on proficiency at the end of compulsory schooling, and that the magnitude of this effect depends on the amount of exposure, weighted by the age-dependent learning rate. This hypothesis would not find support under two conditions. First, if early starters' proficiency level at the end of elementary schooling is so low that it precludes cumulative learning processes at secondary level, then late starters may soon catch up with them. Second, if secondary-level English instruction does not build on what has been learned at elementary school, but starts again from the beginning, early-start students will be systematically underchallenged. In the former case, elementary-level English instruction is ineffective; in the latter, secondary-level English instruction is maladaptive.

The tracked secondary school system in Germany offers a prime setting for a natural experiment investigating the (mal)adaptivity of secondary-level foreign language instruction. Despite ongoing reform processes, students in Germany are still allocated to different secondary school tracks depending on their achievement. The number of tracks differs across states (see Appendix S2, Table S2.1, in the Supporting Information online), as does the age of transfer (usually 10 or 11 years). Findings suggest that early-start students in Germany achieve a modal level of proficiency corresponding to at least level A1 of the CEFR after 2 years of instruction. At the same time, the mean English proficiency of students allocated to the lowest versus highest track (*Hauptschule* vs. *Gymnasium*) is known to differ by up to $d = 2.0$ at the end of elementary school (Lehmann & Lenkeit, 2008, who provided their data to the authors of the present study for reanalysis). In other words, *Hauptschule* classes on average perform well below A1 level, thus failing to achieve the curriculum targets for elementary language teaching, whereas *Gymnasium* classes have on average already reached A2 level.

For secondary instruction to be adaptive, *Hauptschule* teachers must therefore repeat content covered at elementary level (some of it from the very beginning), provide ample opportunities for practice and repetition, and teach at a slow pace, which would also allow late starters to catch up with early starters relatively quickly due to their faster rates of learning. At *Gymnasium*, on the other hand, where English proficiency levels are much higher (although there may be gaps in linguistic accuracy and morphosyntax), the pace of instruction can be faster from the outset. Late starters in this context may achieve elementary attainment targets in a shorter time than early starters did, but they will not be able to catch up with early starters fully as the early starters will be receiving instruction that builds on their prior knowledge. Technically speaking, if secondary school instruction is adaptive, we can expect to observe an interaction effect between school type and AO on long-term proficiency. Results showing no such interaction effect and, at the same time, showing no main effect of an early start would be a strong indication that English teaching at secondary level was not sufficiently adaptive.

Supplementary Research Questions

Our analyses additionally addressed three further research questions that emerged from the review above.

1. It remains unclear to what extent a sound command of the language of instruction may be associated with early foreign language learning. In this study, we drew on a nonselective sample to address this question.
2. Under certain conditions, bilingualism can be a valuable resource for learning a L3. However, it is possible that learning to read and write in a L2 and a L3 simultaneously may have a negative impact, especially if the students have not yet attained balanced bilingualism in their L1 and L2. In this study, we compared the long-term English proficiency of bilingual early starters who learned to read and write in two languages simultaneously versus consecutively.
3. Early-start programs tend to be implemented partially, that is, introduced in some primary schools but not in others. This situation has the consequence that early starters are mixed with late starters at the move to secondary level. This poses particular challenges for providing sufficiently adaptive instruction, both when early and late starters are taught in the same class and when they are taught in different classes but following the same curriculum. No previous research has investigated the effects of partial implementation of early-start English at elementary level. This study aimed to address this gap. In regions of Germany (states) where there was only partial implementation of an early start, it was expected that secondary instruction would be less likely to be adaptive. In states where there was fuller implementation, it was expected that providing adaptive secondary instruction would be more likely. Thus, we expected an interaction effect between the degree of implementation of early instruction (i.e., in a particular state) and AO on outcomes at the end of Year 9 (age 15 years).

Method

Participants

This study drew on data from the BISTA assessment of the national educational standards (full title: Überprüfung des Erreichens von **Bi**ldungs**sta**ndards), which is conducted throughout Germany every 3 years in May or June of Year 9 with students aged 15 to 16 years (KMK, 2015). Specifically, we used data from the 2008–2009 academic year (Köller, Knigge, & Tesch, 2010). These data became available for wider use in 2018 (see https://www.iqb.hu-berlin.de/fdz/studies/IQB-LV_2008-9). The 2008–2009 student cohort was one of the last in which elementary-level English had not yet been implemented in all German states: 68% of the cohort had English lessons at elementary level, 90% of them starting in Year 3 or 4 (AO: 8–9 years) and 10% in Year 1 or 2 (AO: 6–7 years). The remaining 32% started to learn English in Year 5 (AO: 10 years), after moving to secondary school. The modal starting point for English was Year 3.

A two-stage sampling plan was used. First, a disproportionate stratified random sample of schools was drawn by state and school type; second, one intact class was drawn at random per school. The study sample for English comprises 31,426 students in 1,431 schools/classes. The coverage rate at student level was 94.9% (see Köller et al., 2010).

To test the effects of starting English at different points of the school career, we needed to specify a more restricted study population. Specifically, we used data from the student questionnaire to identify Year 9 students who learned English as their first foreign language, starting in Year 5 at the latest and continuing until Year 9, who did not speak English at home, had not participated in a bilingual education program, and had not spent more than 2 months in an English-speaking country. Shorter stays abroad were controlled for in the analyses. The necessary data were available for 24,741 students, of whom 19,653 were identified as belonging to the defined study population.

Study Instruments

Achievement Tests

Reading and listening comprehension tests developed on the basis of the national educational standards tested students' receptive language proficiency in German and in English as the first foreign language. The tests have not been released to the public because they are used for long-term monitoring purposes (see Stanat, Böhme, Schipolowski, & Haag, 2016). However, details of test development and sample items, including information on proficiency levels, are available online (https://www.iqb.hu-berlin.de/bista/ksm). The BISTA study had a multimatrix design, with about 240 items per domain (i.e., 240 for listening comprehension, 240 for reading comprehension). The items were calibrated in a multidimensional one-parameter item response theory model using the software Conquest 4.0 (Adams, Wu, & Wilson, 2015). The difficulty of items testing English ranged from CEFR level A1 to C1+. Standards were set with reference to the CEFR using the bookmark method, a standard-setting technique that takes empirical item difficulty into account (see Harsch, Pant, & Köller, 2010). Five plausible values were estimated for each student on the basis of a broad background model. The reliabilities of these plausible values estimates ranged from Cronbach $\alpha = .92$ to $\alpha = .93$ for English and from $\alpha = .81$ to $\alpha = .83$ for German (for more details, see Appendix S3, Section S3.1, in the Supporting Information online).

Definition of the Treatment Groups With Different AO

Learning history data assessed by means of a student questionnaire were used to define the study population and treatment groups. We distinguished three treatment groups depending on when students started to learn English: The early starters began in Year 1 or 2 of elementary school (AO: 6–7 years); the middle group, in Year 3 or 4 (AO: 8–9 years); and the late starters, in Year 5 of secondary school (AO: 10 years). Years 1 and 2 were aggregated, as were Years 3 and 4, to ensure that there were sufficient student numbers in all groups. We used a categorical grouping variable rather than treating AO as a continuous predictor because this approach made it possible to detect nonlinear relationships and allowed us to test for specific mean differences and interaction effects and interpret these according to the age categories that are used in the school system itself. English was compulsory for all students until the end of Year 9.

Individual Characteristics and Family Background

Data on students' gender and date of birth were obtained from the school records. All other person-specific data were reported in the student questionnaire. Family socioeconomic status was classified according to the International Socio-economic Index of Occupational Status (ISEI; Ganzeboom, de Graaf, & Treiman, 1992). In cases where the parents' ISEI scores differed, we used the higher value (HISEI). The family's educational level was defined as what the students reported the parents' highest educational qualification to be, and coded according to the International Standard Classification of Education (UNESCO, 2006). In addition, indices were calculated for the family's cultural and economic capital and digital information behavior. Cultural capital was assessed by six items tapping possession of cultural goods ($\alpha = .73$); economic capital, by five items tapping material possessions ($\alpha = .69$); and digital information behavior, by three items tapping internet use ($\alpha = .62$).

Students' immigration background was operationalized using the OECD definition, based on the parents' country of birth (0 = both parents born in Germany, 1 = one parent born abroad, 2 = both parents born abroad). Students were defined as simultaneous bilinguals if they came from families in which German and another language were spoken by their parents or daily carers, and they had learned both languages before starting kindergarten or school. They were defined as sequential bilinguals if they had learned German at kindergarten or school, beginning at age 6 years or later. In Germany, bilingual students in the cohort investigated are practically always from immigrant families, but children from immigrant families are not necessarily bilingual. Immigration background and bilingualism are thus confounded (for further information, see Appendix S3, Section S3.2, in the Supporting Information online). Two categories of stays in an English-speaking country were assessed: stays lasting up to 2 weeks and stays between 2 and 8 weeks. Students who had spent more than 2 months in an English-speaking country were excluded from the study sample. A 14-item scale tapped leisure-time use of English in the last 6 months (see Appendix S3, Table S3.1, in the Supporting Information online) as an indicator of interest in English, with a scale reliability of $\alpha = .87$.

Institutional Characteristics and Estimating the Amount of Exposure to Instruction

We examined covariates at the institutional level as follows: school type; federal state; minimum volume of English instruction stipulated by the federal state; degree of implementation of early-start English within the states (full or partial); and provision of additional English classes—enrichment classes for high-performing students (excluding Content and Language Integrated Learning classes) or remedial classes—as reported by the students. Four secondary school types were distinguished: *Hauptschule*, *Realschule*, multitrack schools, and *Gymnasium* (see Appendix S2, Table S2.1, in the Supporting Information online). Federal states (16) were dummy coded. The minimum number of hours of English instruction to be delivered by Year 9 was calculated by reference to the minimum number of lessons per week stipulated by the KMK (2014) (which differs by school year and by track, i.e., type of school), assuming an average of 35 teaching weeks per year. For Year 1 of elementary school, where early-start English does not begin until the second semester, we assumed 17.5 teaching weeks. The degree of implementation of early-start English was dummy coded at the state level. We classified the reforms as fully implemented if over 90% of students in that state reported having had English lessons in elementary school.

Missing Values and Multilevel Structure

Overall, 21.3% of the total of 31,426 BISTA participants (i.e., 6,685 participants) did not complete the student questionnaire. To test the generalizability of findings from the reduced sample, we compared the English and German test scores of those who had completed the questionnaire with the scores of those who had not. The mean differences in the latent achievement scores corresponded to an effect size of Cohen's $d = 0.02$ (95% CI [−0.16, 0.12]) for English (in favor of questionnaire nonparticipants) and Cohen's $d = 0.03$ (95% CI [−0.10, 0.16]) for German (in favor of questionnaire participants). Neither difference was statistically significant ($t = 0.26$, $p = .79$; $t = 0.46$, $p = .65$, respectively) and the effect sizes were minimal and their CIs passed through zero. It can thus be assumed that the missing data were random with respect to test scores.

Missing values in background variables due to partial item nonresponse were imputed using NORM. Cases that could not be assigned to a treatment group (early, middle, or late starters) (4.4% of the final dataset of $N = 19{,}653$) were excluded from the analyses. Due to BISTA's two-stage sampling plan, individuals were nested within schools. Depending on the intraclass correlation, this approach can lead to underestimation of conventional standard errors. Correct standard errors can be obtained either by explicitly modeling the multilevel structure (e.g., in mixed-effects models) or by estimating robust standard errors. If—as is the case here—no explicit hypotheses are tested on the aggregate level, estimating robust standard errors is the method of choice (e.g., using the sandwich estimator implemented in M*plus*; Muthén & Muthén, 2012). We therefore report robust standard errors throughout.

Statistical Analyses

The multivariate analyses were based on stepwise linear regression analyses in which the latent total score for English was modeled as the dependent variable. This latent total score explained 93% of the variance in the two indicators (English reading and listening comprehension at the end of Year 9). Separate analyses for reading and listening comprehension did not reveal any substantial differences in terms of the patterns of findings. To keep the analyses as parsimonious as possible, we also estimated a latent total score for German. Five plausible values were generated for each student's English and German scores. All analyses were repeated five times, and the results were integrated using

Rubin's rules (Rubin, 1987). The α-level was corrected for multiple comparisons using the Bonferroni procedure. A total of five regression models (four of which are shown in Table 2) were fitted using a maximum likelihood estimator with robust standard errors. All analyses were conducted with weighted data to ensure national representativity. Weights were based on register data about students and schools obtained from the statistical offices of the German federal states.

Results

Descriptive Results

Table 1 summarizes the main characteristics of the study sample by treatment group. The three treatment groups were large, with 1,220 early starters, 12,173 students in the middle group, and 5,392 late starters (for further sample information, see Appendix S4, Table S4.1, in the Supporting Information online). The first block of Table 1 reports the minimum amount of English instruction stipulated by the federal states. On average, a total of at least 540 hours of English was required by the end of Year 9 (age 15–16 years): 638 for early starters, 561 for the middle group, and 471 for late starters.

The next block of Table 1 reports means and standard deviations of the raw scores for English reading and listening comprehension and the distribution of students (as percentages) across the CEFR proficiency levels in Year 9 (age 15–16 years). Contrary to expectations, the descriptive findings indicate that the middle group achieved the best results and the early starters, the poorest. The modal proficiency level for listening comprehension was B1 in all three treatment groups; for reading comprehension it was either A2 or B1. Similar differences across the treatment groups were observed for German (see Table 1), although the groups did not differ in the number of German lessons received. These findings strongly suggest that assignment to groups was not random.

The third block of Table 1 reports interest in English, as reflected by scores on the indicators of leisure-time use of English. The figures indicate that early starters used more English in their free time than late starters did, $p < .001$, $d = 0.18$, 95% CI [0.12, 0.24].

The fourth block of Table 1 presents biographical data. The three treatment groups differed little in terms of key individual and background variables. The proportion of students with an immigration background was about 5 percentage points lower in the middle group than in the other two groups ($p < .001$ in both cases). Because almost all bilingual students in the cohort investigated came from immigrant families (see Methods), the proportion of

Table 1 Parameters of the study population by onset of English instruction (weighted data)

	Early starters (*n* = 1,220)		Middle group (*n* = 12,173)		Late starters (*n* = 5,392)		Total (*N* = 18,785)	
Variables	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*
Volume of instruction								
Total hours (60 minutes) of English (min. volume)	637.96	22.0	560.54	31.06	470.73	18.31	539.92	55.07
Proficiency scores								
English: Reading comprehension	486.43 (5.25)	100.39	507.63 (3.76)	96.02	497.15 (4.98)	92.66	503.23 (3.51)	95.59
A1 CEFR (%)	19.3 (1.80)		13.5 (1.00)		14.9 (1.30)		15.1 (0.90)	
A2 CEFR (%)	36.4 (2.20)		32.6 (1.30)		35.7 (2.10)		33.8 (1.20)	
B1 CEFR (%)	31.8 (2.90)		36.3 (1.10)		35.6 (2.00)		35.3 (1.00)	
B2 CEFR (%)	11.1 (1.70)		15.9 (1.10)		12.6 (1.50)		14.3 (1.00)	
C1 CEFR (%)	1.4 (0.70)		1.7 (0.30)		1.1 (0.40)		1.4 (0.20)	
English: Listening comprehension	482.66 (5.41)	98.15	493.06 (3.89)	95.32	494.00 (5.39)	84.47	498.43 (3.64)	95.43
A1 CEFR (%)	8.6 (1.40)		5.7 (0.50)		6.5 (0.80)		6.7 (0.50)	
A2 CEFR (%)	39.8 (2.30)		34.6 (1.50)		36.5 (2.30)		35.9 (1.40)	
B1 CEFR (%)	41.8 (2.30)		46.0 (1.40)		45.3 (2.10)		44.9 (1.30)	
B2 CEFR (%)	9.3 (1.50)		12.8 (1.10)		11.0 (1.70)		11.8 (1.00)	
C1 CEFR (%)	0.2 (0.20)		0.3 (0.10)		0.2 (0.10)		0.2 (0.10)	
German: Reading comprehension	490.41 (4.90)	90.50	505.58 (3.27)	88.32	497.40 (4.76)	87.26	502.23 (3.05)	88.30
German: Listening comprehension	488.24 (5.12)	98.84	508.19 (3.42)	95.20	499.56 (5.24)	96.02	504.40 (3.27)	95.86
German: Spelling	488.14 (5.28)	99.94	511.42 (3.63)	95.14	496.13 (5.54)	96.58	505.50 (3.49)	96.23

(*Continued*)

Table 1 Continued

	Early starters (n = 1,220)		Middle group (n = 12,173)		Late starters (n = 5,392)		Total (N = 18,785)	
Variables	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*
Interest in English								
Leisure-time use of English (1–5)	2.43 (0.04)	0.63	2.30 (0.02)	0.88	2.27 (0.02)	0.81	2.30 (0.01)	0.82
Biographical characteristics								
Age	15.21 (0.03)	0.68	15.16 (0.01)	0.65	15.24 (0.3)	0.74	15.19 (0.02)	0.68
Gender (male) (%)	45.60 (1.90)		48.70 (1.10)		53.00 (1.50)		49.80 (0.90)	
Parents' SES (HISEI)	47.57 (0.73)	16.47	48.53 (0.43)	15.48	47.78 (0.63)	15.68	48.25 (0.40)	15.61
Parents' education (HISCED)	3.95 (0.06)	1.50	3.93 (0.04)	1.49	3.89 (0.05)	1.49	3.92 (0.03)	1.49
Cultural capital of family	0.04 (0.03)	0.63	0.01 (0.02)	0.65	−0.02 (0.02)	0.65	0.00 (0.02)	0.65
Material possessions of family	0.02 (0.03)	0.71	0.07 (0.02)	0.67	0.02 (0.02)	0.68	0.05 (0.01)	0.67
Internet use in household	0.00 (0.02)	0.74	0.04 (0.01)	0.67	0.03 (0.02)	0.70	0.04 (0.01)	0.68
Immigration background BP or OP (%)	26.7 (2.00)		21.7 (1.00)		26.9 (1.80)		23.8 (0.90)	
Simultaneous bilingual BP (%)	11.9 (1.40)		10.5 (0.60)		12.3 (1.10)		11.3 (0.60)	
Sequential bilingual BP (%)	4.0 (0.20)		2.3 (0.20)		4.3 (0.60)		3.1 (0.30)	
Simultaneous bilingual OP (%)	9.5 (1.10)		6.9 (0.40)		8.0 (0.70)		7.4 (0.30)	
Stayed abroad 1–2 weeks ESC (%)	37.6 (1.9)		35.5 (0.80)		32.5 (1.4)		34.4 (0.60)	
Stayed abroad 3–8 weeks ESC (%)	10.6 (1.0)		8.9 (0.40)		7.9 (0.60)		8.6 (0.30)	

(*Continued*)

Table 1 Continued

	Early starters ($n = 1{,}220$)		Middle group ($n = 12{,}173$)		Late starters ($n = 5{,}392$)		Total ($N = 18{,}785$)	
Variables	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*	*M* or % (*SE*)	*SD*
Institutional characteristics								
English enrichment classes (%)	3.30 (0.80)		3.60 (0.40)		6.30 (1.20)		4.40 (0.50)	
English remedial classes (%)	1.90 (0.50)		2.90 (0.40)		3.80 (0.80)		3.10 (0.40)	
Hauptschule (%)	21.10 (2.40)		18.20 (1.90)		23.10 (3.20)		19.80 (1.90)	
Realschule (%)	19.00 (2.50)		23.80 (2.40)		27.40 (3.70)		24.50 (2.30)	
Gymnasium (%)	30.50 (2.80)		36.90 (2.60)		30.50 (4.00)		34.60 (2.50)	
Multitrack schools (%)	29.40 (2.20)		21.10 (1.40)		19.10 (2.10)		21.00 (1.40)	
Federal state (range in %)	1.1–13.3		6.9–88.7		2.2–92.0		–	
Total	6.60 (0.30)		64.7 (1.50)		28.6 (1.50)		100.0	

Note. We report either the mean value or the percentage, depending on the variable. English enrichment classes exclude Content and Language Integrated Learning classes. CEFR = Common European Framework of Reference for Languages; SES = socioeconomic status; HISEI = higher of the two parents' values on the International Socio-economic Index of Occupational Status; HISCED = parents' highest educational qualification in the International Standard Classification of Education; BP = both parents born abroad; OP = one parent born abroad; ESC = in an English-speaking country.

bilingual students was also lower in the middle group than in the other two groups (for information on parents' country of origin, see Appendix S3, Section S3.3, in the Supporting Information online). Differences between groups in terms of bilingualism and immigrant background were statistically significant ($p < .001$). There were noticeable group differences in the gender ratio; according to the descriptive results, boys seemed to start learning English somewhat later than girls. These differences in terms of gender between late starters and the other two groups were statistically significant ($p < .001$). Finally, as might be expected, early starters were somewhat more likely to have spent time in an English-speaking country than late starters. Group differences were statistically significant ($p < .001$).

The fifth block reports institutional characteristics. In all three treatment groups, participation in additional English classes was rare. This applies to both enrichment classes for high-performing students (excluding Content and Language Integrated Learning classes) and remedial support for low-performing students. There were, however, clear group differences in distribution to the secondary tracks. The proportion of academic-track students was highest in the middle group, at 37% ($p < .001$ for comparisons with both other groups), and that of students in multitrack schools was highest in the early-starters group, at 29% ($p < .001$ for comparisons with both other groups). Finally, the federal states differed considerably, not only in terms of the implementation of elementary-level English (as indicated by the ranges reported in Table 1), but also in the mean English proficiency level achieved by Year 9 (not reported in Table 1; see Köller et al., 2010).

Multivariate Analyses

As the descriptive findings have shown, students were evidently not allocated to the three treatment groups at random. Rather, group membership was confounded with individual achievement and/or institutional context. In multivariate analyses, we controlled stepwise for key variables that can be expected to covary with allocation to treatment group and English proficiency. The dependent variable in these analyses was the latent total score for English receptive proficiency at the end of Year 9 (i.e., listening and reading combined in a latent variable with two indicators). The standard deviation of the latent dependent variable in the study sample was 89.7 points.

The starting point for the multivariate analyses was the nonadjusted effects of starting English at elementary relative to secondary level (not presented in Table 2). The mean proficiency level of early starters (Year 1 or 2) seemed slightly lower than that of late starters, although this was not statistically

Table 2 Linear regression of English scores (latent total score) at the end of Year 9 on the onset of English instruction and individual and institutional characteristics, showing unstandardized regression coefficients (b), robust standard errors (SE), p values, and R^2

	Model 1			Model 2			Model 3			Model 4		
Predictors	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Age of onset (AO; reference: Year 5, late, AO 10 years)												
Year 1 or 2 (early, AO 6–7 years)	−10.27	5.0	.04	−2.85	3.70	.44	−2.48	6.88	.72	−7.13	8.13	.38
Year 3 or 4 (middle, AO 8–9 years)	3.63	3.77	.34	0.37	2.37	.88	−1.04	4.57	.82	−2.02	4.88	.68
Individual characteristics												
Age (*z*)	−16.90	1.40	<.001	−5.02	1.02	<.001	−3.46	0.90	<.001	−3.33	0.90	<.001
Gender (male)	−5.29	2.60	.04	3.03	1.76	.09	1.43	1.57	.36	1.79	1.58	.26
Parents’ SES (HISEI; *z*)	16.14	1.20	<.001	4.82	0.92	<.001	1.17	0.87	.18	1.67	0.88	.06
Parents’ education (HISCED; *z*)	10.77	1.16	<.001	3.18	0.97	.001	1.55	0.83	.06	1.31	0.84	.11
Cultural capital of family (*z*)	19.33	1.67	<.001	3.17	0.92	.001	1.31	0.86	.12	1.58	0.85	.06
Material possessions of family (*z*)	−5.56	1.19	<.001	−2.06	0.72	.01	−2.37	0.68	.001	−2.34	0.72	.001
Internet use in household (*z*)	4.32	1.01	<.001	1.90	0.78	.01	0.77	0.70	.27	1.23	0.69	.08
Simultaneous bilingual BP	−7.72	5.06	.13	16.25	3.57	<.001	8.37	3.59	<.001	10.48	3.50	.003
Sequential bilingual BP	−16.16	7.17	.02	15.20	5.84	.01	6.74	5.16	.19	9.14	5.13	.08
Simultaneous bilingual OP	−5.69	5.08	.26	9.01	4.16	.03	8.15	3.74	.02	9.11	3.76	.01
Interaction bilingual ∗ early	−12.40	11.21	.27	−13.00	8.49	.13	−16.56	7.65	.03	−14.56	7.60	.05
Interaction bilingual ∗ middle	1.13	5.11	.87	0.88	3.81	.82	0.22	3.63	.95	0.28	3.52	.94
Stayed abroad 1–2 weeks ESC	2.39	2.08	.25	1.57	1.66	.34	1.04	1.53	.50	0.95	1.53	.54
Stayed abroad 3–8 weeks ESC	4.25	3.27	.19	5.94	2.76	.03	5.59	2.51	.93	5.84	2.56	.02
German proficiency (GER; *z*)				62.11	2.22	<.001	43.94	2.21	<.001	44.06	2.19	<.001
Interaction GER ∗ early				4.65	2.92	.11	6.48	3.51	.06	7.94	3.57	.03
Interaction GER ∗ middle				8.49	1.97	<.001	7.19	2.37	.002	9.25	2.44	<.001
Interest in English (leisure-time use; *z*)				9.44	0.89	<.001	8.87	0.81	<.001	9.09	0.83	<.001

(*Continued*)

Table 2 Continued

	Model 1			Model 2			Model 3			Model 4		
Predictors	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Institutional characteristics												
Additional English classes												
Enrichment							7.69	3.95	.05	6.45	3.74	.09
Remedial							−19.23	4.73	<.001	−18.16	5.17	<.001
School type (ref.: Academic track *Gymnasium*)												
Hauptschule (HS)							−77.64	7.64	<.001	−71.62	7.03	<.001
Multitrack (MTS)							−62.16	5.72	<.001	−67.74	5.76	<.001
Realschule (RS)							−39.72	5.59	<.001	−35.41	5.20	<.001
Interaction HS ∗ early							7.70	11.85	.52	10.85	11.82	.36
Interaction MTS ∗ early							14.04	8.57	.10	16.70	8.68	.05
Interaction RS ∗ early							7.86	8.95	.38	9.62	9.09	.29
Interaction HS ∗ middle							1.93	7.72	.80	5.51	7.47	.46
Interaction MTS ∗ middle							8.06	6.09	.19	10.68	6.35	.09
Interaction RS ∗ middle							8.22	5.94	.17	9.76	5.71	.09
Implementation (IMP; fully imp.)										−20.65	5.65	<.001
Interaction IMP ∗ early										11.78	6.32	.06
Interaction IMP ∗ middle										8.19	5.40	.13
Federal state controlled	No			No			Yes			Partly[a]		
R^2	.28	0.01	<.001	.64	0 .01	<.001	.62	0.01	<.001	.61	0.01	<.001

Note. English enrichment classes exclude Content and Language Integrated Learning classes. SES = socioeconomic status; HISEI = higher of the two parents' values on the International Socio-economic Index of Occupational Status; HISCED = parents' highest educational qualification in the International Standard Classification of Education; BP = both parents born abroad; OP = one parent born abroad; ESC = in an English-speaking country; fully imp. = implementation of English instruction at elementary level.

[a] As the degree of implementation is confounded with federal state, IMP partly controls for federal state.

meaningful, $b = –11.00$ ($t = 1.86, p = .06$) and $d = -0.12$, 95% CI [−0.25, 0.01]. The middle group seemed to slightly outperform the late starters, although again this was not statistically meaningful, $b = 9.47$ ($t = 1.86, p = .06$) and $d = 0.11$, 95% CI [−0.01, 0.22]. Thus, the counterintuitive descriptive difference between early and late starters and the difference between the middle group and the late starters did not quite reach the alpha level of .05 and the 95% CIs around a very small *d* passed through zero. However, the difference between the early starters and the middle group was significant, with $b = 20.47$ ($t = 4.49, p < .001$) and $d = 0.23$, 95% CI [0.13, 0.33], in favor of the middle group.

In Model 1 of Table 2, stable individual and background characteristics were entered as covariates. All covariates predicted English proficiency in the expected direction, but associations with treatment group membership were inconsistent. The negative effect of an early start relative to the reference group of late starters was now significant, with $b = -10.27$ ($t = -2.02, p = .04$) and $d = -0.12$, 95% CI [−0.23, −0.00]; the positive effect of belonging to the middle group decreased yet remained significant relative to early starters, with $b = 13.90$ ($t = 3.24, p = .001$) and $d = 0.15$, 95% CI [0.01, 0.25]. Because school age was held constant (with all students being assessed in Year 9), chronological age is an indicator of a delayed school career, and was negatively related to English proficiency, with $b = -16.90$ ($t = -12.08, p < .001$). Boys had somewhat lower English scores than girls, with $b = –5.29$ ($t = -2.03$, $p = .04$). Likewise, family background variables made specific contributions to explaining English proficiency. Bilingual students—almost all of whom had an immigrant parent background—tended to lag behind their German monolingual peers in English, although the patterns of results were not consistently statistically significant, as follows. For students whose parents had *both* been born abroad, the coefficient was, for simultaneous bilinguals, $b = -7.72$ ($t = -1.53, p = .13$) with an effect size of $d = -0.09$, 95% CI [−0.34, 0.03], and, for sequential bilinguals, $b = -16.16$ ($t = -2.25, p = .02$) and $d = -0.18$, 95% CI [−0.34, −0.02]. A negative interaction between bilingualism and AO emerged for the early starters—that is, students who learned to read and write in the L2 and L3 simultaneously—but this interaction was not significant, with $b = -12.40$ ($t = 1.11, p = .27$). No effect was found for time spent in an English-speaking country. The total amount of variance in English proficiency explained in Model 1 was $R^2 = .28$ ($t = 22.55, p < .001$).

Model 2 additionally included the latent German proficiency score along with out-of-school interest in English as a motivational variable. With $b = 62.11$ ($t = 27.97, p < .001$) and $d = 0.77$, 95% CI [0.74, 0.80], the German score was by far the strongest individual predictor of English proficiency (as also found by Pfenninger, 2016, and Pfenninger & Singleton, 2017). The positive interaction between German proficiency and AO was not significant in the early starters, with $b = 4.65$ ($t = 1.59, p = .11$), but it was significant in the larger middle group, with $b = 8.49$ ($t = 4.31, p < .001$). Out-of-school interest in English also contributed to explaining the variance, with $b = 9.44$ ($t = 10.61, p < .001$) and $d = 0.12$, 95% CI [0.10, 0.14]. The amount of variance explained increased by 36 percentage points from Model 1 to Model 2, reaching $R^2 = .64$ ($t = 37.55, p < .001$).

In Model 2, the differences between the treatment groups disappeared almost entirely, and the weight of the other covariates changed as expected. Gender differences were no longer significant; the girls' lead in English was evidently mediated by L1 proficiency (see Courtney, Graham, Tonkyn, & Marinis, 2017, for similar findings). When we controlled for German proficiency, having stayed abroad for between 3 and 8 weeks became significant. Moreover, bilingual students no longer lagged behind in English proficiency; on the contrary, they now clearly outperformed their monolingual peers, with coefficients of bilingual subgroups ranging between $b = 9.01$ ($t = 2.17, p = .03$) and $b = 16.25$ ($t = 4.53, p < .001$), corresponding to effect sizes of between $d = 0.10$, 95% CI [0.01, 0.19], and $d = 0.18$, 95% CI [0.10, 0.26]. Note, however, that there may have been very tentative evidence that this advantage of bilingualism may not apply to early starters, given the negative although not significant interaction term between bilingualism and an early start of $b =$ –13.00 ($t = -1.53, p = .13$) that is included in the regression equation.

Model 3 additionally included institutional characteristics, such as secondary school type. *Gymnasium* served as the reference group for the dummy-coded school types. (The results for the 16 federal states, which were also dummy coded, are not listed separately in Table 2 due to space constraints; patterns of results as a function of state did not reveal any specific interpretable findings in the context of the present study.) All institutional characteristics made a statistically and practically significant contribution to explaining the variance in English proficiency. Although we controlled for important background variables and additionally for proficiency in German, and thus for selective entry to the different secondary tracks, track membership was still a very strong predictor of English proficiency. *Hauptschule* students lagged behind their peers in *Gymnasium* schools by $b = -77.64$ ($t = -10.16$, $p < .001$) points and $d = -.87$, 95% CI [−1.03, −0.70]. The figures comparing students from multitrack schools and students from *Realschule* with *Gymnasium* students were, respectively, $b = -62.16$ ($t = -10.87, p < .001$)

points with $d = -0.69$, 95% CI $[-0.82, -0.57]$, and $b = -39.72$ ($t = -7.10, p < .001$) points with $d = -0.44$, 95% CI $[-0.57, -0.32]$.

In testing the interaction between school type and AO, we examined the adaptivity of secondary-level language instruction. As noted above, the mean English proficiency of early starters transferring to *Hauptschule* vs. *Gymnasium* can differ by up to $d = 2.0$ at the end of elementary school. To provide instruction that is adaptive to this lower level of proficiency at the start of secondary school, *Hauptschule* teachers must repeat content covered at elementary level and teach at a slow pace, which is likely to allow late starters to catch up relatively quickly due to their faster rates of learning. At *Gymnasium*, the pace can be faster from the outset because proficiency levels tend to be much higher. Accordingly, if instruction is adaptive, we can expect interaction effects between school type and AO on long-term proficiency. In specifying the six interaction terms in Model 3 between school types, on the one hand, and early starters/middle group, on the other hand, we used *Gymnasium* as the reference group. We therefore expected negative interactions to emerge. However, none of the interactions between school type and AO were significant, and all coefficients were positive. If secondary schools do succeed in building on the previous knowledge of elementary students at all, this seemed to be very slightly, although not statistically meaningfully, more the case at multi-track schools and with early starters ($b = 14.04, t = 1.64, p = .10; d = 0.16$, 95% CI $[-0.03, 0.34]$) than at *Gymnasium*. According to this model, the main effect of elementary-level English instruction—whether early or later onset—was zero.

With control for institutional context (Model 3), the negative interaction between bilingualism and an early start observed in Models 1 and 2 became significant ($b = -16.56, t = -2.17, p = .03$). This finding indicates that learning a L2 (German as the language of instruction) and a L3 (English as a foreign language) *simultaneously*, after literacy in a L1 home language has been established, may indeed be a risk factor for proficiency in L3 receptive skills, even after controlling for proficiency in German (as in Models 2 and 3). This interpretation is consistent with the finding that an interaction was no longer present for the middle group ($b = 0.22, t = 0.06, p = .95$; consistent with the findings of Haenni Hoti et al., 2009). As individual and institutional characteristics were confounded, the overall explanatory power of Model 3 did not increase relative to Model 2; however, the locations of the effects could be pinpointed more accurately.

Unless all elementary schools in the area served by the same secondary school offer early-start English, early starters are often taught at secondary school in mixed classes with late starters or—in a better-case scenario—in parallel classes although often following the same curriculum. Under these conditions, cumulative effects can hardly be expected. Model 4 in Table 2 tested this assumption. We expected to see an interaction between degree of implementation (partial vs. full) and AO, and possibly a small positive main effect of full implementation due to routinization. This expectation rests on the assumption that initial difficulties in ensuring secondary schools adapt to cohorts arriving with an early start should be overcome by the time early-start programs have been fully implemented. That is, it would be easier for teaching in secondary schools served by elementary schools with full implementation to build on the knowledge gained at elementary level. In contrast, teaching in schools with a mixed student body would be more challenging and more likely to be maladaptive.

Contrary to our expectations, a very surprising main effect emerged from Model 4. In states where early-start English was implemented across the board, English scores were significantly lower than in states with slower and gradual implementation, with $b = -20.65$ ($t = -3.66$, $p < .001$) and $d = -0.28$, 95% CI [−0.42, −0.13]. In line with our expectations, there was some tentative indication for positive interaction effects between the degree of implementation and early start and middle start, respectively, but the coefficients were not statistically significant ($b_{\mathrm{IMP} \times \mathrm{early}} = 11.78$, $t = 1.86$, $p = .06$; $b_{\mathrm{IMP} \times \mathrm{middle}} = 8.19$, $t = 1.52$, $p = .13$), and the effect sizes were very small with 95% CIs that passed through zero ($d = 0.16$, 95% CI [−0.01, 0.33], and $d = 0.11$, 95% CI [−0.03, 0.25]).

Discussion

This study drew on a nationally representative sample of Year 9 students (age 15–16 years) in Germany to examine the effects of early-start English on receptive language skills at the end of compulsory schooling. We compared the reading and listening comprehension of early starters (English from Year 1 or 2; AO: 6−7 years), a middle group (English from Year 3 or 4; AO: 8−9 years), and late starters (English from Year 5; AO: 10 years). Without control for other variables, the proficiency of the early starters and the middle group differed only marginally from that of the late starters, but the early starters lagged significantly behind the middle group. Controlling for individual and family background characteristics had little effect on this pattern of results. Thus far, our results seemed to broadly align with one of the overarching findings of Jaekel et al. (2017): that an early start may potentially confer a disadvantage in some contexts. However, when we controlled for German proficiency and important institutional parameters such as school type and federal state, this impression was corrected: In general, the point at which students started to learn English—whether in Year 1/2 or Year 3/4 of elementary school or in Year 5 of secondary school—no longer made any overarching difference by the end of Year 9 (age 15–16).

In sum, this study found no evidence for the expected positive effects of early-start English. After 5 years of English at secondary level, the exposure advantage of students who learned English at elementary school was eroded. This result is consistent with the findings of Pfenninger and Singleton (2017), who—without controlling for relevant covariates—found that early starters' (A0: 8 years) exposure advantage (440 hours) was no longer detectable after 6 years of English at Swiss *Gymnasium* schools.

This pattern of results can emerge from one of two situations, or from a mix of both. In one situation, secondary teachers respond adaptively to new students' insufficient levels of English by increasing repetition and practice or by making a fresh start as they do for late starters. In the other situation, they respond maladaptively by not acknowledging and building on what students actually know and can do. In both cases—or with a combination of both—the proficiency levels of early and late starters can be expected to converge.

How well do English teachers succeed in bridging the gap between elementary and secondary school? Continuity of instruction at the transition to secondary level has been a hot topic in educational science and administration for some time now. From a range of countries, reports (e.g., Ofsted, 2011) and studies/reviews (e.g., Bolster, Balandier-Brown, & Rea-Dickens, 2004; Courtney, 2014; Galton, Gray, & Rudduck, 2003; Galton & McLellan, 2018; Muñoz, Tragant, & Camuñas, 2015; Richardson, 2014) underline three main problems at the elementary–secondary interface: secondary teachers systematically underestimating the knowledge and abilities of their new students; curricular objectives and learning content not being properly aligned between the elementary and secondary schools; and a mutual lack of acceptance of differences in teaching methods. Another relevant factor is that secondary school class grouping practices do not accommodate students' different levels of proficiency.

In this study, we tested whether secondary-level English instruction is well adapted to different levels of student achievement quasi-experimentally, by examining the interaction between AO and ability tracking, which produces learner groups with very different levels of English proficiency (mean differences up to $d = 2.0$) across the German secondary tracks. If instruction is adaptive, the long-term effects of early-start English at *Gymnasium* (where English scores are higher on entry) versus *Hauptschule* (where English scores are extremely low on entry) should differ systematically. The finding that none of the specified interactions between AO and school type were significant (Table 2, Model 3) is a strong indicator that secondary-level English instruction failed to respond adaptively to students' different proficiency levels at entry to the school. These problems seem to be most pronounced in *Gymnasium* schools (see the signs of the interactions in Table 2, Model 3 and Model 4). This result is consistent with the findings of Pfenninger and Singleton (2017), who found evidence for processes of convergence between early and late starters in Swiss *Gymnasium* schools.

Another finding also points to insufficient adaptivity of language instruction at secondary level, namely, that there was no significant interaction between AO and degree of implementation of early-start English. It is arguably more difficult to adapt secondary teaching to the existing knowledge of early starters in mixed classes where teachers also have to cater for complete beginners. This should be reflected in significant interactions between AO and degree of implementation in the schools feeding into the secondary schools. The lack of such interaction indicates that the English instruction delivered at secondary level was generally standardized independently of the make-up of the class (as also suggested by Bolster et al., 2004; Muñoz et al., 2015). In fact, the level of English attained at age 15–16 years was lower in federal states where early-start English was implemented across the board than in those where it was not. It is possible that states that introduced the reforms gradually and more cautiously paid more attention to maintaining achievement standards, including, perhaps, being aware of or being more prepared for the need for adaptivity in the secondary schools. Alternatively, or in combination with this explanation, there may simply not be enough qualified teachers to ensure high-quality instruction at elementary level. In this study, it was not possible to consider the proportion of elementary-level English lessons delivered by teachers of other subjects (i.e., non-specialist English teachers).

Another key finding of our study is that a large amount of the difference in English proficiency was explained by German proficiency (incremental $R^2 \approx .35$). The cross-sectional design of our study does not allow a causal interpretation of this result. However, the command of the language of school instruction is known to play an important role in both early immersion programs (Gebauer, Zaunbauer, & Möller, 2013) and regular schooling (Barr, Uccelli, & Phillips Galloway, 2019; Sparks et al., 2008; Ströbel, Kerz, & Wiechmann, 2020). A good command of the language of schooling may help learners benefit from elementary-level English lessons, as shown by the significant interaction effects between AO and German proficiency on English receptive skills that were measured many years later. This interaction was found both for early starters and for the middle group (Table 2, Models 2 and 3). This finding casts the results of Pfenninger (2016) and Pfenninger and Singleton (2017) in a new light: An interaction of German (L1) proficiency and AO probably failed to emerge in their highly selected group of Swiss *Gymnasium* students due to low variance in the English proficiency data and, potentially, small sample size.

Bilingual students, almost all of whom in the cohort studied came from immigrant families, lagged behind their peers on the English receptive tests in all treatment groups, even when we controlled for measures of social and cultural capital. This is a standard finding in Germany (Köller et al., 2010; Stanat et al., 2016). The gap seemed to be amplified in the early-start group: There was a significant negative interaction effect between growing up bilingually and an early start (AO: 6–7 years) in English ($d = -0.22$, 95% CI [−0.42, −0.02]; Table 2, Models 3 and 4) on English scores 9 years later. The negative interaction may indicate that learning German as the second L1 or as a L2 (the language of schooling) plus English as a foreign language simultaneously in Years 1 and 2 (AO: 6–7 years), after the establishment of one (or more) home language(s), can interfere with the development of the foreign language. Indeed, there was no interaction effect for students who started English 1 or 2 years later, in Year 3 or 4 (AO: 8–9 years), probably after German as the language of schooling had become more embedded.

However, when we additionally controlled for proficiency in German, bilingual learners' English proficiency was higher than that of students who spoke only German at home. That is, if bilinguals had strong German proficiency (as measured at age 15–16 years), this seemed to mitigate the observed negative effects of bilingualism combined with an early start in English as a foreign language. This finding could be interpreted as supporting Cummins's (1979) threshold hypothesis, in that a positive influence of an existing language repertoire (L1 and L2) may be hampered by a very early start in a foreign language (AO 6–7 years). Reaching a critical threshold in the (written) language of school instruction may benefit the learning of further languages (for further evidence see Fleckenstein et al., 2018; Rolstad & MacSwan, 2014; Sanz, 2008). Notably, the interaction effect between bilingualism and AO did not emerge when the analysis was repeated in just the subsample of *Gymnasium* students, which is consistent with Pfenninger and Singleton's (2019) findings for *Gymnasium* students in Switzerland. This underlines the importance of using nonselective samples when studying the effects of early-start programs.

Overall, these findings refute the assumption—which has long been questioned by researchers—that starting to learn a foreign language at elementary school is certain to have positive effects on students' proficiency levels at the end of secondary schooling. However, our findings do not refute the effectiveness of an early start per se. Rather, they suggest that the lack of long-term impact of an early start in English seems to be attributable to secondary-level language teaching being insufficiently adaptive to students' proficiency levels on arrival at secondary school.

This maladaptivity is due to a range of phenomena. Institutional objectives, curricula, school timetables, and class grouping arrangements in secondary schools have not generally been adapted to account for early-start students and their longer prior exposure times (see Hempel et al., 2018). In Germany, secondary school textbooks have only recently (around 2015) been revised, and most approved textbooks, even when revised, still refer to curricula that were issued between 2003 and 2007 and do not take into account an early start in English in elementary school (e.g., North Rhine-Westphalia: https://www.schulministerium.nrw.de/docs/Schulsystem/Medien/Lernmittel/index.html or Baden-Württemberg: https://www.ls-bw.de/,Lde/Startseite/Service/sbz3). Furthermore, the national education standards for English at secondary level to which all German states are committed have not been changed since they were first published in 2003 (KMK, 2004). Thus, language instruction still seems to follow routines that were established when all students started to learn English on entering secondary school (Bolster et al., 2004; Galton & McLellan, 2018; Keßler, 2006; Muñoz et al., 2015).

Limitations and Future Directions

Several limitations should be taken into account when interpreting these findings. First, as mentioned above, given the quasi-experimental design of our study, in which students of the same age but with different AO were tested at the same point in their school career, AO and amount of exposure were necessarily confounded. Our study thus could not answer the question as to what role biological age or specific developmental periods play in foreign language acquisition. Rather, our aim was to determine the long-term effects of early-start language programs in which exposure and age-dependent rates of learning interact, and our study design and control for an extensive set of possible confounds allowed a sound estimate of these effects.

The cross-sectional design of our study means that no conclusions can be drawn about the dynamics of early and late starters' language development as a function of the nature of the learning opportunities available. Addressing this question would require a longitudinal design in which the points of measurement are planned in a way that is sensitive to student development and is thus able to capture nonlinear developments. In addition, the tests would have to be scaled based on item response theory and vertically equated in order to allow use of the same metric as the difficulty of the test increases—a precondition for the analysis of growth models. Finally, the amount and nature of learning opportunities would have to be assessed at all points of measurement. Experimental intervention studies involving a treatment condition in which students are systematically provided with developmentally appropriate learning opportunities are also warranted.

Conclusion

The introduction of early-start language programs in elementary schools is a good example of how structural reforms implemented in one part of the education system may have unintended consequences for other parts of the system, and how progress in one area may be counteracted by problems in another. Our findings suggest that to ensure cumulative language learning across educational levels, it will be necessary to develop programs spanning the whole school career. This is a challenge that requires elementary and secondary systems to work in concert at institutional and operational levels.

The data of the present study were collected in 2009, at a time when a control group of students who started foreign language learning only at secondary school was still available. Since then, more than 10 years have passed, and the proportion of students with an early start has risen from around 70% to well over 90% (Stanat et al., 2016, p. 178); but the fact that the national educational standards, most of the curricula of the German federal states, and secondary textbooks have not yet been adapted to the changed situation speaks for the continuing relevance of our findings.

Our findings could also inform discussion across many parts of the world about whether foreign language teaching should begin at the age of 6 years, or whether it is better to wait another 2 years until students have learned to read and write in the language of instruction. In this regard, our findings support the idea that the learning of a foreign language improves in tandem with improvements in proficiency in the language of instruction; this positive relationship, measured up to 9 years after the English as a foreign language learning began, appeared to be strongest among those who had been in early-start English programs, whether they began at age 6−7 years or 8−9 years. This finding suggests that children with a weaker command of German faced particular challenges in early-start English, or, at least, were less likely to benefit from this early exposure in the longer term.

An additional nuance emerging from this study was that an early start in English proved to be a particular hurdle for bilingual students. For this group, proficiency in the language of instruction may be particularly important. Reading and listening proficiency in German at age 15–16 years seemed to predict reading and listening proficiency in the foreign language (English, at age 15–16 years) most strongly for bilinguals who had started learning both German (as the language of school instruction) and English (as a foreign language) together, aged 6−7 years, although this descriptive trend was not statistically significant. No such relationship was found for bilinguals who started learning English as a foreign language slightly later, aged 8−9 years. These intriguing results need to be confirmed and replicated before policy implications can be drawn.

Final revised version accepted 22 February 2020

Notes

[1] There is only one exception to this pattern of findings, from Austria (Buchholz, 2007). Note, however, that this study was based on a convenience sample of just 66 elementary students in three classes; it did not meet the basic requirements of test development; and the standard setting was arbitrary.

[2] In the German-speaking countries, the proficiency levels of the CEFR are often broken down further to allow more fine-grained distinctions.

[3] In 2015, the Zurich system was aligned to that of the other German-speaking cantons, with the start of English lessons being moved to Year 3.

[4] Due to marked ceiling effects in the German test at the second point of measurement, the test did not discriminate sufficiently between the groups at this point (see Pfenninger, 2016).

[5] In the 2014–2015 school year, when an early start had been fully implemented in all elementary schools in the canton of Zurich, Pfenninger and Singleton (2017, p. 29 and p. 62) again tested 102 early starters in two *Gymnasium* schools. This follow-up study did not solve the problem of the control group of late starters in the 2004–2005 school year being selective, however, as by 2014 there were no longer any late starters. All comparisons thus related to the same biased control group from 2004.

References

Adams, R., Wu, M., & Wilson, M. (2015). *ACER ConQuest: Generalised item response modelling software (Version 4)* [Computer software]. Camberwell, Australia: ACER.

Al-Thubaiti, K. A. (2010). *Age effects in a minimal input setting on the acquisition of English morpho-syntactic and semantic properties by L1 speakers of Arabic* (Unpublished doctoral dissertation). Essex University, Colchester, UK.

Barr, C. D., Uccelli, P., & Phillips Galloway, E.. (2019). Specifying the academic language skills that support text understanding in the middle grades: The design and validation of the core academic language skills construct and instrument. *Language Learning*, *69*, 978–1021. https://doi.org/10.1111/lang.12365

Barucki, H., Bliesener, U., Börner, O., Böttger, H., Hoffmann, I.-B., Kierepka, A., … Schlüter, N. (2015). *Der Lernstand im Englischunterricht am Ende von Klasse 4. Ergebnisse der BIG-Studie*. Munich, Germany: Domino-Verlag.

Baumert, J., Hohenstein, F., Fleckenstein, J., Preusler, S., Paulick, I., & Möller, J. (2017). Die schulischen Leistungen an der SESB—4. Jahrgangsstufe. In J. Möller, F. Hohenstein, J. Fleckenstein, O. Köller, & J. Baumert (Eds.), *Erfolgreich integrieren—die Staatliche Europa-Schule Berlin* (pp. 95–188). Münster, Germany: Waxmann.

Bennett, S. N. (1975). Weighing the evidence: A review of ‘Primary French in the Balance’. *British Journal of Educational Psychology*, *45*, 337–340. https://doi.org/10.1111/j.2044-8279.1975.tb02976.x

Bolster, A., Balandier-Brown, C., & Rea-Dickens, P. (2004). Young learners of modern foreign languages and their transition to the secondary phase: A lost opportunity? *Language Learning Journal*, *30*, 35–41. https://doi.org/10.1080/09571730485200211

Boyson, B. A., Semmer, M., Thompson, L. E., & Rosenbusch, M. H. (2013). Does beginning foreign language in kindergarten make a difference? Results of one district’s study. *Foreign Language Annals*, *46*, 246–263. https://doi.org/10.1111/flan.12023

Buchholz, B. (2007). *Facts & figures im Grundschul-Englisch: Eine Untersuchung des verbindlichen Fremdsprachenunterrichts ab der ersten Klasse an österreichischen Volksschulen*. Münster, Germany: LIT Verlag.

Burstall, C. (1975). Primary French in the balance. *Educational Research*, *17*, 193–198. https://doi.org/10.1080/0013188750170304

Burstall, C., Jamieson, M., Cohen, S., & Hargreaves, M. (1974). *Primary French in the balance*. Windsor, UK: NFER Publishing.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg, France: Council of Europe.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg, France: Council of Europe.

Courtney, L. M. (2014). *Moving from primary to secondary education: An investigation into the effect of primary to secondary transition on motivation for language learning and foreign language proficiency* (Unpublished doctoral dissertation). University of Southampton, UK.

Courtney, L., Graham, S., Tonkyn, A., & Marinis, T. (2017). Individual differences in early language learning: A study of English learners of French. *Applied Linguistics*, *38*, 824–847. https://doi.org/10.1093/applin/amv071

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*, 222–251. https://doi.org/10.3102/00346543049002222

De Bot, K. (2014). The effectiveness of early foreign language learning in the Netherlands. *Studies in Second Language Learning and Teaching*, *4*, 409–418. https://doi.org/10.14746/ssllt.2014.4.3.2

DeKeyser, R. (2020). Input is not a panacea. *International Journal of Bilingualism*, *24*, 79–81. https://doi.org/10.1177/1367006918768371

DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 88–108). Oxford, UK: Oxford University Press.

Dyssegaard, C. B., de Hemmer Egeberg, J., Sommersel, H. B., Steenberg, K., & Vestergaard, S. (2015). *A systematic review of the impact of multiple language teaching, prior language experience and acquisition order on students' language proficiency in primary and secondary school*. Copenhagen, Denmark: Danish Clearinghouse for Educational Research.

European Commission. (2017a). *Key data on teaching languages at school in Europe: 2017 edition* [Eurydice report]. Luxembourg: Publications Office of the European Union.

European Commission. (2017b). *EU languages*. Retrieved from https://europa.eu/european-union/about-eu/eu-languages_en

Fleckenstein, J., Möller, J., & Baumert, J. (2018). Mehrsprachigkeit als Ressource. *Zeitschrift für Erziehungswissenschaft*, *21*, 97–120. https://doi.org/10.1007/s11618-017-0792-9

Galton, M., Gray, J., & Rudduck, J. (2003). *Transfer and transitions in the middle years of schooling (7–14): Continuities and discontinuities in learning*. London, UK: DfES Publications.

Galton, M., & McLellan, R. (2018). A transition Odyssey: Pupils' experiences of transfer to secondary school across five decades. *Research Papers in Education*, *33*, 255–277. https://doi.org/10.1080/02671522.2017.1302496

Ganzeboom, H. B., de Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*, 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

Gebauer, S. K., Zaunbauer, A. C. M., & Möller, J. (2012). Erstsprachliche Leistungsentwicklung im Immersionsunterricht: Vorteile trotz Unterrichts in einer Fremdsprache? *Zeitschrift für Pädagogische Psychologie*, *26*, 183–196. https://doi.org/10.1024/1010-0652/a000071

Gebauer, S. K., Zaunbauer, A. C. M., & Möller, J. (2013). Cross-language transfer in English immersion programs in Germany: Reading comprehension and reading fluency. *Contemporary Educational Psychology*, *38*, 64–74. https://doi.org/10.1016/j.cedpsych.2012.09.002

Genelot, S. (1997). L'enseignement des langues vivantes à l'école élémentaire: Eléments d'évaluation des effets au collège. *Revue française de pédagogie*, *118*, 27–42. https://doi.org/10.3406/rfp.1997.1173

Goorhuis-Brouwer, S., & de Bot, K. (2010). Impact of early English language teaching on L1 and L2 development in children in Dutch schools. *International Journal of Bilingualism*, *14*, 289–302. https://doi.org/10.1177/1367006910367846

Graham, S., Courtney, L., Marinis, T., & Tonkyn, A. (2017). Early language learning: The impact of teaching and teacher factors. *Language Learning*, *67*, 922–958. https://doi.org/10.1111/lang.12251

Groot-Wilken, B., & Husfeldt, V. (2013). Die Testinstrumente und -verfahren des EVENING-Projekts: Eine empirische Betrachtungsweise. In O. Börner, G. Engel, & B. Groot-Wilken (Eds.), *Hörverstehen, Leseverstehen, Sprechen* (pp. 121–140). Münster, Germany: Waxmann.

Haenni Hoti, E., Müller, M., Heinzmann, S., Wicki, W., & Werlen, E. (2009). *Frühenglisch – Überforderung oder Chance? Eine Längsschnittstudie zur Wirksamkeit des Fremdsprachenunterrichts auf der Primarstufe* (Research report No. 20). Lucerne, Switzerland: PHZ Luzern.

Haenni Hoti, A., & Werlen, E. (2007). Englischunterricht (L2) in den Zentralschweizer Primarschulen: Hat er einen positiven oder negativen Einfluss auf das Leseverständnis der SchülerInnen in Deutsch (L1)? In E. Werlen & R. Weskamp (Eds.), *Kommunikative Kompetenz und Mehrsprachigkeit* (pp. 139–160). Baltmannsweiler, Germany: Schneider Verlag.

Harsch, C., Pant, A., & Köller, O. (Eds.). (2010). *Calibrating standards-based assessment tasks for English as a first foreign language: Standard-setting procedures in Germany*. Münster, Germany: Waxmann.

Heinzmann, S., Müller, M., Oliveira, M., Hoti Haenni, A., & Wicki, W. (2009). *Englisch und Französisch auf der Primarstufe: Verlängerung des NFP-56-Projekts. Schlussbericht* (Research report No. 23). Lucerne, Switzerland: PHZ Luzern.

Hempel, M., Kötter, M., & Rymarczyk, J. (2018). *Fremdsprachenunterricht in der Grundschule in den Bundesländern Deutschlands*. Frankfurt, Germany: Peter Lang.

Hopp, H., Vogelbacher, M., Kieseier, T., & Thoma, D. (2019). Bilingual advantages in early foreign language learning: Effects of the minority and the majority language. *Learning and Instruction*, *61*, 99–110. https://doi.org/10.1016/j.learninstruc.2019.02.001

Huang, B. H. (2016). A synthesis of empirical research on the linguistic outcomes of early foreign language instruction. *International Journal of Multilingualism*, *13*, 257–273. https://doi.org/10.1515/9783110197372.0.1

Jaekel, N., Schurig, M., Florian, M., & Ritter, M. (2017). From early starters to late finishers? A longitudinal study of early foreign language learning in school. *Language Learning*, *67*, 631–664. https://doi.org/10.1111/lang.12242

Kalberer, U. (2007). *Rate of L2 acquisition and the influence of instruction time on achievement* (Unpublished master's thesis). University of Manchester, UK.

Keßler, J.-U. (2006). *Englischerwerb im Anfangsunterricht diagnostizieren*. Tübingen, Germany: Narr.

KMK = Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Schulabschluss*. Retrieved from https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2003/2003_12_04-BS-erste-Fremdsprache.pdf

KMK = Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2013). *Fremdsprachen in der Grundschule: Sachstand und Konzeptionen 2013*. Retrieved from https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2013/2013_10_17-Fremdsprachen-in-der-Grundschule.pdf

KMK = Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2014). *Vereinbarung über die Schularten und Bildungsgänge im Sekundarbereich I*. Retrieved from http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1993/1993_12_03-VB-Sek-I.pdf

KMK = Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring, überarbeitete Fassung 2015*. Retrieved from https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungsmonitoring.html

Köller, O., Knigge, M., & Tesch, B. (2010). *Sprachliche Kompetenzen im Ländervergleich*. Münster, Germany: Waxmann.

Lambelet, A., & Berthele, R. (2015). *Age and foreign language learning in school*. New York, NY: Palgrave Macmillan.

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research*, *24*, 35–63. https://doi.org/10.1177/0267658307082981

Lehmann, R., & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis: Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin*. Berlin, Germany: Humboldt Universität.

Lin, H., Chang, H., & Cheung, H. (2004). The effects of early English learning on auditory perception of English minimal pairs by Taiwan university students. *Journal of Psycholinguistic Research*, *33*, 25–49. https://doi.org/10.1023/B:JOPR.0000010513.34809.61

Mihaljevic Djigunovic, J., Nikolov, M., & Otto, I. (2008). A comparative study of Croatian and Hungarian EFL students. *Language Teaching Research*, *12*, 433–452. https://doi.org/10.1177/1362168808089926

Muñoz, C. (Ed.). (2006a). *Age and the rate of foreign language learning*. Clevedon, UK: Multilingual Matters.

Muñoz, C. (2006b). The effects of age on foreign language learning: The BAF Project. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 1–40). Clevedon, UK: Multilingual Matters.

Muñoz, C. (2011). Input and long-term effects of starting age in foreign language learning. *International Review of Applied Linguistics in Language Teaching*, *49*, 113–133. https://doi.org/10.1515/iral.2011.006

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, *35*, 463–482. https://doi.org/10.1093/applin/amu024

Muñoz, C., & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*, 1–35. https://doi.org/10.1017/S0261444810000327

Muñoz, C., Tragant, E., & Camuñas, M. (2015). Transition: Continuity or a fresh start? *APAC Quarterly*, *89*, 11–16.

Muthén, L., & Muthén, B. (2012). *Statistical analysis with latent variables: Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Ofsted = Office for Standards in Education. (2011). *Modern Languages: Achievement and challenge 2007 to 2010*. Retrieved from https://www.gov.uk/government/publications/modern-languages-achievement-and-challenge-2007-to-2010

Oller, J., Jr., & Nagato, N. (1974). The long-term effect of FLES: An experiment. *Modern Language Journal*, *58*, 15–19. https://doi.org/10.2307/323984.

Peyer, E., Andexlinger, M., Kofler, K., & Lenz, P. (2016). *Projekt Fremdsprachenevaluation BKZ Schlussbericht zu den Sprachkompetenztests*. Fribourg, Switzerland: Institut für Mehrsprachigkeit.

Pfenninger, S. E. (2016). The literacy factor in the optimal age discussion: A five-year longitudinal study. *International Journal of Bilingual Education and Bilingualism*, *19*, 217–234. https://doi.org/10.1080/13670050.2014.972334

Pfenninger, S. E. (2017). Not so individual after all: An ecological approach to age as an individual difference variable in a classroom. *Studies in Second Language Learning and Teaching*, *7*, 19–46. https://doi.org/10.14746/ssllt.2017.7.1.2

Pfenninger, S. E., & Singleton, D. (2016). Affect trumps age: A person-in-context relational view of age and motivation in SLA. *Second Language Research*, *32*, 311–345. https://doi.org/10.1177/0267658315624476

Pfenninger, S. E., & Singleton, D. (2017). *Beyond age effects in instructional L2 learning: Revisiting the age factor*. Bristol, UK: Multilingual Matters. https://doi.org/10.21832/PFENNI7623

Pfenninger, S. E., & Singleton, D. (2019). Starting age overshadowed: The primacy of differential environmental and family support effects on second language attainment in an instructional context. *Language Learning*, *69*, 207–234. https://doi.org/10.1111/lang.12318

Richardson, K. (2014). *The primary-secondary school transition for languages: Pupil and teacher experiences and beliefs* (Unpublished doctoral dissertation). University of Warwick, UK.

Ritter, M., Jaekel, N., Meister, C., & Lewandowska, Z. (2015). Zwischenbilanz der fachdidaktischen Arbeit im Fach Englisch. In H. Wendt & W. Bos (Eds.), *Auf dem Weg zum Ganztagsgymnasium: Erste Ergebnisse der wissenschaftlichen Begleitforschung zum Projekt Ganz In* (pp. 300–323). Münster, Germany: Waxmann.

Rolstad, K., & MacSwan, J. (2014). The facilitation effect and language thresholds. *Frontiers in Psychology*, *5*, 1197. https://doi.org/10.3389/fpsyg.2014.01197

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Sanz, C. (2008). Predicting enhanced L3 learning in bilingual contexts: The role of biliteracy. In C. Pérez-Vidal, A. Bel, & M. Juan-Garau (Eds.), *A portrait of the young in the new multilingual Spain* (pp. 220–240). Clevedon, UK: Multilingual Matters.

Singleton, D., & Muñoz, C. (2011). Around and beyond the critical period hypothesis. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 407–425). London, UK: Routledge.

Singleton, D., & Pfenninger, S. E. (2019). The age debate: A critical overview. In S. Garton & F. Copland (Eds.), *The Routledge handbook of teaching English to young learners* (pp. 30–43). London, UK: Routledge.

Sparks, R., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2008). Early first-language reading and spelling skills predict later second-language reading and spelling skills. *Journal of Educational Psychology*, *100*, 162–174. https://doi.org/10.1037/0022-0663.100.1.162

Stanat, P., Böhme, K., Schipolowski, S., & Haag, N. (Eds.). (2016). *IQB-Bildungstrends. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweitenLändervergeich*. Münster, Germany: Waxmann.

Ströbel, M., Kerz, E., & Wiechmann, D. (2020). The relationship between first and second language writing: Investigating the effects of L1 complexity on L2 complexity in advanced stages of learning. *Language Learning*, Advance online publication. https://doi.org/10.1111/lang.12394

Szpotowicz, M., & Lindgren, E. (2011). Language achievements: A longitudinal perspective. In J. Enever (Ed.), *ELLiE: Early language learning in Europe* (pp. 125–142). London, UK: British Council.

UNESCO = United Nations Educational, Scientific and Cultural Organization. (2006). *International Standard Classification of Education. ISCED 1997*. Paris: UNESCO.

Unsworth, S., Persson, L., Prins, T., & de Bot, K. (2015). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics*, *36*, 527–548. https://doi.org/10.1093/applin/amt052

Van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, *98*, 3–53. https://doi.org/10.1037//0033-295X.98.1.3

Von Ow, A., Husfeldt, V., & Bader-Lehmann, U. (2012). Einflussfaktoren für den Lernerfolg von Englisch an der Primarschule: Eine Untersuchung in fünf Schweizer Kantonen und dem Fürstentum Liechtenstein. *Babylonia*, *22*, 52–57.

Wilden, E., & Porsch, R. (2016). Learning EFL from Year 1 or Year 3? A comparative study on children's EFL listening and reading comprehension at the end of primary education. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 191–212). New York, NY: Springer.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Addendum to Research Review.
Appendix S2. Addendum to The Present Study.
Appendix S3. Addendum to Methods.
Appendix S4. Addendum to Results.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Starting to Learn a Foreign Language Early Does Not Necessarily Have Long-Term Benefits

What This Research Was About and Why It Is Important

Introducing foreign language learning in the early years of elementary school has become a popular educational policy throughout Europe. But does an early start actually have positive long-term effects on students' language skills? If not, the value of this (costly and time-consuming) policy needs to be reconsidered. This article reviews the state of research on early foreign language learning and reports findings from a new study on the long-term effects of an early start. The study drew on a representative sample of Year 9 students in Germany who started learning English at different points in their school lives. In the long term, there were no differences in the English listening and reading skills of early and later starters. It seems likely that this was mainly because secondary schools did not adapt to build on the knowledge gained at elementary level.

What the Researchers Did

- The researchers reviewed previous research on the effects of early foreign language learning, summarizing the methodological strengths and weaknesses as well as the findings of these studies.
- In a new study, they investigated the English as a foreign language skills of about 20,000 Year 9 students, comparing the reading and listening comprehension of early starters (English from Year 1 or 2 of elementary school), a middle group (English from Year 3 or 4 of elementary school), and late starters (English from Year 5, the first year of secondary school).
- They accounted for further differences between the three groups by including information on students' individual characteristics, family background, and institutional context (e.g., school type, federal state in Germany).

What the Researchers Found

- Many previous studies comparing early and late starters have not paid attention to group differences that influence students' language skills above and beyond the start of language instruction at school.
- Taking such group differences into account, the researchers found no differences in the English listening and reading skills of the three groups by the time they reached Year 9.
- This is probably because secondary language education did not systematically build on what early starters had learned at elementary school.

Things to Consider

- The older students learned faster, enabling them to catch up with the early starters.
- Children with weak skills in the majority language may face particular challenges in early-start programs.
- Foreign language education programs need to span the whole school career and bridge the gap between elementary and secondary school. Teacher education, curricula, and teaching practices need to be properly aligned.

Materials and data: Data are publicly available at https://www.iqb.hu-berlin.de/fdz/studies/IQB-LV_2008-9

How to cite this summary: Baumert, J., Fleckenstein, J., Leucht, M., Köller, O., & Möller, J. (2020). Starting to learn a foreign language early does not necessarily have long-term benefits. *OASIS Summary* of Baumert, Fleckenstein, Leucht et al. (2020) in *Language Learning*. https://oasis-database.org