System 106 (2022) 102763

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

The impact of early foreign language learning on language proficiency development from middle to high school

Nils Jaekel [a,*], Michael Schurig [b], Isabelle van Ackern [c], Markus Ritter [c]

[a] University of Oulu, Finland
[b] Technische Universität Dortmund, Germany
[c] Ruhr-Universität Bochum, Germany

ARTICLE INFO

Keywords:
Early foreign language learning
Receptive language skills
Learner characteristics
Linear mixed modeling

ABSTRACT

Early foreign language instruction has become the norm across Europe. Expected benefits for students include linguistic advantages and ease of learning the second language (L2). However, research rarely supports these assertions. The present study investigated the receptive skills of two cohorts of English as a Foreign Language (EFL) learners in Years 5, 7, and 9 in Germany. The cohorts differed in their age of EFL onset in elementary school and, consequently, in the amount of exposure before secondary school. Linear mixed model analyses were employed to account for the hierarchical structure of the data. Learners with an earlier start performed better in Years 5 and 9 than late starters, suggesting possible long-term benefits of an earlier start. In Year 7, late starters scored higher on the proficiency assessment. Across the Year 5–9 span, the effects of learner characteristics' on English proficiency remained stable for gender, L1, grades, cognitive abilities, and cultural capital.

1. Introduction

Learning a foreign language from an early age in pre-or elementary school is generally understood as a valuable foundation for young learners' long-term language proficiency development and development of intercultural awareness (European Commission, 2011; Kohler, 2017). The introduction of early foreign language programs in elementary schools and policies in Europe (Eurostat, 2019), Asia (Butler, 2014), South and Central America (Cronquist & Fiszbein, 2017), and Australia (Kohler, 2017) underscore its global popularity and the belief in its value for education. For example, in the European Union, 82% of elementary school students receive foreign language education (Eurostat, 2019), a development that can be attributed to Europe's 'mother tongue+2' language policy (Council of the European Union, 1997). The popularity of early foreign language learning is similar across South and Central America, Asia, and Australia, which have aimed to broaden access to early foreign language education (Cronquist & Fiszbein, 2017; Bolton & Graddol, 2012).

However, the conception and benefits of early foreign language learning (EFLL) still need to be empirically substantiated to justify an early start of foreign language education in elementary school and time in the curriculum over a later start in middle school. Studies investigating language proficiency longitudinally after an earlier onset of foreign language education in elementary school across middle school remain scarce. Consequently, little is known about the long-term impact and the effects of students' individual

Abbreviations: EFLL, Early Foreign Language Learning.

* Corresponding author.
E-mail address: nils.jaekel@oulu.fi (N. Jaekel).

https://doi.org/10.1016/j.system.2022.102763
Received 9 April 2021; Received in revised form 18 January 2022; Accepted 7 February 2022
Available online 10 February 2022

differences on language proficiency later in middle and high school. Further, the significance of the transition from elementary to secondary school has only very recently been addressed in SLA research (Courtney, 2017; Jaekel et al., 2017; Pfenninger & Lendl, 2017), while it has a longer tradition in general education research (Anderson et al., 2000).

The present study contributes to the puzzle of understanding long-term effects of an early start to foreign language education. This study investigated whether an early start could keep its promise of higher language proficiency. We sought to investigate the effects of an early onset of English as a Foreign Language (EFL) learning in elementary schools on the outcomes of receptive language proficiency from Year 5 to Year 9. The study followed students' language proficiency development. It was conducted in the state of North Rhine-Westphalia, Germany, and consisted of two cohorts of students, one who started EFL in Year 1 (age 6–7 years, early starters (ES)), the other in Year 3 (8–9 years, late starters (LS)). The outcome variables were receptive English proficiency in listening and reading. We employed linear mixed model analyses to account for the hierarchical structure of the data, additionally controlling for sex, L1, cognitive abilities, cultural capital (books), English grade (Year 8), and socioeconomic status (SES).

2. Status quo of education policy, curricula, and research regarding early foreign language learning

The introduction of EFLL into the elementary school curriculum in Europe results from a top-down approach, i.e., guided by European policies (Council of the European Union, 1997). While acknowledging the apparent benefits of multilingualism in a globalized world, parents across Europe have also long been advocating for early foreign language instruction in elementary schools. Across Europe, EFLL picked up momentum with the introduction of the European language policy, but the implementation of EFLL programs has been described as hastily (Piske, 2013), with a lack of teacher preparation and training for elementary and secondary teachers (Edelenbos et al., 2006; Jaekel et al., 2017), and neglecting available SLA research (Schmelter, 2010).

Several prevalent arguments in favor of EFLL are regularly brought forward by various stakeholders, such as policymakers or parents, supporting EFLL (Jaekel et al., 2017). (1) Research from immersion programs in Canada (Lapkin & Swain, 1982; Wesche, 2002) and the fact that multilingual societies such as Luxemburg or Switzerland offer bilingual education. (2) An argument that is often cited in non-academic contexts is that the successful integration of bilingual immigrant children is evidence in support of EFLL. (3) EFLL programs have shown moderate to above-average results for productive and receptive skills (Barucki et al., 2015; Enever, 2011; Engel, 2009). These studies, however, lack a baseline against which language development could be compared, and thus no conclusions can be drawn if an earlier start has particular advantages over a later start. (4) Potential advantages of EFLL have not only been focused on proficiency measures but have also entailed desired motivational benefits (Courtney, 2017; Graham et al., 2016; Pfenninger & Lendl, 2017) and have supported the development of intercultural awareness, which is fundamental in European policies (Demircioglu, 2008; Education, Audiovisual & Culture Executive Agency, 2017; Roos, 2007). (5) In multilingual countries or multi-state unions (EU), multilingualism is supported through policies, although sometimes regionally limited. EFLL has the potential to provide more time to develop advanced proficiency in more than one foreign language and foster a deeper cultural understanding of more than one culture. As such, the European Union's policies, for example, strongly encourage and support EFLL (Council of Europe, 2005). Furthermore, the aspired advanced proficiency of learners would foster geographic mobility within and beyond a country's borders as well as provide opportunities for upward mobility. Multilingualism is necessary for these societies to function and is believed to be a catalyst for social and cultural understanding.

With the many potential advantages of EFLL, it is important to set realistic expectations and draw conclusions from a sound research-based foundation. This requires careful consideration of the available research within the respective educational contexts and societal constellations, which may be more or less supportive of EFLL.

For example, one fallacy in evaluating the merits of EFLL can be drawing conclusions for low-exposure EFLL programs from research into immersion or successful integration of immigrant children. This approach neglects the stark differences of these diverse contexts in contrast to the foreign language classrooms in elementary or secondary schools. Immersion programs, intensive language programs such as Content and Language Integrated Learning (CLIL), or bilingualism due to immigration generally provide children with immense exposure to rich, authentic language input and the opportunity to engage with the L2 in a meaningful way. Providing comparatively minimal input to the second language in a non-immersive foreign language classroom cannot emulate these approaches (Larson-Hall, 2008). Furthermore, in the context of immigration, SLA is essential with a sense of urgency to acquire a new language to survive, integrate into a new culture, and function in society. Altogether, promoting potential EFLL benefits by referencing research on the effects of bilingualism can thus be grossly misleading.

2.1. The earlier, the better?

The popularity of implementing foreign language education before middle or even elementary school finds one of its main roots in the belief that 'earlier is better' when it comes to learning languages. Parallels have been drawn between the context and process of L1 acquisition for babies, toddlers, and young children and the elementary school context and methodology used to teach L2 (Grotjahn, 2005). L1 language acquisition at an early age is natural, i.e., guided by necessity, generally successful, and more effective, e.g., children growing up absorbing language structures and vocabulary like a sponge, as opposed to L2 language learning at an older age potentially posing a higher cognitive burden, increased workload and the pitfall of not sounding native-like (N. Ellis, 2017). Two key differences between the family L1 acquisition versus the educational L2 learning context are the significantly higher exposure to and level of contextuality, i.e., language acquisition focuses on more imminent needs. L1 acquisition follows a generally successful implicit process, but it is not necessarily sufficient for L2 learning as the brain has been attuned to the L1. "Transfer, learned attention, and automatization" that are developed during the L1 acquisition process are potentially in conflict with a successful implicit learning

process for L2 (N. Ellis, 2017, p. 119).

2.2. Curricula constraints

However, curricula for EFLL heavily rely on implicit language learning. The argument is that students at this early age benefit from exposure to the language and may not benefit from explicit instruction as much. However, Lichtman (2016) argued that there are few differences between children and adults learning a new language, except that adults may prefer explicit instruction due to their abundance of exposure to it and their ability to abstract from L1. Therefore, foreign language education at the secondary level is much more explicit than in elementary schools. Transitioning from elementary to high school level foreign language classes, students are likely not able to produce explicit knowledge about the L2 or use meta-language as they may not have learned it, creating a potential mismatch with teacher expectations at the secondary level. The minimal exposure to the target language in many elementary curricula has been pointed out as one of the key problem areas impeding the long-term success of EFLL (Jaekel et al., 2017; Larson-Hall, 2008; Piske, 2017). Regular EFLL programs continue to be constrained by minimal exposure, often with 90 min or less per week allocated to elementary school teaching, large groups with small amounts of individual conversational practice, and the explicit testing of the L2 knowledge and understanding (Edelenbos et al., 2006). However, this cannot be an excuse for secondary teachers to neglect their incoming students' L2 proficiency and adapt their curriculum accordingly.

2.3. Is there a case for a later start?

The rapid progress older learners make has been attributed to their more advanced L1 reading and writing skills (Cummins & Swain, 1986; Lapkin et al., 1980), higher levels of cognitive maturity (Harley, 1986; Muñoz, 2006), and the focus on explicit teaching approaches (DeKeyser & Larson-Hall, 2005). More explicitly focused language teaching relies on students' knowledge about grammatical structures in their L1 or contrasting them with the L2, which allows them to progress faster. In contrast, implicit language learning progresses more slowly as learners need much more exposure to language structures to internalize and reproduce them.

One of the central promises of an early start to language learning is better language proficiency (DeKeyser & Larson-Hall, 2005). However, very few studies have investigated learners' L2 proficiency development through middle or high school. In a comprehensive review, Huang included 42 empirical studies published in the past 50 years and could not "find unequivocal evidence for a younger learner advantage" (Huang, 2016, p. 268). Early on after the transition to secondary school, either the earlier start or increased classroom time favors early starters over late starters concerning receptive language skills (Wilden & Porsch, 2016). The length of exposure to the L2, which is generally tied to an earlier start in school contexts, has benefited learners regarding various language skills, including grammaticality judgment tasks (García Mayo, 2003). However, throughout the middle school years, older learners are at an advantage and make faster progress than younger learners (Jaekel et al., 2017; Muñoz, 2008; Pfenninger, 2014). Cognitive maturation and more advanced literacy skills in the majority language have been identified as likely factors for the rapid development of older learners (Jaekel et al., 2017; Muñoz, 2008; Pfenninger, 2014). Secondary school students may have a favorable attitude towards secondary school practices resulting from a feeling of making faster progress than in elementary school (Courtney, 2017; Graham et al., 2016; Piske, 2017). Courtney (2017) has shown that the perception of the transition and the corresponding development of motivation, attitudes, and competencies are interrelated with pedagogic language practices. Therefore, an empirical difference between EFLL impact on language skills and attitudes or motivation becomes visible.

While research suggests that older learners show advantages in developing L2 language proficiency, EFLL programs have reported that learners meet curricula proficiency thresholds at the end of elementary school and thus develop basic language proficiency (Bos, 2007; de Bot, 2014).

3. Investigating the long-term effects of EFLL

The long-term success of EFLL depends on a plethora of factors in elementary and secondary schools. At both elementary and secondary levels, potentials for improvement have been identified, such as the availability of sufficient numbers of highly qualified and proficient language teachers (Jaekel et al., 2017; Kolb, 2011; Piske, 2017; Unsworth et al., 2015), developing adequate curricula including schoolbooks and other resources as well as ensuring a smooth transition from elementary school into secondary education (Jaekel et al., 2017; Pfenninger & Lendl, 2017). The latter issue may act as an Achilles heel for successful EFLL programs (Courtney, 2017; Jaekel et al., 2017; Kersten & Rohde, 2013). Students reported mixed perceptions of the transition and how it connects to their elementary experience (Jaekel et al., 2021). Research into the long-term development of EFLL comparing early versus late starters, particularly within the school context, remains rare. Generally, language learning in formal school settings does not show that younger learners outperform late starters in contexts with equal (Cenoz, 2002; Holmstrand, 1982) or increased amounts of exposure (Baumert et al., 2020; Jaekel et al., 2017; Muñoz, 2006; Oller & Nagato, 1974).

The Barcelona Age Factor (BAF) Project (Muñoz, 2006) followed the language development of learners with different ages of onset, from 8, 11, 14, and 18 years. All learners were proficient Catalan-Spanish bilinguals, and English constituted their L3. English programs progressed with different levels of intensity. The youngest group (age of onset (AO) 8), for example, required three years to reach 416 h of instruction, while the next oldest group (AO 11) only required two. The results from an extensive test battery ranging from a Cloze test to interviews and grammar tests demonstrated a faster rate of language skill development for late starters during the first 200 h (T1). Older learners for almost all skills outperformed all younger groups initially but slowed down significantly in their improvement into T2 after 416 h. The AO 11 group caught up to proficiency levels of older learners by T3, or after 726 h, for many of

the tested language skills. However, the youngest group (AO 8) continuously trailed the older groups without surpassing them. Muñoz concluded that "age effects are not uniform across measures of language abilities" (Muñoz, 2006, p. 30). Explicit instruction, paired with more advanced cognitive development of older learners, had likely benefited them initially. Overall, the study demonstrated a complex interplay of the age of onset and exposure to the language.

One study with partially positive long-term effects was conducted by Pfenninger and Singleton (Pfenninger & Singleton, 2019) in Switzerland. The study compared early classroom learners from Year 2 with late classroom learners in Year 7. The study is particularly unique as it controlled for different forms of language use at home and literacy in these languages, i.e., bilingualism and biliteracy. Data were collected when students were 13–14 years and 18–19 years. Students were assessed on a variety of receptive and productive tasks, including an argumentative essay and grammatical judgment tasks. Despite five years of additional exposure and after initial advantages in the first assessment for early starters regarding some productive and receptive measures, early starters had no discernible advantages in the second assessment. Only the group of simultaneous bilinguals who also demonstrated biliteracy and also benefited from an overall supportive and advantageous home environment, i.e., high SES, high levels of parental education, and high levels of parent involvement, showed long-term benefits of an early start. The authors concluded that contextual factors trump age of onset and that AO lacks stability as a factor over time (Pfenninger & Singleton, 2019).

More recently, Baumert et al. (2020) drew on data from a large, nationwide assessment in Germany. The cross-sectional study assessed English listening and reading performance in Year 9. The analyses controlled for a wide range of background variables using stepwise linear regression. The authors identified three groups of English students: early, middle, and late starters. These students started in Years 1 and 2, Years 3 and 4, and Year 5, respectively. None of the four models, which successively added background variables, German scores, school tier, and state, significantly differed between the three groups' receptive English skills. Across models, proficiency in the language of instruction, i.e., German, was shown to be the strongest predictor of English receptive skills in Year 9. The authors concluded that there is no evidence for an advantage of an early start after five years of secondary school. Similarly to other studies in SLA (Courtney, 2017; Jaekel et al., 2017; Pfenninger & Lendl, 2017), the study identified the transition to secondary school and, more specifically, the lack of adapting the curriculum and teaching practices to the learners English skills as a significant impediment for the success of EFLL.

4. Study context

The present study was part of a multi-school project in the state of North-Rhine Westphalia (NRW), Germany, called *Ganz In - All-Day Schools for a Brighter Future [Mit Ganztag mehr Zukunft],* which ran from 2010 until 2018 in two phases. The project investigated the impact of all-day schooling at grammar schools and was a multidisciplinary project, the data presented here focus on English proficiency development.

Students' development of receptive EFL proficiency, i.e., listening and reading, was the primary interest for the present study. Data were collected from two cohorts of students in Years 5, 7, and 9. The late starter (LS) cohort began EFL in Year 3 of elementary school (i.e., age 8–9 years, 2008), while the early starter (ES) cohort began EFL in the second half of Year 1 (i.e., age 6–7 years, 2008, see Fig. 1). The quantitative difference in exposure to EFL lessons between both cohorts was 105 lessons, generally 45-min lessons, for a total of 78 ¾ hours (see Table 1; MSW - NRW, 2008).

Germany's secondary school system differs between states. Generally, students are streamed into different secondary schools after Year 4. In North-Rhine Westphalia, Germany, the *Hauptschule* (lower secondary), *Realschule* (middle secondary), and the *Sekundarschule* (middle secondary) constitute options for middle years that finish with a middle year's degree. The *Gymnasium* (grammar school) and *Gesamtschule* (comprehensive school) offer high school qualifications that provide full access to tertiary education, which requires proficiency in two foreign languages offered in school. With the focus on grammar schools, participants in this study will have shown better than average academic aptitude at their elementary school as grammar schools generally attract students with better grades or more promising academic development.

As discussed above in the literature review, during the first years after introducing English in Year 1, few English teachers at elementary schools were trained during their initial licensure phase at the university. Elementary school teachers who did not have prior licensure were able to qualify through a language certificate. Similarly, teaching materials at the elementary and secondary level geared towards the early start in Year 1 were not widely available. Training for middle school teachers was also scarce, and teachers had to adapt their classes to understand the effects of an early start better.

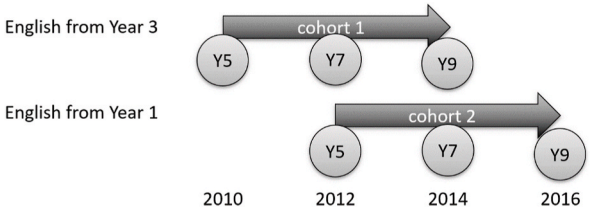

Fig. 1. Outline of the project and assessments.

Table 1
Amount of EFL instruction in lessons (a lesson is generally 45 min long).

	LS	ES
By Year 5 (T1)	140	245
By Year 7 (T2)	444	549
By Year 9 (T3)	710	815

5. Research questions and hypotheses

The study aimed to investigate the English receptive language proficiency of early (AO Year 1) in contrast to late starters (AO Year 3) in Year 9, a difference of 105 h of EFL instruction. The study was designed as a natural experiment. A key aim of this study was to evaluate the EFLL policy introduced in the state. We were thus evaluating the success of this educational policy in its natural context. Analyses controlled for learner characteristics, sex, SES, L1, cultural capital, household income, the last grade in English in Year 8, and cognitive abilities.

Research question: How do very early (AO Year 1) and early starters (AO Year 3) differ in their receptive skills in Year 9 after controlling for confounding variables?

6. Methods

6.1. Participants

The sample consisted of 12 schools in North-Rhine Westphalia, Germany, that participated in assessments from 2010 to 2016 (see Fig. 1). Data for LS students were collected in Years 5, 7, and 9 in 2010, 2012, and 2014, respectively. Students' age at the initial data point in Year 5 was between 10 and 11 years of age, by Year 9, students were between 14 and 15 years old. The data collection between 2010 and 2014 was part of the *Ganz In* Project. A follow-up assessment in 2016 only focused on English language assessments. Consequently, the assessment day in 2016 was shorter, with 60 min of testing in contrast to a full testing day covering other content and general pedagogical assessments. Data was collected on at least one proficiency test in one of the three assessment years from 2827 Students (Cohort 1 = 1418; Cohort 2 = 1409). 1519 students (Cohort 1 = 805; Cohort 2 = 714) took part in all three assessments in Years 5, 7, and 9. In order to avoid imputing outcome variables, students were only included in analyses if they participated in all assessments. Of the students that were excluded, across both cohorts, 502 students (19.9%) participated in Year 5, but not in Year 7, and 855 (33.9%) participated in Year 5 but not in Year 9. Another 602 students across both cohorts participated in Year 7 but not in Year 9 (27%; see Table 3 in Appendix).

Student participation was voluntary; written consent was obtained from parents before data collection commenced, and students assented their participation at the time of the assessment. The tests were administered between weeks five and nine at the beginning of the new school year. Data collection was conducted during regular school lessons.

6.2. Instruments

Across the project, both cohorts received identical language and cognitive assessments as well as the same background survey. See Fig. 1 for a detailed overview of the assessment period.

6.2.1. Receptive language assessments

English listening and reading comprehension tests were standardized and were previously used and validated in large-scale assessments in Germany (Engel & Ehlers, 2013; Institut zur Qualitätssicherung im Bildungswesen, 2014; Wilden et al., 2013). English proficiency at the end of Year 9, according to the state of NRW's curriculum, is Common European Framework of Reference for Language Learning A2 to B1. The tasks selected included items that were rated as CEFR A2-C1.

The language assessments included multiple-choice, sentence completion, and open-answer items (see Table 2 for more

Table 2
Overview of test items per grade level assessment.

Grade	Skill	Items	Cronbach's alpha, α	Tasks
5	Listening	28	.69	Multiple Choice, listening for gist, listening for details, understanding core message
5	Reading	24	.74	Multiple Choice, open-answer items with a short written response (1–3 words), understanding core message, extracting specific information
7	Listening	26	.79	Multiple Choice, open-answer items, sentence completion, listening for details and specific information
7	Reading	26	.89	Multiple Choice, open-answer items, sentence completion, understanding core message, extracting specific information
9	Listening	18	.73	Multiple Choice, open-answer items, sentence completion, listening for specific information
9	Reading	12	.78	Multiple Choice, open-answer items, sentence completion, extracting specific information

information). The assessments covered a wide range of themes in accordance with the curriculum covering age-appropriate topics relevant for students' lives ranging from everyday activities, family life, leisure time activities, story-telling, vacation, or job application. Responses were coded dichotomously (correct vs. incorrect). Missing values were coded as incorrect. When the reading or the listening test was missing completely, the respective items were coded as missing. Two-dimensional, two-parameter logistic (2 PL) item response models for binary items (Adams & Wu, 2007; Rasch, 1980; Warm, 1989) were applied to estimate the models and the proficiency scores. Item parameters were checked for item conformity. Only items falling within the thresholds of acceptable fit based on mean squared errors (MNSQ) were retained (Adams & Wu, 2002; Wright & Linacre, 1994). To obtain comparable estimates for the differences within cohorts, Year 5, 7, and 9 were scaled simultaneously, resulting in three different estimations of the person parameters.

Internal consistency (standardized Cronbach's alpha of the raw scores) was tested as a baseline to analyze the six tests by domain and by cohort and ranged from $\alpha = 0.69$ to 0.89 (see Table 2). After Items Response Theory (IRT) models were applied, the item parameters and Infit and Outfit values were evaluated. This was done first for the three one-dimensional and second for the three two-dimensional models. Three items from Year 7 and one from Year 9 were excluded due to extreme outfit values in the first step. Additionally, three items were excluded after the second step. The mean Infit of the final two-dimensional 2 PL model within Year 5 was 1.029 ($SD = 0.070$). The mean outfit was 0.0998 ($SD = 0.005$). For the model within Year 7, the mean outfit was 0.998 ($SD = 0.070$), and the mean infit 1.001 ($SD = 0.009$). For Year 9, the mean outfit was 1.188 ($SD = 0.365$), and the mean infit 0.984 ($SD = 0.024$).

The one-dimensional models were calculated to check if the assumption of dimensionality held true. All two-dimensional models proved to fit the data better in likelihood ratio tests and regarding information criteria.

6.2.2. Cognitive abilities

The Figural Analogy subtest of the *Kognitiver Fähigkeitstest* [cognitive functions test] (KFT; Heller & Perleth, 2000) was used to assess cognitive abilities. This test is particularly suited to estimate students' general cognitive abilities independent of their L1 proficiency as it is a non-verbal test. Reliability levels in the remaining Year 9 sample were good and in line with the norm-sample values ($\alpha_{norm} = 0.94$, $\alpha_{LS} = 0.91$, $\alpha_{ES} = 0.89$).

6.2.3. Demographic variables

All demographic variables were based on student and parent responses. Where possible, missing student responses were filled in from parent responses. The scales and categories for cultural capital and family income were adapted from Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS; Bos et al., 2009).

Cultural capital was estimated by asking students how many books their family-owned. The five response categories for cultural capital were 0–10, 11–25, 26–100, 101–200, or more than 200. A visual representation of bookshelves and the corresponding number of books/categories was used to support their response. The families' income was based on parents' responses who responded with one of the following eight categories: below €10,000, €10,000–19,999, €20,000–29,999, €30,000–39,999, €40,000–49,999, €50,000–59.999, €60,000–69,999€, and €70,000 or more. Students also provided their last grade in English on the most recent report card from Year 8. The previous grades ranged from 1 to 6, with 1 representing a very good grade and 6 unsatisfactory.

Table 3
Sociodemographic Characteristics of Students (including missing data).

Baseline characteristic	Year 5			Year 7			Year 9			Full sample		
	n	%	*% missing*	*n*	%	*% missing*	*n*	%	*% missing*	*n*	%	*% missing*
Sex												
female 1	1290	51.2	0.1	1166	52.8	0.9	987	53.6	1.5	1435	51.5	1.5
male 2	1230	48.8		1042	47.2		856	46.4		1350	48.5	
Sex*Cohort												
female Cohort 1	632	51.5	0.2	610	52.9	1.4	518	54.1	2.4	713	51.5	2.3
male Cohort 1	595	48.5		544	47.1		440	45.9		672	48.5	
female Cohort 2	658	50.9	<0.1	556	52.8	0.5	469	53.0	0.4	722	51.6	0.6
male Cohort 2	635	49.1		498	47.2		416	47.0		678	48.4	
L1[a]												
German	1912	76.0		1699	76.7		1421	77.1		2084	74.7	
non-German	605	24.0	0.2	516	23.3	0.6	422	22.9	1.5	705	25.3	1.3
Cohort 1												
German	950	77.4		892	76.8		744	77.4		1048	75.3	
non-German	277	22.6	0.2	269	23.2	0.8	217	22.6	2.1	343	24.7	1.9
Cohort 2												
German	962	74.6		807	76.6		677	76.8		1036	74.1	
non-German	328	25.4	0.2	247	23.4	0.5	205	23.2	0.7	362	25.9	0.8
Cohort [a]												
Cohort 1	1229	48.7		1170	52.5		982	52.5		1418	50.2	
Cohort 2	1293	51.4		1059	47.5		889	47.5		1409	49.8	

[a] Reflects the number and percentage of participants answering whether German was their mother tongue.

6.3. Statistical analyses

To address the research question, a mixed-effects modeling approach was chosen to account for individual differences and random effects (Cunnings, 2012; Linck & Cunnings, 2015; Sahai & Ageel, 2000). Data were analyzed using R (R Core Team, 2018). IRT modeling was done with the TAM software package (Robitzsch et al., 2020). Mixed models were estimated with GAMLj (Gallucci, 2019) a statistics package, using the jamovi software suit (Jamovi, 2019).

The mixed-effects models were estimated by a restricted maximum likelihood approach. Categorical covariates, as well as the cohort variable, were coded as dummy variables. Continuous and ordinal covariates were centered.

For each language domain, two models were estimated. In models 1.1 (listening) and 2.1 (reading), only time, i.e., Year 5, 7, 9, cohort, random intercepts of the individuals and schools were introduced. In models 1.2 (listening) and 2.2 (reading), covariates including sex, L1, cognitive abilities, parental income, and cultural capital (books at home) were added. This stepwise approach was chosen to control for interactions between main effects and effects of covariables.

7. Results

Table 3 describes the sociodemographic background of learners. The composition of both cohorts with regard to sex and L1 was similar for both cohorts (sex: X^2 (1, $N = 2785$) = 0.002, $p = .962$; L1: X^2 (1, $N = 2789$) = 0.564, $p = .453$). The number of non-German students in Year 5 was slightly higher in Cohort 2, but not significantly (X^2 (1, $N = 2517$) = 2.799, $p = .094$). Across time, in both cohorts, slightly more boys and non-L1 German students did not return for assessments. These two groups of students are more likely to have been retained or switched schools due to academic rigor at grammar schools (Krohne and Meier, 2004; Statistisches Bundesamt, 2018). As shown in Table 4, ES (Cohort 2) and LS (Cohort 1) students differed significantly with regard to cognitive abilities and income, while the effect size for the income was negligible. Cultural capital and grades in English at the end of Year 8 did not differ between cohorts.

In a first step we examined the proficiency scores for both cohorts. In Year 5, the ES cohort outperformed the LS cohort in both listening (t (2522) = 7.47, p < .001) and reading (t (2522) = 7.97, p < .001). Two years later in Year 7, the LS cohort caught up with the ES cohort in listening (t (2229) = 1.78, p = .075) and outperformed their peers in reading (t (2229) = 8.98, p < .001). By Year 9, the ES cohort had reconsolidated their strong performance from Year 5 and outperformed their LS peers in both listening (t (1871) = 27.22, p < .001) and reading (t (1871) = 11.07, p < .001).

Correlations between proficiency scores were calculated to assess the connections across time (see Table 5). Tables for each cohort are provided in the Appendix (Tables 10 and 11).

The correlations ranged from small to moderate. As expected, correlations within years are the strongest ($r > 0.4$). The correlations between reading in Year 7 and both domains in Year 9 as well as Listening in Year 5 were especially small ($r < 0.2$). The largest difference between Cohort 1 and Cohort 2 was $\Delta r = 0.06$ (Tables 10 and 11, see Appendix).

In the second step, we conducted mixed-effects models. All mixed-effects models (1.1, 1.2, 2.1, and 2.2) demonstrated significant effects of the cohorts, time, and the interaction of cohort and time could be observed. Time showed a significant negative effect between Year 5 and 7 and a significant positive effect between Year 7 and 9. In all models, the cohort showed a significant interaction with the shift in time. Between models 1.1 and 1.2 (see Table 6) as well as models 2.1 and 2.2 (see Table 7), those interactions could be interpreted as cohort-specific differences between the years that shift due to the introduction of control variables. Regarding the control variables, the L1, cognitive abilities, and English grades significantly affected listening. Better grades and cognitive skills, as well as being a native German speaker, positively predicted listening. Similarly, the cognitive abilities and English grade predicted reading scores. Cultural capital, income, and sex showed no significant mean effects. Interaction terms were added for robustness. For reading, no interaction showed significant effects. For listening, an interaction effect between cohort and L1 as well as cohort and cognitive abilities could be observed. The negative effect of the L1 might indicate a slight closure of the gap between German and non-German students in Cohort 2.

Table 4
Mean comparisons of student background variables and descriptive findings of proficiency scores between cohorts.

Group Descriptives	Cohort 1			Cohort 2			*U/t*	*p*	Cohen's *d*
	n	M	SD	n	M	SD			
cultural capital	1385	3.50	1.159	1403	3.46	1.118	949544	.283	0.033
cognitive abilities	1163	18.14	5.277	1046	16.73	5.353	503142	<.001	0.265
income	1210	5.27	2.205	1244	5.61	2.204	684073	<.001	0.156
grade[a]	931	2.63	0.892	866	2.59	0.928	395321	0.453	0.039
Listening Year 5	1229	485	89.8	1293	514	106.9	7.358	<.001	0.294
Reading Year 5	1229	484	96.7	1293	515	100.8	7.874	<.001	0.314
Listening Year 7	1170	504	97.2	1059	496	102.9	−1.887	.059	0.080
Reading Year 7	1170	518	97.5	1059	480	99.1	−9.118	<.001	0.387
Listening Year 9	982	449	76.8	889	556	92.5	27.313	<.001	1.259
Reading Year 9	982	476	98.4	889	526	95.2	11.147	<.001	0.516

Note. *M* = Mean; *SD* = Standard Deviation; *U* = Mann-Whitney *U* Test statistic; *t* = *t*-test statistic for language proficiency scores.
[a] English grade at the end of Year 8 ranges from 1 (very good) – 6 (unsatisfactory).

Table 5
Correlations for students' proficiencies in English reading and listening.

Variable	1	2	3	4	5	6
1. Listening Year 5 [a]	–					
2. Reading Year 5 [a]	.46***	–				
3. Listening Year 7 [a]	.28***	.31***	–			
4. Reading Year 7 [a]	.18***	.28***	.44***	–		
5. Listening Year 9 [a]	.26***	.31***	.21***	.13***	–	
6. Reading Year 9 [a]	.21***	.25***	.22***	.16***	.46***	–

Note. Pearson Correlation Coefficients were applied.
*$p < .05$. **$p < .01$. ***$p < .001$.
[a] All proficiency scores were estimated as wle-scores. The scores were afterward censored to ± 3 SD. Then the scores were transformed to a mean of 500 and a standard deviation of 100.

The intraclass correlation (*ICC*) within students across time ranged from 24.1% in listening to 25.9% in reading. The school-level showed just 3.6% (reading) to 6.2% (listening) of variance explanation. When controlling for covariates in the second model, the *ICC* diminished, showing just 16.5% at the student level and 3.6% at the school level in listening and 16.2% and 1.8% in reading.

The R^2 values represent the explained variance of the effects of the expected variance of the dependent variable. The marginal R^2 corresponds to the variance explained by the fixed effects. The conditional R^2 denotes the variance explained by the fixed and the random effects. The R^2 values indicated a small to moderate variance explanation (P. D. Ellis, 2010). In summary, the corrected mean values, i.e., mean values, were controlled for background variables derived from the models are shown in Tables 8 and 9.

In conclusion, by comparing the marginalized means, Cohort 1 demonstrated lower proficiency in both language domains in Year 5 but caught up in listening and slightly surpassed Cohort 2 in Year 7. In Year 9, however, the difference from Year 5 was reconsolidated and even expanded. For reading, the difference within Year 5 was roughly re-established in Year 9.

8. Discussion

This study contributes to the growing international literature on EFLL and its long-term impact on proficiency. The present study investigated the impact of starting EFLL in Year 1 in contrast to Year 3 on Year 9 receptive English proficiency. The study assessed receptive language proficiency three times each at the beginning of Years 5, 7, and 9 while controlling for individual learner characteristics. With the bulk of long-term studies finding late starters to be at an advantage in the long run (Baumert et al., 2020; Jaekel et al., 2017; Muñoz, 2006; Pfenninger, 2014), the current study adds starkly contrasting results. The non-linear development between Year 5 and Year 9 was unexpected and raised questions. However, the results also invoke questions regarding how students are received, i.e., how teachers adjusted their lessons to the particular strengths and areas for improvement in secondary education.

The present study demonstrated that ES students outperformed LS students in listening and reading skills in Year 9, after previously falling behind in Year 7. The correlations indicate measurable but only moderate effects within Years and small effects between Years. These findings do not align well with previous studies, which have shown prior L2 achievement to be the strongest predictor of future achievement (Csapo & Nikolov, 2009). The non-linear change in achievement between cohorts was a likely cause for the lower-than-expected correlations. The linear mixed model showed that some individual learner characteristics or their interactions contributed to explaining the assessment outcomes in Year 9. It is important to view the results in the context of previous outcomes. The initial findings in Year 5, when ES outperformed LS, can easily be attributed to the additional 105 h of English classes in elementary school for ES (Jaekel et al., 2017). The Year 5 language proficiency advantage of Year 1 early starters has been replicated four years after the initial study showing a stable advantage for early starters in the new cohort similar to the ES cohort (Jaekel et al., 2022). The proficiency gains of the LS cohort in Year 7 are in line with current research (Muñoz, 2008; Pfenninger, 2014). This dip in proficiency for the ES may also have been caused by the brevity of the additional exposure of 105 lessons in elementary school. We (Jaekel et al., 2017) proposed several hypotheses to explain the proficiency gains for LS in Year 7. The transition to secondary school could have presented different challenges for students of the ES and LS cohort, including the novelty of EFL in Year 1 in the context of the study, secondary teachers' lack of experience with and the availability of adequate materials to support their advanced level of English proficiency.

Furthermore, the late start could have posed advantages for the LS due to their more advanced explicit grammar knowledge in L1, their higher German literacy skills, and further cognitive developments (Cummins & Swain, 1986; Lapkin et al., 1980). Language classrooms were possibly more attuned to the LS age with a stronger focus on explicit language learning, i.e., teachers may not have been ready for students transitioning with a higher language proficiency from elementary school. The potential lack of changes to the Year 5 curriculum may have been unsupportive to the ES and thus may have benefited LS in Year 7 (DeKeyser & Larson-Hall, 2005; Jaekel et al., 2017; Muñoz, 2006). However, little evidence supports the theory of an abrupt shift to explicitly focused teaching strategies (Pfenninger & Singleton, 2019). Recent evidence, however, suggests that the transition to middle school is not smooth for every student (Jaekel et al., 2021).

The new Year 9 results do not align with research comparing early foreign language learners with those starting later. The existing research body has overwhelmingly demonstrated advantages for students starting later who have been shown to make faster progress, catching up quickly with ES and maintaining the advantage until later in their education (Cenoz, 2002; Huang, 2016; Krashen et al., 1979; Muñoz, 2006). Muñoz (2006), in the BAF study, reported that ES students were able to improve their language skills but never

Table 6
Mixed models for listening.

Effect	*Model 1.1*			*Model 1.2*		
	Estimate	*SE*	95% CI	Estimate	*SE*	95% CI
Fixed effects						
Intercept	477.5***	6.36	[465.1; 490]	469.0***	7.77	[453.8; 484.2]
Cohort	42.0***	2.73	[36.6; 47.3]	64.2***	8.92	[46.7; 81.7]
time_L5 [a]	−19.1***	3.40	[-25.7; −12.4]	19.7	11.11	[-2.1; 41.4]
time_L7 [b]	54.4***	3.62	[47.3; 61.5]	−34.6**	11.13	[-56.4; −12.8]
cultural capital [c]				2.3	1.59	[-0.8; 5.4]
cognitive abilities [c]				1.7***	0.43	[0.9; 2.6]
income [c]				0.6	0.81	[-1.0; 2.2]
grade [c]				−19.5***	1.78	[-23.0; −16.0]
L1 [d]				22.2**	7.34	[7.8; 36.6]
sex [e]				6.0	9.97	[-13.6; 25.5]
Cohort * time_L5 [a]	37.8***	4.84	[28.3; 47.3]	−36.1**	14.44	[-64.4; −7.8]
Cohort * time_L7 [b]	−114.4***	5.26	[-124.7; −104.1]	69.7***	14.37	[41.5; 97.9]
Cohort * L1 [d]				−21.8**	10.23	[-41.9; −1.8]
time_L5 [a] *L1 [d]				12.6	12.32	[-11.5; 36.8]
time_L7 [b] *L1 [d]				7.3	12.32	[-16.9; 31.4]
Cohort * sex [e]				−16.6	14.24	[-44.5; 11.3]
time_L5 [a] * sex				3.3	15.28	[-26.7; 33.2]
time_L7 [b] * sex [e]				2.2	15.13	[-27.4; 31.9]
L1 [d] * sex [e]				−19.3	11.13	[-41.1; 2.5]
Cohort * cognitive abilities				1.6**	0.62	[0.4; 2.8]
Cohort * time_L5 [a] *L1 [d]				−5.3	15.33	[-35.3; 24.8]
Cohort * time_L7 [b] *L1 [d]				4.9	15.23	[-24.9; 34.8]
Cohort * time_L5 [a] * sex [e]				−11.6	12.47	[-36.0; 12.9]
Cohort * time_L7 [b] * sex [e]				−9.8	12.43	[-34.2; 14.5]
Cohort * L1 [d] * sex [e]				21.1	15.94	[-10.1; 52.4]
time_L5 [a] *L1 [d] * sex [e]				−12.9	15.61	[-43.5; 17.7]
time_L7 [b] *L1 [d] * sex [e]				−8.6	15.47	[-38.9; 21.7]
	Variance	*SD*	*ICC*	Variance	*SD*	*ICC*
Random Effects						
Student Intercept	2122	46.1	.241	1273	35.7	.165
School Intercept	440	21.0	.062	241	15.5	.036
Residual	6684	81.8		6428	80.2	
$N_{Observations}$	6606			4081		
$N_{Students}$	2811			1393		
$N_{Schools}$	12			12		
Model Fit						
AIC	78513.26			48060.36		
BIC	78548.15			48116.78		
Marginal R^2	0.086			0.180		
Conditional R^2	0.339			0.337		

Note. REML estimate. CI = confidence interval; *LL* = lower limit; *UL* = upper limit; *SD* = Standard Deviation; *ICC* = Intra-Class-Correlation; *AIC* = Akaike Information Criterion; *BIC* = Bayes Information Criterion; Marginal R^2 = Variance explained by fixed effects; Conditional R^2 = Variance explained by fixed and random effects.

[a] Time_L5 = Listening Score in Year 5 - Listening Score in Year 7.

[b] Time_L7 = Listening Score in Year 7 - Listening Score in Year 9.

[c] Cultural Capital, Cognitive Abilities, Household Income and Last Grade (Year 8) were grand centered.

[d] 0 = not-german, 1 = german.

[e] 0 = female, 1 = male.

fully caught up to or surpassed LS after 726 h of instruction. The hours spent in English classes for the BAF study lie between those of the ES and LS cohorts in this study. In contrast to the BAF study, the ES cohort surpassed the LS students after 710 h of English instruction. To a limited extent, our data confirm the BAF study in that ES students improved their proficiency after the dip in Year 7. However, surpassing their LS peers is difficult to explain without further data. Motivational issues, changes in the methodology, and increased teacher experience after several years of working with ES students could explain some of the gains, but further research through follow-up studies needs to shed more light on this development. Baumert et al. (2020) investigated German Year 9 students' receptive skills of early, middle, and late starters but could not find any discernible differences in Year 9 regarding the onset after controlling for learner characteristics and the educational context. However, the study only considered cross-sectional data in Year 9 and is missing the crucial middle school years link. This limits comparability with the current data as this was when ES students scored lower than their LS peers. Nevertheless, as our study was conducted in the same country, the results are intriguing as they demonstrate an advantage for an early start, although later on.

The non-linear, U-shaped development for both the ES and LS cohorts raises several questions, considering the existing research

Table 7
Mixed model for reading.

Effect	*Model 2.1*			*Model 2.2*		
	Estimate	*SE*	95% CI	Estimate	*SE*	95% CI
Fixed effects						
Intercept	491.9***	5.07	[482.0; 501.8]	498.3***	7.27	[484.0; 512.5]
Cohort	12.9***	2.85	[7.33; 18.5]	21.9**	9.17	[3.9; 39.9]
time_R5 [a]	−33.5***	3.48	[-40.29; −26.6]	−41.8**	11.47	[-64.2; 19.3]
time_R7 [b]	41.7***	3.71	[34.38; 48.9]	43.0***	11.14	[21.2; 64.8]
cultural capital [c]				2.4	1.63	[-0.8; 5.6]
cognitive abilities [c]				2.3***	0.44	[1.4; 3.2]
income [c]				0.7	0.83	[-0.9; 2.3]
grade [c]				−23.0***	1.83	[-26.6; −19.4]
L1 [d]				3.5	7.53	[-11.3; 18.3]
sex [e]				−5.3	10.25	[-25.4; 14.7]
Cohort * time_R5 [a]	68.9***	4.96	[59.18; 78.6]	78.1***	14.92	[48.9; 107.4]
Cohort * time_R7 [b]	−86.7***	5.39	[-97.23; −76.1]	−76.4***	14.67	[-105.1; −47.6]
Cohort * L1 [d]				−5.7	10.51	[-26.3; 14.9]
time_R5 [a] *L1 [d]				−0.3	12.73	[-25.2; 24.7]
time_R7 [b] *L1 [d]				1.7	12.43	[-22.7; 26.0]
Cohort * sex [e]				−3.9	14.63	[-32.6; 24.8]
time_R5 [a] * sex				23.1	15.79	[-7.8; 54.1]
time_R7 [b] * sex [e]				4.2	15.53	[-26.3; 34.6]
L1 [d] * sex [e]				−8.9	11.43	[-31.3; 13.5]
Cohort * cognitive abilities				0.8	0.63	[-0.4; 2.0]
Cohort * time_R5 [a] *L1 [d]				−3.7	15.84	[-34.8; 27.3]
Cohort * time_R7 [b] *L1 [d]				−5.0	15.61	[-35.6; 25.5]
Cohort * time_R5 [a] * sex [e]				5.4	12.88	[-19.8; 30.7]
Cohort * time_R7 [b] * sex [e]				−7.3	12.76	[-32.3; 17.7]
Cohort * L1 [d] * sex [e]				12.7	16.38	[-19.5; 44.8]
Time_R5 [a] *L1 [d] * sex [e]				−10.9	16.12	[-42.5; 20.7]
Time_R7 [b] *L1 [d] * sex [e]				−16.8	15.90	[-47.9; 14.4]
	Variance	*SD*	*ICC*	Variance	*SD*	*ICC*
Random Effects						
Student Intercept	2445	49.4	.259	1324	36.4	.162
School Intercept	259	16.1	.036	125.0	11.2	.018
Residual	7011	83.7		6861	82.8	
$N_{Observations}$	6606			4081		
$N_{Students}$	2811			1393		
$N_{Schools}$	12			12		
Model Fit						
AIC	78933.49			48306.26		
BIC	78968.59			48361.70		
Marginal R^2	0.037			0.132		
Conditional R^2	0.305			0.283		

Note. REML estimate. CI = confidence interval; *LL* = lower limit; *UL* = upper limit; *SD* = Standard Deviation; *ICC* = Intra-Class-Correlation; *AIC* = Akaike Information Criterion; *BIC* = Bayes Information Criterion; Marginal R^2 = Variance explained by fixed effects; Conditional R^2 = Variance explained by fixed and random effects.
*$p < .05$. **$p < .01$. ***$p < .001$.
[a] Time_R5 = Reading Score in Year 5 – Reading Score in Year 7.
[b] Time_R7 = Reading Score in Year 7 – Reading Score in Year 9.
[c] Cultural Capital, Cognitive Abilities, Household Income and Last Grade (Year 8) were grand centered.
[d] 0 = not-german, 1 = german.
[e] 0 = female, 1 = male.

Table 8
Marginalized means for the effect of time by cohort in listening.

Cohort	Listening	Mean	SE	df	95% CI	
					Lower	Upper
1	Year 5	482	6.22	29.3	469	494
2	Year 5	523	6.29	30.5	510	536
1	Year 7	505	6.17	28.4	492	518
2	Year 7	503	6.29	30.4	491	516
1	Year 9	448	6.13	27.7	436	461
2	Year 9	559	6.25	29.7	546	572

Note. Means were estimated averaging across interacting variables.

Table 9
Marginalized means for the effect of time by cohort in reading.

Cohort	Reading	Mean	SE	df	95% CI	
					Lower	Upper
1	Year 5	487	5.48	51.4	476	498
2	Year 5	532	5.56	53.6	521	544
1	Year 7	520	5.41	48.6	509	531
2	Year 7	487	5.55	53.4	475	498
1	Year 9	478	5.37	47.3	468	489
2	Year 9	527	5.51	51.7	516	538

Note. Means were estimated averaging across interacting variables.

body. The interaction effects can be explained by comparing the respective differences in the means. For listening, the interaction effect of L1 and cohort is due to a mean difference within cohorts where non-German students performed lower than their German peers. These effects cannot be found for reading, where both groups perform at a similar level. The interaction effect of cognitive abilities and cohort can be attributed to the significant differences in the cohorts' performances in the cognitive abilities assessment. Concerning the main research focus, a non-linear downturn points to significant changes across schools, possibly systemic issues, and differences between cohorts in the case of the current study. Yet, no significant systemic changes were implemented in the school system during the assessment that would not have impacted both cohorts. All participating schools started implementing all-day, i.e., extended education activities, clubs, remedial courses. Unfortunately, no data was available to determine the offerings focused on English. A potential impact on students' English outcomes cannot be ruled out.

The results for Year 9, however, raise several questions with practical and theoretical implications. Why did ES outperform LS in Year 9 after the Year 7 assessment showed advantages for the LS cohort? Which changes occurred between Years 7 and 9 that are different from previous studies with different outcomes? Did ES develop advantages from learning English more implicitly in elementary school that benefit them later in Year 9? Unfortunately, the questions raised cannot be answered with the current dataset, and more research is necessary to answer these questions adequately. It is possible that English teachers started adapting their lessons better to the younger learners or that more support structures were in place for the extended education programs in the afternoon that facilitated the ES cohort's skill growth. While potential advantages from learning a language implicitly for a longer time in elementary school are conceivable, research specifically focused on implicit versus explicit language learning across elementary and secondary schools is needed to confirm this hypothesis. The data suggest that we have to be careful about closing the book on EFLL and its long-term implications without thoroughly investigating the secondary teaching process. The results presented by Jaekel et al. (2017), which the present data confirm, demonstrate a significant surge in proficiency for late starters from Year 5 to Year 7. EFLL programs have been shown to improve slightly, i.e., increase students' proficiency at the end of elementary school, over time with more experienced teachers (Jaekel et al., 2022). The transition may play a significant role in causing the non-linear development; however, the collected data does not provide information on the transition. The transition to secondary school for EFLL, to a large extent, remains a black box.

From a statistical standpoint, the results were stable across the sample, and all twelve participating schools, situated in different socioeconomic and urban vs. rural contexts, showed a similar development following the Year 7 assessment. Furthermore, the results showed that Year 5 proficiency levels were stronger predictors for Year 9 outcomes.

The results of this study are relevant to the growing international discourse on EFLL. Particularly as many countries have adopted EFLL programs in elementary education often without extensively considering available SLA research, for example, about increased exposure to the L2 for younger learners (Larson-Hall, 2008) or the impact of teachers' L2 language proficiency (Unsworth et al., 2015). Similarly, the long-term efficacy of EFLL remains under-researched, and the present study adds new findings to this limited field.

In light of the presented results, educators and policymakers should avoid making hasty decisions on EFLL policies and curricula changes without more research. Systemic changes take time to implement. Considering unresolved questions regarding the transition from elementary to secondary school, more evidence is needed to make definitive statements on EFLL (Courtney, 2017; Jaekel et al., 2017; Pfenninger & Lendl, 2017). Teachers should incorporate diagnostic assessments for students transitioning to secondary education in their classes to gauge students' proficiency for all language skills. In contexts where the transition from elementary and secondary school is accompanied by a shift from implicit to explicit language learning, it may be beneficial to also gain an understanding of the level of explicit grammatical knowledge and meta-language students already possess. These assessments allow teachers to adjust their curriculum according to each class language profile and differentiate tasks within the class. Particularly in contexts where foreign language education is textbook-based, adjusting or supplementing materials and exercises may be necessary. Where possible, collaborations between elementary and secondary school teachers can also provide additional information on skill level, teaching methodology in elementary schools and facilitate a smoother transition for students.

The explained variance was lower than expected, considering that multiple important indicators have been included in models 1.2 and 2.2. Even in models 1.1 and 2.1, the explained variance should have been higher. However, the previous measures were taken in two-year intervals, which reduced prior learnings' predictive power.

9. Strengths and limitations

The large sample size covering rural and urban settings and follow-up assessments covering five years provide a robust foundation for statistical analyses. Especially the opportunity to assess comparable (though not linked) estimators for the success in SLA is not common for five years in one cohort and three years and a follow-up study in another cohort. While all test items were identical for both cohorts, test days varied. The multi-year, multi-subject *Ganz In* project that this study was part of did not conduct a full-project Year 9 assessment for the ES cohort. While the LS cohorts' English assessment in Year 9 was part of a one-day, multi-assessment battery with randomized testing times, the ES cohort's assessment only lasted 60 min. Due to the reduced cognitive load and potential effects of students getting tired, the difference in test intensity may have had a small impact on the data, i.e., the language proficiency outcome.

The change in assessment structures for the last Year 9 assessment is one significant difference between the two cohorts. While test theory cautions to assure consistent assessment conditions (Shadish et al., 2002), the context of the overarching project did not conduct another assessment and, as such, could not be controlled. The impact of the difference in test day schedule and cognitive load should not have a large effect on overall test scores. A similar contextual difference was part of a replication study for Year 5 with a third cohort starting in Year 1 (Jaekel et al., 2022). The data in this study suggests that the long test day has a negligible impact on language proficiency outcomes and thus should not be the cause for the rapid increase in language proficiency by the ES cohort. Generally, large-scale assessments with prolonged periods focused on testing may not only cause fatigue in participants, but they raise concerns regarding research ethics. It is crucial to balance the time and cost constraints of large-scale, standardized research, school scheduling, and the needs and respect for participants. Under ideal conditions, assessments for research would be incorporated into school testing. However, that may be rather impractical with many teachers and schools involved. Instead of isolated whole-day testing spreading it out across a week or two with fewer tests each day may be more ethical.

Age of onset (AO) and the amount of exposure are conflated in this study and, as such, cannot be disentangled (Lambelet & Berthele, 2015). Considering the small gap of only 1 ½ years between cohorts, the additional exposure is likely to have had a greater effect on outcomes. However, lacking prior research, the authors assume that AO and the amount of exposure cannot be disentangled.

The study includes manifold variables covering individual and contextual factors. However, out-of-school technology use, including the ownership of e-books, was not assessed. Therefore, these factors were not considered in our analyses.

School attrition was another issue with the second Year 9 assessment. Due to the voluntary nature of the study, 19 grammar schools did not participate in the Year 9 assessment for the ES cohort. This reduced the overall sample size considerably, leaving 12 schools that participated. However, an overall large sample size remained, and descriptive characteristics in Years 5 and 7 were the same as in previous publications based on the larger sample. Additionally, students were excluded from analyses if they did not participate in all assessments to avoid imputing outcome variables.

10. Conclusion

Early starters outperformed late starters in the EFL listening assessment in Year 9, making up the assessment gap from the Year 7 assessment. The results raise several questions that the collected data fails to explain. However, while not supported by extensive international research on EFLL, the results show that there were long-term benefits to EFLL in this study.

Evidently, in the context of the existing research body, these new results call for carefully designed, longitudinal research in various educational settings to better understand the long-term implications of EFLL. Future studies should carefully follow students' proficiency development longitudinally, including speaking and writing competence, at least from the beginning of secondary school, and, if possible, start in elementary school to investigate the transition, which many studies identify as a likely obstacle. Studies on secondary foreign language teachers' pedagogical approaches to addressing students' existing L2 strengths and weaknesses could also shed more light on their role in the transition process. There may be other contextual or systemic factors, curricular changes, motivational issues, and other factors that have not been identified yet. Considering the international push for EFLL and that most recent longitudinal studies on EFLL have been conducted in Europe, other research contexts would enlighten the field and could provide new perspectives.

Language learners are best served by schools and districts that facilitate cooperation across elementary and secondary schools. Early diagnostic assessments after the transition should inform teachers how to best support their students. Rigid curriculum progression neglects the strengths or areas of growth of classes. Research-based continuous professional development for language teachers is essential, particularly in a quickly developing field such as EFLL.

EFLL needs also to be held accountable and show sufficient progress across the transition to justify the curricular time allocated to language learning. However, we want to emphasize that the value of EFLL should not be solely determined by students' level of language proficiency alone but must also consider the impact language education has on intercultural understanding and the development of attitudes concerning the target language, other cultures, and language learning in general.

Funding

This research was in part funded by the Mercator Stiftung.

Appendices.

Table 10
Correlations for Students' Proficiencies in English Reading and Listening within Cohort 1

Variable	1	2	3	4	5	6
1. Listening Year 5 [a]	–					
2. Reading Year 5 [a]	.47***	–				
3. Listening Year 7 [a]	.30***	.33***	–			
4. Reading Year 7 [a]	.22***	.32***	.41***	–		
5. Listening Year 9 [a]	.23***	.26***	.27***	.26***	–	
6. Reading Year 9 [a]	.17***	.20***	.25***	.28***	.43***	–

Note. Pearson Correlation Coefficients were applied.
*$p < .05$. **$p < .01$. ***$p < .001$.
[a] All proficiency scores were estimated as wle-scores. The scores were afterward censored to ± 3 SD. Then the scores were transformed to a mean of 500 and a standard deviation of 100.

Table 11
Correlations for Students' Proficiencies in English Reading and Listening within Cohort 2

Variable	1	2	3	4	5	6
1. Listening Year 5 [a]	–					
2. Reading Year 5 [a]	.43***	–				
3. Listening Year 7 [a]	.28***	.32***	–			
4. Reading Year 7 [a]	.22***	.33***	.47***	–		
5. Listening Year 9 [a]	.19***	.24***	.27***	.27***	–	
6. Reading Year 9 [a]	.18***	.23***	.21***	.18***	.37***	–

Note. Pearson Correlation Coefficients were applied.
*$p < .05$. **$p < .01$. ***$p < .001$.
[a] All proficiency scores were estimated as wle-scores. The scores were afterward censored to ± 3 SD. Then the scores were transformed to a mean of 500 and a standard deviation of 100.

References

Jamovi. (2019). The jamovi project (Version 1.1) [Computer software] https://www.jamovi.org.

Adams, R., & Wu, M. (2002). *PISA 2000 technical report. OECD*.

Adams, R., & Wu, M. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In C. H. Carstensen, & M. von Davier (Eds.), *Statistics for social and behavioral sciences. Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 55–76). Springer.

Anderson, L. W., Jacobs, J., Schramm, S., & Splittgerber, F. (2000). School transitions: Beginning of the end or a new beginning? *International Journal of Educational Research, 33*(4), 325–339. https://doi.org/10.1016/S0883-0355(00)00020-3

Barucki, H., Bliesener, U., Boerner, O., Böttger, H., Hoffmann, I.-B., Kierepka, A., Kronisch, I., Legutke, M., Lohmann, C., Mueller, T., & Schlueter, N. (2015). Der Lernstand im Englischunterricht am Ende von Klasse 4: Ergebnisse der BIG-Studie. *Domino-Verl*.

Baumert, J., Fleckenstein, J., Leucht, M., Köller, O., & Möller, J. (2020). The long-term proficiency of early, middle, and late starters learning English as a foreign language at school: A narrative review and empirical study. *Language Learning. Advance online publication*, 1091–1135. https://doi.org/10.1111/lang.12414

Bolton, K., & Graddol, D. (2012). English in China today: The current popularity of English in China is unprecedented, and has been fuelled by the recent political and social development of Chinese society. *English Today, 28*(3), 3–9. https://doi.org/10.1017/S0266078412000223

HANSE - Hamburger Schriften zur Qualität im Bildungswesen. In Bos, W. (Ed.), *KESS 4-Lehr- und Lernbedingungen in Hamburger Grundschulen* (vol. 2), (2007). Waxmann http://fox.leuphana.de/portal/de/publications/kess-4-lehr-und-lernbedingungen-in-hamburger-grundschulen(af3fbf63-1bcb-4e30-ac3f-a794f040c82d).html.

Bos, W., Bonsen, M., Gröhlich, C., Guill, K., & Scharenberg, K. (2009). KESS 7: Skalenhandbuch zur Dokumentation der Erhebungsinstrumente. *HANSE - Hamburger Schriften zur Qualität im Bildungswesen, 4* (Waxmann).

de Bot, K. (2014). The effectiveness of early foreign language learning in The Netherlands. *Studies In Second Language Learning And Teaching, 3*, 409–418. https://doi.org/10.14746/ssllt.2014.4.3.2

Statistische Bundesamt. (2018). Schulen auf einen Blick: Ausgabe 2018. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Schulen/Publikationen/Downloads-Schulen/broschuere-schulen-blick-0110018189004.pdf?_blob=publicationFile.

Butler, Y. G. (2014). Current issues in English education for young learners in East Asia. *English teaching, 69*(4), 3–25. https://doi.org/10.15858/engtea.69.4.201412.3

Cenoz, J. (2002). Age differences in foreign language learning. *ITL - International Journal of Applied Linguistics, 135–136*, 125–142. https://doi.org/10.1075/itl.135-136.06cen

Council of Europe. (2005). *Language education policy profile*. Strasbourg https://rm.coe.int/language-education-policy-profile-luxembourg/16807b3c27.

Council of the European Union. (1997). Council Resolution on the early teaching of European Union languages. Council of the European Union. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31998Y0103(01)&from=EN.

Courtney, L. (2017). Transition in modern foreign languages: A longitudinal study of motivation for language learning and second language proficiency. *Oxford Review of Education, 43*(4), 462–481. https://doi.org/10.1080/03054985.2017.1329721

Cronquist, K., & Fiszbein, A. (2017). English Language Learning in Latin America. The Dialogue - Leadership for the Americas. https://www.thedialogue.org/wp-content/uploads/2017/09/English-Language-Learning-in-Latin-America-Final-1.pdf.

Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences, 19*(2), 209–218. https://doi.org/10.1016/j.lindif.2009.01.002

Cummins, J., & Swain, M. (1986). *Bilingualism in education: Aspects of theory, research and practice. Applied linguistics and language study*. Routledge.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*(3), 369–382. https://doi.org/10.1177/0267658312443651
DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll, & A. M. B. Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 88–108). Oxford Univ. Press [Jenifer].
Demircioglu, J. (2008). Englisch in der Grundschule - Auswirkungen auf Leistungen und Selbstbewertung in der weiterführenden Schule. *Lebenslang lernen: Brain & Development, 7* (Logos- Verl).
Edelenbos, P., Johnstone, R., & Kubanek, A. (2006). The main pedagogical principles underlying the teaching of languages to very young learners: Languages for the children of Europe: Published Research, Good Practice & Main Principles [Final Report of the EAC 89/04, Lot 1 study]. *European Commission*. http://ec.europa.eu/languages/policy/language-policy/documents/young_de.pdf.
Education Audiovisual & Culture Executive Agency. (2017). Key data on teaching languages at school in Europe: Eurydice Report (2017 edition). *Publications Office of the European Union*.
Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results (9th print)*. Cambridge Univ. Press.
Ellis, N. (2017). Implicit and explicit knowledge about language. In J. Cenoz, D. Gorter, & S. May (Eds.), *Language awareness and multilingualism* (pp. 113–124). Springer International Publishing. https://doi.org/10.1007/978-3-319-02240-6_7.
Enever, J. (2011). ELLiE: Early language learning in Europe. *British Council*.
Engel, G. (2009). Evening: Konsequenzen für die Weiterentwicklung des Englischunterrichts in der Grundschule. In G. Engel, B. Groot-Wilken, & E. Thürmann (Eds.), *Englisch in der Primarstufe - chancen und Herausforderungen: Evaluation und Erfahrungen aus der Praxis* (1st, pp. 197–215). (Cornelsen).
Engel, G., & Ehlers, G. (2013). Hören – Zuhören – Verstehen. Möglichkeiten der Analyse, Diagnose und gezielten Förderung des Hörverstehens. In O. Börner, G. Engel, & B. Groot-Wilken (Eds.), *Hörverstehen, Leseverstehen, Sprechen: Diagnose und Förderung von sprachlichen Kompetenzen im Englischunterricht der Primarstufe* (1st, pp. 44–69). Waxmann.
European Commission. (2011). Language learning at pre-primary school level: Making it efficient and sustainable. *A policy handbook*. http://ec.europa.eu/languages/policy/language-policy/documents/early-language-learning-handbook_en.pdf.
Eurostat. (2019). Foreign language learning statistics. *Eurostat*. https://ec.europa.eu/eurostat/statistics-explained/index.php/Foreign_language_learning_statistics#Primary_education.
Gallucci, M. (2019). GAMLj [Computer software]. https://gamlj.github.io/.
García Mayo, M. (2003). 5. Age, length of exposure and grammaticality judgements in the acquisition of English as a foreign language. In M. García Mayo, & M. García Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 94–114). Bristol, Blue Ridge Summit: Multilingual Matters. https://doi.org/10.21832/9781853596407-006.
Graham, S., Courtney, L., Tonkyn, A., & Marinis, T. (2016). Motivational trajectories for early language learning across the primary-secondary school transition. *British Educational Research Journal, 42*(4), 682–702. https://doi.org/10.1002/berj.3230
Grotjahn, R. (2005). Je früher, desto besser? Neuere Befunde zum Einfluss des Faktors „Alter" auf das FremdsprachenlernenH. Pürschel, & T. Tinnefeld (Eds.). *Reihe: Fremdsprachen in Lehre und Forschung (FLF): Bd. 38. Moderner Fremdsprachenerwerb zwischen Interkulturalität und Multimedia: Reflexionen und Anregungen aus Wissenschaft und Praxis*, 186–202 (AKS-Verl).
Harley, B. (1986). Age in second language acquisition. In *Multilingual matters, 22*Multilingual Matters.
Heller, K. A., & Perleth, C. (2000). KFT 4-12+R - Kognitiver Fähigkeits-Test für 4. bis 12. *Klassen* (Revision. Beltz).
Holmstrand, L. (1982). *English in the elementary school: Theoretical and empirical aspects of the early teaching of English as a foreign language. Uppsala studies in education* (vol. 18). Univ.; Almqvist & Wiksell international (distr.).
Huang, B. H. (2016). A synthesis of empirical research on the linguistic outcomes of early foreign language instruction. *International Journal of Multilingualism, 13*(3), 257–273. https://doi.org/10.1080/14790718.2015.1066792
Institut zur Qualitätssicherung im Bildungswesen. (2014). VERA – Ein Überblick. https://www.iqb.hu-berlin.de/vera.
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: From early starters to late finishers? *Language Learning, 67*(3), 631–664. https://doi.org/10.1111/lang.12242
Jaekel, N., van Ackern, I., Schurig, M., & Ritter, M. (2021). Investigating the transition from elementary to secondary school: Student perspectives of the continuum and receptive language proficiency. *The Language Learning Journal, 0*(0), 1–16. https://doi.org/10.1080/09571736.2021.1879906
Jaekel, N., Schurig, M., & Ritter, M. (2022). *Investigating cohort effects of early foreign language learning*.
Kersten, K., & Rohde, A. (2013). On the road to nowhere? The transition problem of bilingual teaching programs. In D. Elsner, & J.-U. Kessler (Eds.), *Narr studienbücher. Bilingual education in primary school: Aspects of immersion, CLIL, and bilingual modules* (pp. 93–134) (Narr Francke Attempto GmbH).
Kohler, M. (2017). Review of languages education policies in Australia. the Multicultural Education and Languages Committee (MELC). https://www.education.sa.gov.au/sites/default/files/melc-review-of-languages-education-policies-australia.pdf?acsf_files_redirect.
Kolb, A. (2011). Kontinuität und Brüche: Der Übergang von der Primar- zur Sekundarstufe im Englischunterricht aus der Perspektive der Lehrkräfte. *Zeitschrift Für Fremdsprachenforschung, 22*(2), 145–175.
Krashen, S. D., Long, M. A., & Scarcella, R. C. (1979). Age, rate and Eventual attainment in second language acquisition. *TESOL Quarterly, 13*(4), 573. https://doi.org/10.2307/3586451
Krohne, J. A., Meier, U., & Tillmann, K.-J. (2004). Sitzenbleiben, Geschlecht und Migration - Klassenwiederholungen im Spiegel der PISA-Daten. *Zeitschrift für Padagogik, 50*(3), 373–391. https://www.pedocs.de/volltexte/2011/4816/pdf/ZfPaed_2004_3_Krohne_Meier_Tillmann_Sitzenbleiben_Geschlecht_Migration_D_A.pdf.
Lambelet, A., & Berthele, R. (2015). Age and Foreign Language Learning in School. Palgrave Macmillan UK. https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=4001237.
Lapkin, S., & Swain, M. (1982). *Evaluating bilingual education: A Canadian case study. Multilingual matters (multilingual matters, clevedon): Vol. 2*. Multilingual Matters.
Lapkin, S., Swain, M., Kamin, J., & Hanna, G. (1980). *Report on the 1979 evaluation of the Peel County late French immersion program, grades 8, 9, 10, 11 and 12*. Toronto: Ontario Institute for Studies in Education.
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*(1), 35–63. https://doi.org/10.1177/0267658307082981
Lichtman, K. (2016). Age and learning environment: Are children implicit second language learners? *Journal of Child Language, 43*(3), 707–730. https://doi.org/10.1017/S0305000915000598
Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185–207. https://doi.org/10.1111/lang.12117
MSW - NRW. (2008). Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen: Deutsch, Sachunterricht, Mathematik, Englisch, Musik, Kunst, Sport, Evangelische Religionslehre, Katholische Religionslehre. *Schule in NRW, 2012* (Ritterbach).
Muñoz, C. (2006). *Age and the rate of foreign language learning. Second Language Acquisition* (vol. 19). Multilingual Matters.
Muñoz, C. (2008). Age-related differences in foreign language learning. *Revisiting the empirical evidence. IRAL - International Review of Applied Linguistics in Language Teaching, 46*(3). https://doi.org/10.1515/IRAL.2008.009
Oller, J. W., & Nagato, N. (1974). The long-term effect of FLES: An experiment. *The Modern Language Journal, 58*(1/2), 15. https://doi.org/10.2307/323984
Pfenninger, S. E. (2014). The misunderstood variable: Age effects as a function of type of instruction. *Studies in Second Language Learning and Teaching, 3*, 529–556. https://doi.org/10.14746/ssllt.2014.4.3.8
Pfenninger, S. E., & Lendl, J. (2017). Transitional woes: On the impact of L2 input continuity from primary to secondary school. *Studies in Second Language Learning and Teaching, 7*(3), 443–469. https://doi.org/10.14746/ssllt.2017.7.3.5
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*(3), 207–234. https://doi.org/10.1111/lang.12318

Piske, T. (2013). Frühbeginn allein ist nicht genug.: Welchen Einfluss haben Faktoren wie Alter, sprachlicher Input, Geschlecht und Motivation auf die Aussprechentwicklung und die grammatischen Kenntnisse von Zweitsprachenlernern? In C. Bürgel, & D. Siepmann (Eds.), *Thema Sprache - wissenschaft für den Unterricht: Vol. 6. Sprachwissenschaft - fremdsprachendidaktik: neue Impulse* (pp. 117–144) (Schneider Verl. Hohengehren).

Piske, T. (2017). The earlier the better? Some critical remarks on current EFL teaching to young learners and their implications for foreign language teacher education. In E. Wilden, & R. Porsch (Eds.), *The professional development of primary EFL teachers: National and international research* (pp. 45–57) (Waxmann).

R Core Team. (2018). *R* [Computer software]. https://cran.r-project.org/.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Univ. of Chicago Pr (Expanded).

Robitzsch, A., Kiefer, T., & Wu, M. (2020). TAM (Version R package version 3.5-19) [Computer software]. https://CRAN.R-project.org/package=TAM.

Roos, J. (2007). Spracherwerb und Sprachproduktion: Lernziele und Lernergebnisse im Englischunterricht der Grundschule. *Giessener Beiträge zur Fremdsprachendidaktik* (Narr).

Sahai, H., & Ageel, M. I. (2000). The analysis of variance: Fixed, random, and mixed models. *Birkhäuser*.

Schmelter, L. (2010). (K)eine Frage des Alters - Fremdsprachenunterricht auf der Primarstufe. *Zeitschrift für Interkulturellen Fremdsprachenunterricht, 15*(1), 26–41.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin. http://www.loc.gov/catdir/enhancements/fy1105/2001131551-d.html.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Wesche, M. (2002). Early French Immersion: How has the original Canadian model stood the test of time? In P. Burmeister, T. Piske, & A. Rohde (Eds.), *An Integrated view of language development: Papers in honor of Henning Wode* (pp. 357–379) (Wissenschaftlicher Verlag Trier).

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: Springer.

Wilden, E., Porsch, R., & Ritter, M. (2013). Je früher desto besser? Frühbeginnender Englischunterricht ab Klasse 1 oder 3 und seine Auswirkungen auf das Hör- und Leseverstehen. *Zeitschrift Für Fremdsprachenforschung, 24*(2), 171–201.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. www.rasch.org/rmt/rmt83b.htm.