This postprint was originally published by Elsevier as:
Preusler, S., Zitzmann, S., Baumert, J., & Möller, J. (2022). Development of German reading comprehension in two-way immersive primary schools. *Learning and Instruction, 79*, Article 101598. https://doi.org/10.1016/j.learninstruc.2022.101598.

Development of German reading comprehension in two-way immersive primary schools

Sandra Preusler[a,*], Steffen Zitzmann[b], Jürgen Baumert[c], Jens Möller[a]

[a] *Institute for Psychology of Learning and Instruction at Kiel University, Germany*
[b] *Hector Research Institute of Education Sciences and Psychology at University of Tübingen, Germany*
[c] *Max Planck Institute for Human Development, Berlin, Germany*
* Corresponding author. Institute for Psychology of Learning and Instruction at Kiel University, Olshausenstraße 75, 24118, Kiel, Germany.
E-mail address: spreusler@ipl.uni-kiel.de

Abstract

Two-way immersion (TWI) is a variant of the increasingly popular bilingual instruction. Most TWI research lacks longitudinal data or the consideration of background variables to control for possible selection effects. This article examines the development of German reading comprehension of TWI students (N = 984) from fourth to sixth grade compared to conventionally taught students (N = 992). The latent growth curve models showed that immersion students reached the same level of German reading comprehension over the three measurement points, even if background variables like first language, socioeconomic background, and cognitive ability were included. Despite reduced instruction in German, TWI students showed the same reading comprehension level as students in regular instruction while having the advantage of learning an additional language. Although the level of reading comprehension differed between language groups (L1 German speakers, L1 partner language speakers, simultaneous bilinguals), the learning trajectories of reading comprehension were similar.

Article Info

Keywords:
Two-way immersion
Bilingual education
Dual language immersion
Reading comprehension
Language proficiency
Language education

1. Introduction

Bilingual school programs are becoming increasingly popular. They offer the chance to learn another language besides the majority language. One promising concept is two-way immersion (TWI). TWI is one of the most widespread forms of bilingual instruction among other forms, such as Content and Language Integrated Learning (CLIL)—a European approach, usually a form of foreign language teaching (Coyle, 2007)—and one-way immersion (Genesee, 1981). TWI, also called dual language education or bilingual immersion, is a dual language program in which majority- and minority-language speakers were originally intended to be approximately equally represented in the classroom, and instruction is given in both languages (Baker, 2011; Lindholm-Leary, 2001). Today, with the increasing diversification of the student body, there are also many bilingually raised students who make up the majority of the student body in some programs as well as some students who do not speak any of the languages as a first language (Baumert, Hohenstein, Fleckenstein, & Möller, 2017; de Jong, 2016; Howard et al., 2018). Bilingual programs in which the proportion of foreign language instruction is up to 50% are usually referred to as CLIL, whereas in one-way immersion, typically, more than 50% of the instruction occurs in the second language (Möller et al., 2018; Dalton-Puffer, 2011). In TWI, the proportion of both languages is balanced, at least over time (Lindholm-Leary, 2001). The language allocation plans for the languages of instruction vary widely in this regard; in some TWI programs, each subject is allocated to one language, whereas in others, all subjects are taught in both languages (Howard et al., 2018). TWI can also be distinguished from other forms of immersion by the student composition, which in TWI, unlike one-way immersion, consists of students with one or the other language of instruction or both as their first language(s) (Lindholm-Leary, 2001). In some lessons, the language of instruction is a student's L1; in others, it is their L2. Because all students are first-language speakers of one or both of the two languages of instruction, competent role models for both languages are always at hand (Christian et al., 2000), whether in or outside of lessons. Moreover, in many TWI programs, the teachers instruct in their first language.

TWI programs aim to achieve additive bilingualism. Therefore, the goal for students is to become proficient in an additional language, without any detrimental effects on L1 development (Lambert, 1984). Fundamental to bilingual programs is the interdependence hypothesis, according to which there is a positive correlation between the development of the L1 and the L2 (Cummins, 1979, 1984). Like an iceberg, the linguistic abilities in the respective languages show on the surface, but underneath the surface, they have a common basis—the common

underlying proficiency (Cummins, 1981). With adequate motivation and contact with the other language, instruction in one of the languages then leads to an improvement in the other language as well. Another important assumption is that using the first foreign language (L2) as the language of instruction leads to more effective and extensive acquisition of the L2 than in traditional foreign language instruction because the language is learned implicitly (Krashen, 1982). According to Krashen, implicit means that language learning does not occur explicitly through vocabulary learning and grammar rules as is common in traditional foreign language instruction, but that language learning occurs more unconsciously and in a less controlled manner in everyday, natural language situations.

Reading is an important skill for participating in society. Reading comprehension is understood as a complex interactive and constructive process, resulting in the mental representation of the text (Mullis & Martin, 2015; van den Broek & Helder, 2017). It involves interactions between processes at different levels (Castles et al., 2018). At a lower level, these are the basal processes of decoding, the acquisition of word meanings, and the semantic and syntactic relation of parts of sentences and clauses; these processes are usually automated for "experienced" readers (McNamara & Magliano, 2009). At a higher level, comprehensive reading is based on active constructions of overarching coherence formation that lead to an adequate situation model (mental representation; van Dijk & Kintsch, 1983). Moreover, while reading, passive and reader-initiated processes occur that interact with each other and with the evolving mental representation of the text (van den Broek & Helder, 2017).

In elementary school, the basics of comprehensive reading are usually acquired, and especially lower-level processes are routinized by the end of grade 4 so that children can read and comprehend age-appropriate text fluently (Klicpera & Gasteiger-Klicpera, 1993). After completing literacy instruction, reading literacy development slows down (see Baumert, Nagy, et al., 2012; Chall, 1983; Francis et al., 1996); its development then depends less on language instruction in school and more on the individual's amount of reading (Becker et al., 2014; Guthrie et al., 1999). Acquiring the basics of reading comprehension relies on explicit instruction. The basic instruction at the elementary school level is therefore of particular importance. This may apply especially to language-minority speakers if their development in the majority language lags behind that of language majority speakers in semantics, grammar, and syntax.

TWI offers students the opportunity to learn to read in two languages simultaneously. One advantage over mainstream schools is that (at least) one language of instruction is the first language for each student.

This study aims to analyze the development of reading proficiency in the majority language, in this case, German, of TWI students compared to students in regular instruction. The article also deals with a specific type of TWI program, where schools offer nine different language combinations with the majority language. The present study is part of a larger project. In a previous article, reading comprehension in German and a partner language was analyzed and compared cross-sectionally but not longitudinally (Preusler et al., 2019). Before describing our study and the research questions in this article, we will first provide an overview of previous research on reading comprehension in the majority language in TWI research.

1.1. Research on TWI

In this passage, we review the research on reading in the majority language in TWI, focusing on minority-language students. Then we present research on the native language effect (i.e., the effect that L1 students are more proficient than L2 students are) and the longitudinal effects of TWI on reading comprehension.

Several studies have shown the benefits of immersive instruction (for an overview, see Baker & Lewis, 2015). Large-scale studies have explored the positive effects of one- and two-way immersion on academic achievement and on developing language skills in both the L1 and the L2 (for an overview, see Fleckenstein et al., 2019; Kim et al., 2015; Krashen, 2005; Watzinger-Tharp et al., 2018). According to numerous longitudinal studies with US public schools (Collier & Thomas, 2017; Thomas & Collier, 2002), TWI programs have been shown to be superior to English-only and transitional bilingual programs. Transitional bilingual programs are a type of bilingual education in which children first learn to read in their non-English first language and later – usually between the second and the fourth grade – progress to English reading instruction (Slavin & Cheung, 2005).

Considering the majority language proficiency, Lindholm-Leary (2011) showed that the students in Chinese TWI programs in the US performed at or above grade level in English. A study of a Mandarin/English TWI program reported no statistically significant differences between TWI and non-immersion students in an English language arts test in third and fifth grades (Padilla et al., 2013). However, the sample size was very small, and no background variables were included in the analyses. Similarly, Marian et al. (2013) found that the test scores in reading in English of majority-language students in TWI programs were higher than or equal to those of mainstream classes. Berens et al. (2013) found that second- and third-grade children from English-only households who were enrolled in TWI programs performed better or equally well in English reading tasks than such children in monolingual classes. Their results also showed an advantage of TWI students from English-only homes over TWI students who grew up with both languages of instruction. Steele et al. (2017) showed positive effects of TWI on reading in English in grades five and eight and comparable performances in other subjects and grades.

Students whose L1 is the minority language seem to benefit particularly from becoming proficient in the majority language (for an overview, see Genesee & Lindholm-Leary, 2013; Krashen, 2005). Lindholm-Leary and Block (2010) found that the English proficiency of English Language Learners (ELL) in TWI programs was higher than that of corresponding students under regular instruction. When the reading achievement was compared with that of students in transitional programs, likewise, minority-language students in TWI appeared to perform better (Marian et al., 2013). A study of Latino students in TWI programs by Lindholm-Leary and Hernandéz (2011) allocated students to groups by language proficiency (native English speakers, current ELLs, former ELLs). There was no difference between native English speakers and former ELLs, yet the current ELLs scored significantly lower in an English reading proficiency test than the other two groups. In their summary of several meta-analyses, McField and McField (2014) showed that Spanish instruction improves performance in the majority language English among Latin American immigrants. Sufficient input in the majority language, alphabetization, and subject teaching in the minority language were beneficial for majority language competencies. Jepsen (2010; see also Conger, 2010) initially found negative effects of bilingual programs on competence in the majority language for members of a minority language. However, these effects were absent when the majority language was used in equal proportion to the minority language Spanish. Studies examining ELs in various instructional programs – including English-only and dual-language immersion – from elementary to middle school also found some negative effects in bilingual programs in the early grades. Thus, while Latino ELs in the bilingual programs were slower to reclassify to "fluent English proficient" in elementary school, by the end of high school, a higher number of ELs had been reclassified, and their English proficiency was higher than in English-only classes (Umansky & Reardon, 2014). ELs in all bilingual programs were also found to show faster growth rates in English language arts than their monolingually taught peers (Valentino & Reardon, 2015).

Several studies have shown that students who acquired the test language as L1 generally seem to score higher than students who acquired the test language as L2 (or L3). This effect is called the native language effect (Lindholm-Leary & Howard, 2008). In Spanish/English

TWI programs, native speakers seem in general to outperform second-language learners, as concluded in the summary of research by Howard et al. (2003). Nevertheless, the differences decreased in the course of the study over three years. The native language effect was also found for reading in the third grade, with native English speakers scoring higher in English and native Spanish speakers scoring (slightly) higher in Spanish (Howard et al., 2004). However, it was not controlled sufficiently for the selectivity of the program. The students in the various language groups and programs differed in terms of socioeconomic background as measured by free/reduced lunch status and parent education. Therefore, this was a critical limitation. In an analysis of English and Spanish writing performance, Neugebauer and Howard (2015) also found an advantage for native speakers when controlled for gender and free and reduced-price lunch. Howard and Neugebauer (2015), who looked at the English and Spanish writing skills of TWI students from the second to the fifth grade, found similar results. A greater use of a language at home was associated with higher writing proficiency in that language. In English, the performance of students with a greater use of Spanish at home approached the performance of those who spoke more English at home over time. In a preceding article on a German TWI program (Preusler et al., 2019), 939 immersion students worked on reading tests from international large-scale studies in German and a partner language. Results of multivariate regression analyses in which socioeconomic status (SES), cognitive abilities, age, and gender were controlled for supported the native language effect, with L1 speakers outperforming L2 speakers for both languages. The study also provided initial findings for an extension of the native language effect to simultaneous bilinguals: Native speakers outperformed not only students who learned the test language as L2 but also students who grew up bilingually and thus spoke the test language and another language.

Children's socioeconomic background in TWI programs often varies greatly depending on different language backgrounds (Howard & Sugarman, 2001). Children with the minority language as their first language tend to come from families with a lower SES than children who are native speakers of the majority language (Marian et al., 2013). Although the children's background varies widely, control variables to adjust for selection bias are rarely included in studies (Steele et al., 2017). This limitation has been common among studies on TWI. There are few longitudinal studies on TWI in which students were randomized or in which background variables were also used for analysis. In a study by Steele et al. (2017), students were randomly assigned to either a TWI group or a control group that received regular instruction. In grade 5, the immersion students showed significantly better reading proficiency in the majority language than the students did in regular instruction, and the gap between the two groups increased over the time of the study as shown in the reading test in grade 8. The results indicated a benefit for two-way immersion students. Another recent study that controlled for several student-level background variables showed that elementary students in TWI programs acquired the majority language English faster than students who received monolingual instruction (Serafini et al., 2020).

Although a considerable body of research has investigated the outcomes of majority- or minority-language students in immersion programs (for an overview, see Lindholm-Leary & Genesee, 2014), few studies have sought to examine the proficiency of simultaneous bilinguals as a third language group (e.g., Berens et al., 2013; Howard & Neugebauer, 2015). Moreover, almost all TWI studies have originated from the North American region and have generally been limited to the language combination English-Spanish, which has been the most widespread combination in the United States (Howard & Sugarman, 2001). To summarize, there is a need for longitudinal studies on reading development in the majority language from countries other than the US and Canada that evaluate reading comprehension in the majority language in TWI while taking important background information into account.

1.2. The present study

Although there are many TWI studies, the few existing longitudinal studies do not clarify how exactly the different language groups' reading literacy develops in TWI compared to regular instruction. Therefore, in this study, we examined the development of German reading comprehension of TWI students from fourth to sixth grade compared to students that are conventionally taught. To this end, latent growth curve models were estimated and, in addition to group status (TWI or regular schooling), background variables were included. Our design allows us to examine whether there were differences between the two groups in fourth-grade achievement and development in the following two school years, including the role of students' first language(s). In grade four, students have already mastered basic reading and writing skills, which makes it exciting to compare their language proficiency. Grades 4 to 6 are also an interesting period because the transition to secondary school takes place after the sixth grade in Berlin, although an earlier transition is also possible starting after the fourth grade.

With our analysis, we address four research questions. The first research question is how the German (as the majority language) reading proficiency developed between Grades 4 and 6. With the second research question, we want to determine the program's effect, that is, whether there were differences between a TWI program with nine different language combinations and a control group with regular instruction that was similar in terms of relevant characteristics. The third question is how the German reading comprehension development varied depending on different linguistic backgrounds, that is, how it differed between students with German as their L1 or L2 and simultaneous bilinguals. Therefore, question 3 is about the effect of the home language. Finally, the fourth research question focuses on whether there was a combined effect for the TWI program and the home language. In our analyses, we examine two hypotheses about the level of reading proficiency in the fourth grade and two hypotheses about the changes in the course of the following grades. Based on the empirical findings (Preusler et al., 2019; Howard & Neugebauer, 2015), students who grew up speaking both languages (simultaneous bilinguals) should show lower German reading comprehension than L1 German-speaking students (Hypothesis 1a) but higher German reading comprehension than students with the minority language as L1 (Hypothesis 1b). As students with a minority language as L1 seem to benefit particularly from TWI (Genesee & Lindholm-Leary, 2013; Krashen, 2005), we expect that this group of students would perform better in fourth grade in the TWI program than in regular instruction in fourth grade (Hypothesis 2). After grade 4, we expect the advantage to decrease and to show a tendency toward parallel developments. Therefore, according to Hypothesis 4, no differential learning trajectories are expected based on students' L1 in either the TWI or the mainstream school. After grade 4, we expect the advantage to decrease and to show a tendency toward parallel developments. Therefore, according to Hypothesis 4, no differential learning trajectories are expected based on students' L1 in either the TWI or the mainstream school. Previous studies of majority language proficiency found that TWI students performed as well as or better than comparable students in mainstream classes (Berens et al., 2013; Lindholm-Leary, 2011; Marian et al., 2013; Steele et al., 2017). Therefore, Hypothesis 3 is that TWI students perform better or equally well than students in mainstream schools, although German instruction was significantly reduced (see below).

2. Methods

2.1. The state Europe school of Berlin

The State European School of Berlin (SESB) is an example of a TWI program (Möller et al., 2017 ; Preusler et al., 2019). At the SESB, children whose L1 is German are taught together with children who have another language as L1. Both languages have equal status as languages of

instruction. There are nine language combinations at the SESB: German with either English, French, Greek, Italian, Polish, Portuguese, Russian, Spanish, or Turkish, with one language combination offered per location. Half of the lessons are held in German and half in the partner language. The teachers are native speakers of the respective language. The language of instruction differs depending on the subject. For example, mathematics is always taught in German, and natural sciences/biology is always taught in the partner language. Thus, the German-speaking instructional input is reduced compared to regular schools in Germany.

In theory, the SESB classes consist of 50% native speakers of German and 50% native speakers of the partner language. In practice, a considerable proportion of pupils are raised bilingually with two L1s: German and the partner language (Baumert, Hohenstein, Fleckenstein, & Möller, 2017). All children who have a sufficient level of proficiency in German or the partner language (as defined in a language test conducted by the SESB) are entitled to apply for a place in a SESB school. In general, admission to SESB takes place in grade 1. A later entry is possible if spaces become available and if a level of proficiency in both languages sufficient for participation is established in an admission interview (Senatsverwaltung für Bildung, Jugend und Senatsverwaltung für Bildung Jugend und Familie, 2018). Students are taught in both languages from first grade to school graduation. From the first to the eighth grade, the classes are divided into two groups for language lessons (in German and the partner language) depending on the (more dominant) L1. For the rest of the subjects and after eighth grade, the students are taught as a mixed language group (Abgeordnetenhaus Berlin, 2010; Möller et al., 2017). The goal of the SESB is to develop proficiency in both languages, resulting in additive bilingualism. This study is part of the evaluation of the SESB commissioned and approved by the Berlin Senate to establish SESB as a school with a special educational character.

2.2. Participants

The sample consisted of $N = 984$ TWI students from 17 SESB schools and $N = 992$ students with regular instruction from a similar amount of classes. The study entailed an exhaustive assessment of all TWI classes in the 2013/14 school year and a follow-up assessment of TWI schools with low student numbers in the 2014/15 school year. In addition, monolingual parallel classes taught at the same primary school locations were included in the study as a control group of maximum similarity. In cases where the SESB locations were managed as purely bilingual schools, control classes were selected at primary schools with comparable catchment areas. All of the students were assessed three times, at the end of fourth grade, fifth grade, and sixth grade.

2.3. Variables

2.3.1. Individual and background characteristics

Information about the L1 was assessed in both a parent and a student questionnaire, using the questions: "Which language(s) did the child for whom you are completing the questionnaire acquire first in the family?" and "Which language(s) did you acquire in your family from the beginning?". Multiple answers were possible. Priority was given to the parent responses; we consulted the student responses only if the parent response was missing. From this information, we have formed three subgroups for the analyses: children raised monolingually in German (L1 German speakers) or the partner language (L1 partner language speakers) and children raised bilingually in German and the partner language from infancy (simultaneous bilinguals). In the group with regular instruction, *partner language* refers to any language other than German.

Previous research has shown that students in German immersion programs are often positively selected in terms of socioeconomic background and cognitive performance compared to students in mainstream schools (Baumert, Köller, et al., 2012 ; Baumert, Hohenstein, Fleckenstein, & Möller, 2017). Therefore, these variables were included as control variables to statistically control for the influence of entrance selectivity.

The International Socioeconomic Index of occupational status (ISEI; Ganzeboom & Treiman, 1996) was used to operationalize the family's *socioeconomic status* (SES). For this purpose, the parents were asked about their occupation, their occupational activities, their authority to give instructions, and their education. From these occupational data, the ISEI was formed. When the parents' ISEI scores differed, we used the higher value (HISEI = highest ISEI).

Children's cognitive abilities in fourth grade were assessed using the figural subtest of the Test of Cognitive Abilities (KFT; Heller & Perleth, 2000) for fourth-grade students. The test comprises 25 items; its internal consistency was very good (Cronbach's $\alpha = 0.94$ for TWI students, Cronbach's $\alpha = 0.93$ for students with regular instruction).

The *child's gender* was coded as 0 = male and 1 = female.

As shown in Table 1, the TWI group children were positively selected regarding socioeconomic background (higher SES) but not in terms of their cognitive abilities (KFT). There were considerably more children who grew up as simultaneous bilinguals or with a non-German language as L1 in the TWI group.

2.4. Achievement characteristics

Reading comprehension was assessed longitudinally. Texts and tasks for all three measurement points were taken from the longitudinal ELEMENT study (Lehmann & Lenkeit, 2008), which tracked the development of students' reading and mathematics proficiency from Grades 4 to 6 in Berlin. The fourth-grade test from the ELEMENT study was originally taken from the Progress in International Reading Literacy Study (PIRLS; Mullis et al., 2003). It consisted of one short story, followed by 13 questions (seven multiple-choice items & six open-ended items). The total maximum score was 17. In fifth grade, the test comprised two short stories (one of which was the same as in fourth grade) and 14 related questions (of which nine were multiple-choice, and five were open-ended). A maximum score of 18 could be achieved. The sixth-grade test consisted of three short stories and one non-literary text with 24 associated multiple-choice questions and a maximum score of 24. Trained test administrators administered the tests to the whole class.

A joint scaling with the ELEMENT study data (Lehmann & Lenkeit, 2008) had to be used to achieve an anchor items structure since there were linking items between the fourth and fifth grade, but not between the fifth and sixth grades. The test data, which comprised both dichotomous and ordinal (partial-credit) variables, were scaled to fit a one-parameter generalized item response theory (IRT) model (Adams & Wu, 2007):

Table 1
Participant demographics.

Variable	TWI group	Regular instruction
Gender (female)	53.7%	50.1%
Language group		
L1 German	20.3%	56.2%
Simultaneous bilinguals	40.5%	17.2%
L1 partner language	39.2%	26.6%
Parents' SES [*M* (*SD*)]	60.23 (21.19)	53.56 (22.63)
Age (grade 4) [*M* (*SD*)]	10.13 (0.45)	10.26 (0.54)
Cognitive abilities (figural, grade 4) [*M* (*SD*)]	16.20 (7.11)	16.24 (6.94)
Total *N*	984	992

1. We used the item difficulties of PIRLS from 2001 (Mullis et al., 2003) and 2006 (Mullis et al., 2007) as anchors to estimate a joint IRT model from our fourth-grade data with the data from the fourth grade of the ELEMENT study.
2. We freely estimated joint IRT models for our data together with the ELEMENT data for the fifth and sixth grades.
3. According to the Stocking-Lord method (Kolen et al., 2014), we linked the data from all three-measurement dates with a joint linking approach.
4. Weighted Likelihood Estimates (WLEs; Warm, 1989) were calculated as estimates of individual proficiency.

Scaling, linking, and computation of the WLEs were conducted in R (R Core Team, 2021) using the TAM package (Robitzsch et al., 2020). The reading test's reliability was satisfactory to good and thus sufficiently high for group comparisons, between $r_{\mathrm{WLE}} = 0.76$ and $r_{\mathrm{WLE}} = 0.82$. We adjusted the WLEs for unreliability by adopting a latent variable approach.[1]

2.4.1. Treatment of missing values and hierarchical data structure

To ensure the fullest possible sample size and prevent bias, we performed multiple imputation using the *mice* package in R (R Core Team, 2021; van Buuren & Groothuis-Oudshoorn, 2011). In this procedure, missing values are estimated based on available background variables (Lüdtke et al., 2017; Schafer & Graham, 2002). Specifically, we used the background variables school class, student age and gender, language background, immigration background, parental SES, KFT score, test scores from other parts of the study (science, vocabulary, English, math), and school grades. As the students were enrolled in several classes and locations in Berlin, the data structure was hierarchical (students nested within classes). To account for the hierarchical data structure, we used the imputation method "2l.pan" from the R package "pan" (Zhao & Schafer, 2018). We generated 65 imputed data sets and integrated the results according to the rules of Rubin (1987).

2.4.2. Statistical analyses

The research questions were examined using latent growth curve models (LGCM; Bollen & Curran, 2006). These models were run in M*plus* 7.4 (Muthén & Muthén, 1998-2012). Latent growth curve models are based on two latent variables: The intercept that allows the baseline level to be estimated and the slope that is used to estimate growth. For this article, linear growths were assumed. The factor loadings at the three points of measurement were fixed to zero at T1, one at T2, and two at T3, as tests were carried out every year and the intervals between tests were thus equal. The LGCMs were built stepwise by sequentially entering predictors. In Model 1, a common model was first estimated for all students. In order to examine whether the initial status and the learning trajectory in reading achievement differed between the TWI and the students with regular instruction, the TWI status was included in Model 2. The TWI status was a dummy variable for TWI, with the regular instruction group serving as the reference group. In Model 3, we extended the model by including the following background variables: linguistic background, gender, SES, cognitive abilities, and age. We used two dummy variables to code students' language group membership (simultaneous bilinguals = SB, L1 partner language speakers = L1PLS), with L1 German speakers serving as the reference group. Finally, in Model 4, the estimation of Model 3 was repeated using the simultaneous bilinguals as the reference group and, again, using two dummy variables for the language groups (L1 German speakers = L1GS, L1 partner language speakers = L1PLS). The model without background variables (Model 2) compares the students' actual performance in TWI versus mainstream schools. In contrast, the models with background variables (Models 3 and 4) allow a fairer comparison because they correct for important, potentially biasing background variables. As students were nested in classes, we estimated cluster-robust standard errors in all analyses using the option TYPE = COMPLEX, which corrects the standard errors for clustering effects in the data.

Four fit indices were used to evaluate the goodness of fit of the models: The Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). For the CFI and TLI, values greater than 0.90 are considered acceptable and values greater than 0.95 indicate a very good fit of the model to the data (Marsh et al., 2004). Like the RMSEA value, the SRMR should not exceed 0.08 for an acceptable fit (Hu & Bentler, 1999; Kline, 2005; Schreiber et al., 2006). There is no theoretical basis for calculating chi-square tests using multiply imputed data. Therefore, no chi-square values are reported.

3. Results

3.1. Descriptive findings

Table 2 shows the unadjusted means, standard deviations, and intercorrelations for the variables based on the imputed data. There was a significant increase in both groups' mean values over the years for the performance variables. Significant correlations were found between all performance variables. The covariates SES and cognitive abilities also showed significant correlations with the performance variables.

3.2. Results of the latent growth curve analyses

We specified latent growth curve models to analyze the influence of the type of schooling (TWI vs. regular instruction) and the impact of background variables on the learning trajectories of reading proficiency. For reasons of reducing a plethora of interactions, we limit ourselves to the overall representation across all program languages, which is a good approximation of the individual findings. The results of the analyses for the individual school languages can be found in the appendix. Table 3 shows the model fit information for all illustrated models. A clear deviation from the cutoff values applied for an acceptable model fit occurred for the intercept-only (no change) model. Models 1–4 showed an almost perfect fit.

The parameters of the latent growth curve models are reported in Table 4. With the first research question, we wanted to examine the general growth pattern of reading proficiency in German. The results of Model 1 showed, as expected, positive absolute growth for all students from grades 4 to 6 ($\mu_{\mathrm{SLOPE}} = 0.22$, $p < .001$). While there were statistically significant interindividual differences in the initial level of reading proficiency ($\psi_{\mathrm{Intercept}} = 0.75$, $p < .001$), there were no statistically significant interindividual differences in the linear increase over time. There was also no significant correlation between the initial value and the developmental trajectory.

The second research question was about the effect of TWI. Therefore, Model 2 includes the TWI group status to examine the influence of the type of schooling. Neither the initial level nor the learning trajectory showed significant differences between the TWI students and the students with regular instruction, supporting Hypothesis 3 according to which TWI students were expected to perform better or equally well than students in regular schools.

With the third research question, we wanted to investigate what effect the home language had, that is, whether the development of German reading comprehension differed depending on the linguistic background. For this purpose, in Models 3 and 4, the language group and background variables were entered as predictors for the reading

[1] Specifically, for each point of measurement, we modeled the true score of the WLE by a latent variable. As the indicator, we used the observed WLE. The loading of this indicator was fixed to 1, and the error variance was set to the mean of the squared WLE standard errors, which we computed prior to the analysis. Furthermore, we equated the intercepts of the observed WLEs across the three points of measurement in order to identify the model.

Table 2
Non-adjusted means (M), standard deviations (SD), group differences, and correlations of the variables based on the imputed data (above the diagonal for TWI students, below the diagonal for comparison students).

Variable	TWI		Regular instruction		Group differences						
	M	*SD*	*M*	*SD*	*t*	*P*	(1)	(2)	(3)	(4)	(5)
(1) Reading grade 4	0.52	1.09	0.48	1.09	0.86	.39	–	**.64**	**.59**	**.51**	**.32**
(2) Reading grade 5	0.78	1.14	0.71	1.07	1.12	.26	**.62**	–	**.62**	**.49**	**.28**
(3) Reading grade 6	0.96	1.23	0.90	1.18	0.75	.45	**.56**	**.59**	–	**.44**	**.26**
(4) Cognitive abilities	16.20	7.11	16.24	6.94	−0.03	.97	**.53**	**.47**	**.47**	–	**.32**
(5) Parents' SES	60.23	21.19	53.56	22.63	5.74	<.001	**.42**	**.31**	**.35**	**.40**	–

Note: Cognitive abilities were measured in fourth grade. Correlations significantly different from 0 ($p < .001$) are printed in bold.

Table 3
Model fit information for all illustrated models.

Model	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Intercept-only model	5	0.93	0.95	0.14	0.09
Model 1	8	0.99	0.99	0.03	0.01
Model 2	10	0.99	0.99	0.03	0.01
Model 3	26	0.99	0.98	0.04	0.01
Model 4	26	0.99	0.98	0.04	0.01

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual

proficiency and developmental trajectory, reducing the interindividual differences of the initial level to $\psi_{\text{Intercept}} = 0.44$, $p < .001$. Here, the SES and cognitive indicators are control variables to highlight the linguistic results. Model 3 is graphically represented in Fig. 1. First, the intercept findings will be described. Not surprisingly, a higher SES and better cognitive performance were accompanied by higher reading performance in fourth grade ($b_{\text{Intercept, HISEI}} = 0.21$, $p < .001$; $b_{\text{Intercept, KFT}} = 0.31$, $p < .001$).

Regarding the linguistic background, the native language hypothesis found support: students who grew up with German as L1 outperformed students whose L1 was a partner language ($b_{\text{Intercept, L1PLS}} = -0.40$, $p < .001$). We expected students who grew up speaking both languages to show lower German reading comprehension than L1 German-speaking students (Hypothesis 1a) but higher German reading comprehension than students with the minority language as L1 (Hypothesis 1b). As assumed by Hypothesis 1a, L1 German speakers performed better than students who grew up bilingually ($b_{\text{Intercept, SB}} = -0.25$, $p < .001$). Contrary to Hypothesis 1b, Model 4 showed no statistically significant differences between simultaneous bilinguals and L1 partner language speakers in fourth-grade reading achievement.

Regarding the slope findings, the addition of the background variables did not change compared to Model 2: There were no significant differences in growth between the TWI group and the students with regular instruction. Thus, the finding had proven to be robust against the inclusion of background variables, which further supports Hypothesis 3.

Consistent with Hypothesis 4, which expected no differential learning trajectories based on students' L1, there were no differences among the three language groups in terms of growth rates in either the TWI or the mainstream school.

The fourth research question addressed whether there was a combined effect for the TWI program and the language background. There was no effect of TWI on fourth-grading reading performance and therefore no advantage for students whose L1 was a partner language at TWI. This indicates no support for Hypothesis 2 according to which students with a minority language as L1 should perform better in fourth grade in the TWI program than in regular instruction.

4. Discussion

This article investigated the development of reading proficiency in the German language of TWI students from fourth to sixth grade in comparison to conventionally taught students in Berlin. In particular, we aimed at investigating whether German reading comprehension development differed between students with different first languages. In fourth grade, we found that TWI students performed at the same level as students in mainstream schools, even though their amount of German instruction was reduced. The addition of background variables to the analysis did not change this result. This finding is in line with past studies documenting that TWI students performed similarly in the national language to conventionally monolingually taught students (Berens et al., 2013; Lindholm-Leary, 2011; Marian et al., 2013; Padilla et al., 2013). However, there are also studies in which TWI students performed better than monolingually taught students (e.g., Steele et al., 2017; Thomas & Collier, 2002).

Looking at the learning trajectories between grades 4 and 6, again, there was no difference between the TWI students and the students with regular instruction. One explanation for the TWI students' comparable reading performance in German despite a reduced amount of overall instruction time in German compared to conventionally taught students might be positive transfer effects of bilingual language learning. According to Cummins (1979, 1981, 1984), the languages have a common basis that leads to the interdependence of the two languages and, given sufficient language contact, contributes to reciprocal improvement in both languages. Thus, it is possible that TWI students improve their reading performance in their first language (German) by learning another language, or that students who have a non-German first language also improve their reading ability in German by receiving instruction in their first language. The requirement of sufficient input in the partner language is given by the program.

Another reason for comparable performance with reduced German instruction could be selectivity effects. On average, immersion students come from families with higher socioeconomic status. While we controlled for this covariate, other variables could also vary due to selection bias and thus influence reading achievement. One example of such possible influences is parental support for student learning, which could be higher for immersion students because their parents tend to be very positive about the programs (Lindholm-Leary, 2001).

Even though the TWI students in our study showed comparable performance in German reading comprehension, the question arises why they did not perform better than the monolingually taught students did as other TWI studies have shown (Serafini et al., 2020; Steele et al., 2017). One possible explanation is that assessment at fourth to sixth grade level is too early to detect a positive effect of TWI on performance in the majority language. However, an earlier cross-sectional survey of ninth-grade students in the same TWI program also showed comparable, not better, reading achievement in German compared to students in regular instruction (Fleckenstein et al., 2017). A unique feature of this TWI program is the separation of students in language lessons in grades one through eight by the (dominant) first language. This could also explain why stronger outcomes for TWI could not be shown because it is not clear what the effect of this separation is. On the one hand, the separation allows students to be better supported according to their level of language proficiency. Therefore, the separation could positively influence German performance in the TWI group. However, on the other

Table 4
Latent growth curve models for reading proficiency.

Means										
Intercept	0.51 (0.08)[d]	0.48 (0.12)[d]			0.57 (0.06)[d]			0.32 (0.10)[c]		
Slope	0.22 (0.03)[d]	0.22 (0.04)[d]			0.21 (0.05)[d]			0.20 (0.05)[d]		
Variances										
Intercept	0.75 (0.12)[d]	0.75 (0.12)[d]			0.44 (0.07)[d]			0.44 (0.07)[d]		
Slope	0.01 (0.05)	0.01 (0.05)			0.04 (0.04)			0.04 (0.04)		
Covariances										
Intercept x Slope	0.01 (0.07)	0.01 (0.07)			−0.01 (0.04)			−0.01 (0.04)		
Effects										
TWI										
Intercept	–	0.05 (0.16)	0.06	[-0.30, 0.42]	0.08 (0.07)	0.10	[-0.07, 0.26]	0.14 (0.11)	0.16	[-0.09, 0.41]
Slope	–	0.01 (0.04)	0.07	[-0.55, 0.68]	0.06 (0.05)	0.56	[-0.36, 1.47]	0.00 (0.05)	0.02	[-0.90, 0.95]
SB										
Intercept	–	–	–	–	−0.25 (0.08)[c]	−0.28	[-0.47, −0.10]	–	–	–
Slope	–	–	–	–	−0.01 (0.05)	−0.05	[-0.91, 0.80]	–	–	–
L1PLS										
Intercept	–	–	–	–	−0.40 (0.08)[d]	−0.46	[-0.63, −0.29]	−0.15 (0.09)	−0.18	[-0.39, 0.04]
Slope	–	–	–	–	0.06 (0.05)	0.52	[-0.37, 1.42]	0.07 (0.06)	0.58	[-0.42, 1.57]
L1GS										
Intercept	–	–	–	–	–	–	–	0.25 (0.08)[c]	0.28	[0.10, 0.47]
Slope	–	–	–	–	–	–	–	0.01 (0.05)	0.05	[-0.80, 0.91]
TWI x SB										
Intercept	–	–	–	–	0.06 (0.11)	0.07	[-0.17, 0.31]	–	–	–
Slope	–	–	–	–	−0.06 (0.06)	−0.54	[-1.64, 0.57]	–	–	–
TWI x L1PLS										
Intercept	–	–	–	–	−0.16 (0.09)	−0.18	[-0.38, 0.02]	−0.21 (0.11)[b]	−0.25	[-0.49, 0.00]
Slope	–	–	–	–	−0.09 (0.07)	−0.81	[-2.05, 0.44]	−0.03 (0.07)	−0.27	[-1.47, 0.93]
TWI x L1GS										
Intercept	–	–	–	–	–	–	–	−0.06 (0.11)	−0.07	[-0.31, 0.17]
Slope	–	–	–	–	–	–	–	0.06 (0.06)	0.54	[-0.57, 1.64]
Gender (1 = female)										
Intercept	–	–	–	–	0.23 (0.05)[d]	0.27	[0.16, 0.37]	0.23 (0.05)[d]	0.27	[0.16, 0.37]
Slope	–	–	–	–	−0.01 (0.03)	−0.13	[-0.68, 0.42]	−0.01 (0.03)	−0.13	[-0.68, 0.42]
Parents' SES[a]										
Intercept	–	–	–	–	0.21 (0.03)[d]	0.24	[0.18, 0.30]	0.21 (0.03)[d]	0.24	[0.18, 0.30]
Slope	–	–	–	–	−0.02 (0.02)	−0.10	[-0.31, 0.12]	−0.02 (0.02)	−0.10	[-0.31, 0.12]
Cognitive abilities[a]										
Intercept	–	–	–	–	0.31 (0.03)[d]	0.35	[0.29, 0.40]	0.31 (0.03)[d]	0.35	[0.29, 0.40]
Slope	–	–	–	–	0.00 (0.02)	0.00	[-0.18, 0.17]	0.00 (0.02)	0.00	[-0.18, 0.17]
Age[a]										
Intercept	–	–	–	–	−0.15 (0.03)[d]	−0.17	[-0.24, −0.10]	−0.15 (0.03)[d]	−0.17	[-0.24, −0.10]
Slope	–	–	–	–	−0.02 (0.02)	−0.09	[-0.33, 0.14]	−0.02 (0.02)	−0.09	[-0.33, 0.14]

Note. Unstandardized solution. Standard errors are in parentheses. CI = confidence interval; TWI = Two-way immersion; SB = simultaneous bilinguals; L1PLS = L1 Partner language speakers; L1GS = L1 German speakers. Cognitive abilities were measured in fourth grade.
[a] centered.
[b] $p < .05$.
[c] $p < .01$.
[d] $p < .001$.

hand, a separation in the language lessons speaks against the idea of TWI according to which competent language role models are always present (Christian et al., 2000). Furthermore, it is logistically difficult to assign children without a dominant first language (i.e., the simultaneous bilinguals) to one of the two language groups for instruction. Therefore, it should be considered to discontinue the separation and to teach the children together in all subjects as it is common in the USA. Certainly, it would be desirable if future studies could examine whether students in TWI are more effectively taught together or separated by language group.

We expected that students whose L1 is a minority language would perform better in the TWI program than in regular instruction as these students seem to benefit particularly from TWI (Genesee & Lindholm-Leary, 2013; Krashen, 2005). Contrary to these expectations, no differences emerged between L1 partner language speakers in TWI and mainstream schools in grade 4. This finding has proven to be robust against the inclusion of background variables. A possible explanation is that, unlike some other studies (e.g., Lindholm-Leary & Block, 2010), we included students' background variables like SES, gender, and cognitive abilities in the analyses. It is possible that the positive effects for language minority students in other studies are due to biases.

Thus, our results suggest that TWI students show the same reading performance development in the majority language as conventionally taught students, regardless of their language background. However, this also means that TWI is confronted with the same unresolved pedagogical challenges as the monolingual mainstream schools. First and foremost, these are the large gaps in the reading skills of L1 partner language speakers, which are substantial even when controlling for cognitive abilities and socioeconomic status. Like mainstream schools, TWI schools should focus more on targeting these students, such as providing increased language input in the form of positively evaluated interventions (e.g., Stanat et al., 2012). The meta-analyses of reading interventions for English Language Learners by Cho et al. (2021) and Ludwig et al. (2019) suggest providing interventions for older elementary students delivered in medium-size groups (6–15 students per group) and focusing on reading strategies; shorter interventions are successful as well. However, it should be noted that students in grades 4 to 6 were assessed in this study. It is unclear and will be worth looking at in future

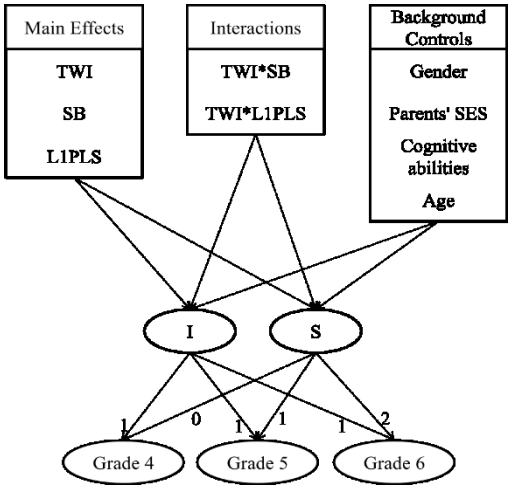

Fig. 1. Graphical representation of Model 3, Note. (Residual) correlations are not shown to improve readability. TWI = Two-way immersion; SB = simultaneous bilinguals; L1PLS = L1 Partner language speakers.

studies how TWI affects students in higher grades.

Our results replicate the native language effect, which predicts that native speakers perform better than students who learned the test language as L2 (Lindholm-Leary & Howard, 2008). This finding is attributable to the earlier age of onset in language acquisition and higher exposure to the language in the family setting. The extension of the native language effect according to which native (monolingual) speakers perform better than bilingually raised students (Preusler et al., 2019) was also confirmed. de Jong (2016) argued that a binary categorization of TWI students into members of the minority or majority language group is not appropriate. Our study shows that students who grew up bilingually perform quite differently from students who grew up monolingually. Therefore, future studies should treat bilingually raised students as a separate language group, as we did in this study.

There are several reasons why our study contributes to immersion research. It represents a longitudinal TWI study in Europe, in which multiple background variables were also collected. We examined a program that offers numerous language combinations, which goes beyond the previous focus on English-Spanish. Also, students who grew up bilingually were treated and analyzed as a distinct language group so that statements could also be made specifically for this group.

While our study only deals with the subject German, it would be interesting to investigate the effect of the test language in non-language subjects such as science. Due to the different linguistic levels, it would be possible to offer tests in the respective first language and the majority language at TWI schools to determine the students' potential in the respective subject. For instance, Canz et al. (2021) suggest that using the language of instruction to assess content knowledge may underestimate students' achievement in bilingual education. Similarly, for TWI, students' test scores were lower when tested in the non-German language, even when the subject was taught in the non-German language (Baumert, Hohenstein, Fleckenstein, Preusler, et al., 2017).

The limitations of this study are, of course, to be considered. We did not differentiate between the various language combinations in the analyses because the subgroups' sample sizes would have been too small. Thus, it remains unclear whether differences exist between individual language combinations. Unfortunately, random assignment to TWI or the control group was not possible. As with all quasi-experimental studies, causal inferences should be drawn with caution from the analyses presented in this paper. However, we used rigorous methods (i.e., background variables) to make fair comparisons between TWI students and the regular instruction students.

Our analysis supports the assumption that TWI does not lead to a disadvantage in the development of reading proficiency in primary schools. Neither in grade 4 nor in the development between grade 4 and grade 6 there were any significant differences in German reading comprehension between the TWI and the regular instruction students. This statement applies regardless of students' language background, socioeconomic status, and cognitive abilities. Beyond learning an additional language (Preusler et al., 2019), TWI students show the same levels of reading comprehension in German.

Author statement

Manuscript title: Development of German reading comprehension in two-way immersive primary schools, Sandra Preusler: Conceptualization, Data Curation, Methodology, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Steffen Zitzmann: Methodology, Formal analysis, Writing - Review & Editing, Jürgen Baumert: Writing - Review & Editing, Funding acquisition, Project administration, Jens Moller: Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Funding

We have no conflicts of interest to disclose. This study is a part of the "Europe study", funded by the Senate of Berlin and the Stiftung Mercator.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://hdl.handle.net/21.11116/0000-000A-211C-2.

References

Abgeordnetenhaus Berlin. (2010). Staatliche Europa-Schule Berlin – Bewährten Schulversuch abschließen und Europaschulzentren schaffen! [State Europe School Berlin – finalize the established school pilot project and create Europe School centers] (Drucksache 16/3575). Berlin: Abgeordnetenhaus Berlin, Vol. 16. Wahlperiode. http://www.parlament-berlin.de/ados/16/IIIPlen/vorgang/d16-3575.pdf.

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the rasch model. In M. von Davier, C. H. Carstensen, & M. Davier (Eds.), *Statistics for social and behavioral sciences. Multivariate and mixture distribution rasch models: Extensions and applications* (pp. 57–75). Springer New York. https://doi.org/10.1007/978-0-387-49839-3_4.

Baker, C. (2011). *Foundations of bilingual education and bilingualism (5. ed.). Bilingual education & bilingualism* (Vol. 79). Multilingual Matters.

Baker, C., & Lewis, G. (2015). A synthesis of research on bilingual and multilingual education. In W. E. Wright, S. Boun, & O. García (Eds.), *The handbook of bilingual and multilingual education* (pp. 109–126). Wiley-Blackwell. https://doi.org/10.1002/9781118533406.ch7.

Baumert, J., Hohenstein, F., Fleckenstein, J., & Möller, J. (2017b). Wer besucht die Staatliche Europa-Schule Berlin? Sprachlicher, ethnischer und sozioökonomischer Hintergrund sowie kognitive Grundfähigkeiten der Schülerinnen und Schüler [Who attends the State Europe School Berlin? Linguistic, ethnic and socioeconomic background as well as basic cognitive abilities of the pupils]. In J. Möller, F. Hohenstein, J. Fleckenstein, O. Köller, & J. Baumert (Eds.), *Erfolgreich integrieren - die Staatliche Europa-Schule Berlin* (pp. 75–94). Waxmann.

Baumert, J., Hohenstein, F., Fleckenstein, J., Preusler, S., Paulick, I., & Möller, J.(2017a). Die schulischen Leistungen an der SESB – vierte Jahrgangsstufe. In J. Möller, F. Hohenstein, J. Fleckenstein, O. Köller, & J. Baumert (Eds.), *Erfolgreich integrieren - die Staatliche Europa-Schule Berlin* (pp. 95–187). Waxmann.

Baumert, J., Köller, O., & Lehmann, R. (2012a). Leseverständnis im Englischen und Deutschen und Mathematikleistungen bilingual unterrichteter Schülerinnen und Schüler am Ende der Grundschulzeit. Ergebnisse eines Zwei-Wege- Immersionsprogramms. *Unterrichtswissenschaft, 40*(4), 290–314.

Baumert, J., Nagy, G., & Lehmann, R. (2012b). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools? *Child Development, 83*(4), 1347–1367. https://doi.org/10.1111/j.1467-8624.2012.01779.x

Becker, M., McElvany, N., Lüdtke, O., & Trautwein, U. (2014). Lesekompetenzen und schulische Lernumwelten. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 46*(1), 35–50. https://doi.org/10.1026/0049-8637/a000104

Berens, M. S., Kovelman, I., & Petitto, L.-A. (2013). Should bilingual children learn reading in two languages at the same time or in sequence? *Bilingual Research Journal, 36*(1), 35–60. https://doi.org/10.1080/15235882.2013.779618
Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Wiley series in probability and statistics. Wiley-Interscience.
van den Broek, P., & Helder, A. (2017). Cognitive processes in discourse comprehension: Passive processes, reader-initiated processes, and evolving mental representations. *Discourse Processes, 54*(5–6), 360–372. https://doi.org/10.1080/0163853X.2017.1306677
van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03
Canz, T., Piesche, N., Dallinger, S., & Jonkmann, K. (2021). Test-language effects in bilingual education: Evidence from CLIL classes in Germany. *Learning and Instruction, 75*, 101499. https://doi.org/10.1016/j.learninstruc.2021.101499
Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest, 19*(1), 5–51. https://doi.org/10.1177/1529100618772271
Chall, J. S. (1983). *Stages of reading development*. McGraw-Hill.
Cho, Y., Kim, D., & Jeong, S. (2021). Evidence-based reading interventions for English Language Learners: A multilevel meta-analysis. *Heliyon, 7*(9), Article e07985. https://doi.org/10.1016/j.heliyon.2021.e07985
Christian, D., Howard, E. R., & Loeb, M. I. (2000). Bilingualism for all: Two-way immersion education in the United States. *Theory Into Practice, 39*(4), 258–266. https://doi.org/10.1207/s15430421tip3904_9
Collier, V. P., & Thomas, W. P. (2017). Validating the power of bilingual schooling: Thirty-two years of large-scale, longitudinal research. *Annual Review of Applied Linguistics, 37*, 203–217. https://doi.org/10.1017/S0267190517000034
Conger, D. (2010). Does bilingual education interfere with English-language acquisition? *Social Science Quarterly, 91*(4), 1103–1122. https://doi.org/10.1111/j.1540-6237.2010.00751.x
Coyle, D. (2007). Content and Language integrated learning: Towards a connected research agenda for CLIL pedagogies. *International Journal of Bilingual Education and Bilingualism, 10*(5), 543–562. https://doi.org/10.2167/beb459.0
Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*(2), 222–251. https://doi.org/10.3102/00346543049002222
Cummins, J. (1981). The role of primary language development in promoting educational success for Language Minority Students. Schooling and language minority students: A theoretical framework. In *California state department of education* (pp. 3–49). Evaluation, Dissemination, and Assessment Center, California State University. https://doi.org/10.13140/2.1.1334.9449.
Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. Multilingual Matters.
Dalton-Puffer, C. (2011). Content-and-Language Integrated learning: From practice to principles? *Annual Review of Applied Linguistics, 31*, 182–204. https://doi.org/10.1017/S0267190511000092
van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.
Fleckenstein, J., Gebauer, S. K., & Möller, J. (2019). Promoting mathematics achievement in one-way immersion: Performance development over four years of elementary school. *Contemporary Educational Psychology, 56*, 228–235. https://doi.org/10.1016/j.cedpsych.2019.01.010
Fleckenstein, J., Möller, J., Hohenstein, F., Radmann, S., Becker, M., & Baumert, J. (2017). Die schulischen Leistungen an der SESB - 9. Jahrgangsstufe und 15-Jährige [Academic performance at SESB - 9th grade and 15-year-olds.]. In J. Möller, F. Hohenstein, J. Fleckenstein, O. Köller, & J. Baumert (Eds.), *Erfolgreich integrieren - die Staatliche Europa-Schule Berlin* (pp. 189–252). Waxmann.
Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*(1), 3–17. https://doi.org/10.1037/0022-0663.88.1.3
Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. *Social Science Research, 25*(3), 201–239. https://doi.org/10.1006/ssre.1996.0010
Genesee, F. (1981). A comparison of early and late second language learning. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement, 13* (2), 115–128. https://doi.org/10.1037/h0081168
Genesee, F., & Lindholm-Leary, K. (2013). Two case studies of content-based language education. *Journal of Immersion and Content-Based Language Education, 1*(1), 3–33. https://doi.org/10.1075/jicb.1.1.02gen
Guthrie, J. T., Wigfield, A., Metsala, J. L., & Cox, K. E. (1999). Motivational and cognitive predictors of text comprehension and reading amount. *Scientific Studies of Reading, 3* (3), 231–256. https://doi.org/10.1207/s1532799xssr0303_3
Heller, K. A., & Perleth, C. (2000). Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+R) [Cognitive Abilities Test for grades 4 to 12, revision (KFT 4-12+R)]. Hogrefe.
Howard, E. R., Christian, D., & Genesee, F. (2004). *The development of bilingualism and biliteracy from grades 3 to 5: (Research report 13): A summary of findings from the CAL/ CREDE study of two-way immersion education*. Research Report 13. Santa Cruz, CA and Washington, DC: Center for Research on Education, Diversity & Excellence https://www.cal.org/ndlf/pdfs/publications/development-of-bilingualism-and-biliteracy-from-grade-3-to-5.pdf.
Howard, E. R., Lindholm-Leary, K., Rogers, D., Olague, N., Medina, J., Kennedy, B., Sugarman, J., & Christian, D. (2018). *Guiding principles for dual language education* (3rd ed.). Center for Applied Linguistics.
Howard, E. R., & Neugebauer, S. R. (2015). Moving towards biliteracy: Varying paths of bilingual writers in two-way immersion programs. *Revista Miriada Hispanica, 10*, 83–106.
Howard, E. R., & Sugarman, J. (2001). *Two-way immersion programs: Features and statistics*. ERIC Clearinghouse on Languages and Linguistics. https://www.cal.org/content/download/1577/16838/file/TwoWayImmersionPrograms.pdf.
Howard, E. R., Sugarman, J., & Christian, D. (2003). *Trends in two-way immersion education*. A review of the research (Report 63). Baltimore, MD. CRESPAR/Johns Hopkins University. Center for Research on the Education of Students Placed At Risk.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118
Jepsen, C. (2010). Bilingual education and English proficiency. *Education Finance and Policy, 5*(2), 200–227. https://doi.org/10.1162/edfp.2010.5.2.5204
de Jong, E. J. (2016). Two-way immersion for the next generation: Models, policies, and principles. *International Multidisciplinary Research Journal, 10*(1), 6–16. https://doi.org/10.1080/19313152.2016.1118667
Kim, Y. K., Hutchison, L. A., & Winsler, A. (2015). Bilingual education in the United States: An historical overview and examination of two-way immersion. *Educational Review, 67*(2), 236–252. https://doi.org/10.1080/00131911.2013.865593
Klicpera, C., & Gasteiger-Klicpera, B. (1993). *Lesen und Schreiben – Entwicklung und Schwierigkeiten: Die Wiener Längsschnittuntersuchungen über die Entwicklung, den Verlauf und die Ursachen von Lese- und Schreibschwierigkeiten in der Pflichtschulzeit* [Reading and Writing - development and Difficulties: The Vienna Longitudinal Studies on the Development, Course, and Causes of Reading and Writing Difficulties in Compulsory Schooling.]. Huber.
Kline, R. B. (2005). *Principles and practice of structural equation modeling*. Methodology in the social sciences. The Guilford Press.
Kolen, M. J [Michael J.], & Brennan, R. L [Robert L.] Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*, 2014 (3rd ed.). New York: Springer. https://doi.org/10.1007/978-1-4939-0317-7. Statistics for Social and Behavioral Sciences.
Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.
Krashen, S. D. (2005). The acquisition of academic English by children in two-way programs: What does the research say? In V. Gonzales, & J. Tinajero (Eds.), *Review of research and practice* (pp. 1–19). Lawrence Erlbaum Associates.
Lambert, W. E. (1984). An overview of issues in immersion education. In *Studies on immersion education: A collection for United States educators* (pp. 8–30). California State Department of Education.
Lehmann, R., & Lenkeit, J. (2008). *ELEMENT. Erhebung zum Lese- und Mathematikverständnis. Entwicklungen in den Jahrgangsstufen 4 bis 6 in Berlin Abschlussbericht über die Untersuchungen 2003, 2004 und 2005 an Berliner Grundschulen und grundständigen Gymnasien [ELEMENT. Survey on reading and mathematics comprehension. Trends in grades 4 to 6 in Berlin: Final report on the 2003, 2004, and 2005 surveys of Berlin elementary schools and high schools.]*. Berlin: Humboldt Universität zu Berlin. https://www.iqb.hu-berlin.de/fdz/studies/Element/element6_bericht.pdf.
Lindholm-Leary, K. (2001). Dual language education. *Bilingual education and bilingualism, 28* (Multilingual Matters).
Lindholm-Leary, K. (2011). Student outcomes in Chinese two-way immersion programs: Language proficiency, academic achievement and student attitudes. In D. J. Tedick, D. Christian, & T. W. Fortune (Eds.), *Immersion education: Practices, policies, possibilities* (pp. 81–103). Multilingual Matters.
Lindholm-Leary, K., & Block, N. (2010). Achievement in predominantly low SES/ Hispanic dual language schools. *International Journal of Bilingual Education and Bilingualism, 13*(1), 43–60. https://doi.org/10.1080/13670050902777546
Lindholm-Leary, K., & Genesee, F. (2014). Student outcomes in one-way, two-way, and indigenous language immersion education. *Journal of Immersion and Content-Based Language Education, 2*(2), 165–180. https://doi.org/10.1075/jicb.2.2.01lin
Lindholm-Leary, K., & Hernandéz, A. (2011). Achievement and language proficiency of Latino students in dual language programmes: Native English speakers, fluent English/previous ELLs, and current ELLs. *Journal of Multilingual and Multicultural Development, 32*(6), 531–545. https://doi.org/10.1080/01434632.2011.611596
Lindholm-Leary, K., & Howard, E. R. (2008). Language development and academic achievement in two-way immersion programs. Pathways to multilingualism: Evolving perspectives on immersion education. In T. W. Fortune, & D. J. Tedick (Eds.). *Bilingual education and bilingualism* (Vol. 66, pp. 177–200). Multilingual Matters.
Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods, 22*(1), 141–165. https://doi.org/10.1037/met0000096
Ludwig, C., Guo, K., & Georgiou, G. K. (2019). Are reading interventions for English language learners effective? A meta-analysis. *Journal of Learning Disabilities, 52*(3), 220–231. https://doi.org/10.1177/0022219419825855
Marian, V., Shook, A., & Schroeder, S. R. (2013). Bilingual two-way immersion programs benefit academic achievement. *Bilingual Research Journal, 36*(2). https://doi.org/10.1080/15235882.2013.818075
Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11* (3), 320–341. https://doi.org/10.1207/s15328007sem1103_2
McField, G. P., & McField, D. R. (2014). The consistent outcome of bilingual education programs: A meta-analysis of meta-analyses. In G. P. McField (Ed.), *The Miseducation of English learners: A tale of three states and lessons to be learned (267-299)*. Information Age Publishing.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 51, pp. 297–384). Elsevier Science. https://doi.org/10.1016/S0079-7421(09)51009-2.
Möller, J., Fleckenstein, J., Hohenstein, F., Paulick, I., Preusler, S., & Baumert, J. (2018). Varianten und Effekte bilingualen Lernens in der Schule. *Zeitschrift Für Erziehungswissenschaft, 21*(1), 4–28. https://doi.org/10.1007/s11618-017-0791-x
Möller, J., Hohenstein, F., Fleckenstein, J.,Köller, O., & Baumert, J. (Eds.). (2017). *Erfolgreich integrieren - die Staatliche Europa-Schule Berlin [Successful Integration – the State Europe School Berlin].* Waxmann.
Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework. International association for the evaluation of educational achievement*. http://timssandpirls.bc.edu/pirls2016/framework.html.
Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary school in 35 countries.* International Study Center.
Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (Eds.). (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries.* Boston College: TIMSS & PIRLS International Study Center.
Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.
Neugebauer, S. R., & Howard, E. R. (2015). Exploring associations among writing self-perceptions, writing abilities, and native language of English-Spanish two-way immersion students. *Bilingual Research Journal, 38*(3), 313–335. https://doi.org/10.1080/15235882.2015.1093039
Padilla, A. M., Fan, L., Xu, X., & Silva, D. (2013). A Mandarin/English two-way immersion program: Language proficiency and academic achievement. *Foreign Language Annals, 46*(4), 661–679. https://doi.org/10.1111/flan.12060
Preusler, S., Zitzmann, S., Paulick, I., Baumert, J., & Möller, J. (2019). Ready to read in two languages? Testing the native language hypothesis and the majority language hypothesis in two-way immersion students. *Learning and Instruction, 64*. https://doi.org/10.1016/j.learninstruc.2019.101247
R Core Team. (2021). *R: a language and environment for statistical computing*. https:// www.R-project.org/.
Robitzsch, A., Kiefer, T., & Wu, M. L. (2020). *TAM: Test analysis modules*. https://CRAN.R-project.org/package=TAM.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley. https://doi.org/10.1002/9780470316696
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037//1082-989X.7.2.147
Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338
Senatsverwaltung für Bildung, Jugend und Familie. (2018). *Rahmenvorgaben der staatlichen europa-schule Berlin (SESB) als Schule besonderer pädagogischer prägung [framework of the staatliche europa-schule Berlin (SESB) as a school with a special pedagogical character].* Berlin.
Serafini, E. J., Rozell, N., & Winsler, A. (2020). Academic and English language outcomes for DLLs as a function of school bilingual education model: The role of two-way immersion and home language support. *International Journal of Bilingual Education and Bilingualism, 12*(1), 1–19. https://doi.org/10.1080/13670050.2019.1707477
Slavin, R. E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research, 75*(2), 247–284. https://doi.org/10.3102/00346543075002247
Stanat, P., Becker, M., Baumert, J., Lüdtke, O., & Eckhardt, A. G. (2012). Improving second language skills of immigrant students: A field trial study evaluating the effects of a summer learning program. *Learning and Instruction, 22*(3), 159–170. https://doi.org/10.1016/j.learninstruc.2011.10.002
Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (2017). Effects of dual-language immersion programs on student achievement. *American Educational Research Journal, 54*(1S), 282S–306S. https://doi.org/10.3102/0002831216634463
Thomas, W. P., & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. UC Berkeley: Center for Research on Education, Diversity & Excellence. http://escholarship.org/uc/item/65j213pt.
Umansky, I. M., & Reardon, S. F. (2014). Reclassification patterns among Latino English learner students in bilingual, dual immersion, and English immersion classrooms. *American Educational Research Journal, 51*(5), 879–912. https://doi.org/10.3102/0002831214545110
Valentino, R. A., & Reardon, S. F. (2015). Effectiveness of four instructional programs designed to serve English learners. *Educational Evaluation and Policy Analysis, 37*(4), 612–637. https://doi.org/10.3102/0162373715573310
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627
Watzinger-Tharp, J., Swenson, K., & Mayne, Z. (2018). Academic achievement of students in dual language immersion. *International Journal of Bilingual Education and Bilingualism, 21*(8), 913–928. https://doi.org/10.1080/13670050.2016.1214675
Zhao, J. H., & Schafer, J. L. (2018). *pan: Multiple imputation for multivariate panel or clustered data.*