EMPIRICAL STUDY

First and Second Language Acquisition in German Children Attending a Kindergarten Immersion Program: A Combined Longitudinal and Cross-Sectional Study

Kirstin Bergström, Maria Klatte, Claudia Steinbrink, and Thomas Lachmann

This study investigated first (L1) and second (L2) language acquisition in two age-matched groups of 2- to 6-year-old kindergarten children over the course of 2.5 years. The immersion group participated in a partial English immersion program whereas the conventional instruction group received a conventional L2 course (30 minutes per week); the groups were comparable with respect to control variables (e.g., socioeconomic status). L1 and L2 competencies were assessed at four time points (pretest and three posttests). No detrimental effects of immersion on L1 development were found. Concerning L2 receptive skills, the immersion group outperformed the conventional instruction group in each posttest. Contact quality and quantity were found to relate to L2 acquisition through immersion. Immersion appears to be a successful method for early L2 learning; it fosters L2 receptive skills without any cost for the L1.

Keywords language development; input quantity and quality; second language learning; preschool; language instruction

This study was supported by a grant from the Ministry of the Interior, for Sports and Infrastructure of the German Federal State of Rhineland-Palatinate. Thanks are due to the city of Kaiserslautern, Department for Youth and Sports, for supporting us and handling the personnel recruitment and administration of the English educator. We want to thank the children, parents, and educators for their continuous cooperation over the course of the study. Many thanks are also due to Annette Rieksmeier, Birgit Daniel, Stella Klatte, and Anna Eifer for help in data collection.

Correspondence concerning this article should be addressed to Kirstin Bergström, Cognitive and Developmental Psychology, University of Kaiserslautern, Erwin-Schrödinger-Str. 57, 67663 Kaiserslautern, Germany. E-mail: kirstin.bergstroem@sowi.uni-kl.de

Language Learning 66:2, June 2016, pp. 386–418
© 2016 Language Learning Research Club, University of Michigan
DOI: 10.1111/lang.12162
Introduction
Mastery of foreign languages is becoming increasingly important in the modern world. Thus, education systems in Europe are attaching increasing importance to foreign language instruction (see Common European Framework of Reference for Languages; Council of Europe, 2009). Because there is evidence that learners who start second language (L2) learning early in life achieve a higher level of L2 proficiency when compared to older starters (Birdsong, 2005; DeKeyser & Larson-Hall, 2005; Johnson & Newport, 1989; see Genesee, 2004, for a discussion of the role of age in L2 acquisition), there is a trend to start L2 instruction as early as possible, often during preschool years (Education, Audiovisual and Culture Executive Agency, European Commission, 2012).

The diverse L2 instructional approaches can be assigned to a continuum from language to content driven (Met, 1998). In language-driven approaches, the goal is language learning (e.g., vocabulary, grammar) while content (e.g., nonacademic text material) is used only as a medium for teaching target language skills. Language-driven approaches are often called “conventional L2 instruction” (e.g., Genesee, 2004) and are typically given in the form of courses with a specified amount of lesson time per week. In content-based approaches, on the contrary, the L2 is used as the medium of instruction and is learned incidentally along with acquisition of knowledge in other domains, such as mathematics, science, and arts (Cloud, Genesee, & Hamayan, 2000).

Immersion programs in which instruction is provided in the L2 represent content-driven instructional approaches. Immersion programs differ with respect to a number of characteristics, including the starting point (early: beginning in kindergarten or first grade; middle: middle elementary grades; late: late elementary or high school), the amount of instruction given in the L2 (total immersion with 100% of instruction in the L2 vs. partial immersion with a minimum of 50% of instruction in the L2), and the ratio of first language (L1) instruction to L2 instruction at different stages within the program (increasing vs. decreasing extent of L2 instruction; for further characteristics, see Genesee, 1983, 2004).

Typically, immersion programs are provided for majority-language children, that is, for children whose L1 is the majority language of the country (Genesee, 2004). The aim of immersion has been defined as additive bilingualism, where the L2 is learned while maintaining the L1. Additive bilingualism is a core feature of a prototypical immersion program (for further core features, see Swain & Johnson, 1997) and has to be distinguished from subtractive bilingualism, in which the L2 develops at the cost of the L1. Subtractive bilingualism can occur when minority-language children learn the majority language of
the country where they reside (e.g., Herkenrath, 2012; Leseman, 2000; Wong Fillmore, 1991).

Since the first Canadian French immersion program (Lambert & Tucker, 1972), in which L2 instruction endured from kindergarten until the end of primary school, immersion programs have been implemented all over the world (e.g., Gebauer, Zaunbauer, & Möller, 2013; Jared, Cormier, Levy, & Wade-Woolley, 2011; Johnson & Swain, 1997). The existing research on the efficacy of foreign language immersion programs focuses on language and academic learning at schools, and only few of the studies include German children (e.g., Zaunbauer, Gebauer, & Möller, 2012). Therefore, the present study investigated language development in 2- to 6-year-old children in an immersion kindergarten. Specifically, we focused on early partial immersion for majority-language German children learning English as an additional language. The main goal of the present study was to show that, after careful control for a variety of confounding variables and elimination of a potential selection bias, L1 learning is not sacrificed when young children are in an immersion L2 setting.

Background Literature
L2 Acquisition in Educational Settings
Research has identified both method- and learner-related factors that affect the success of L2 acquisition. Concerning method-related factors, quantity and quality of L2 input play an important role (e.g., Moyer, 2011; Vermeer, 2001). Input quantity typically refers to the frequency and duration of the input, whereas input quality refers to the way in which input is presented to the learner. Examples of input quality variables are vocabulary sophistication, vocabulary diversity, input simplification or enhancement, and the form of corrective feedback (e.g., Lyster, 1998; Stafford, Bowden, & Sanz, 2012). For example, in a study with 4- to 7-year-old children, Vermeer showed that input quantity (frequency of oral and written language in elementary schools) was associated with children’s vocabulary knowledge. However, Moyer found that the impact on L2 proficiency was stronger for the quality of L2 experience (context-specific, interactive language use) than for the quantity (weekly hours using the L2). Hence, both input quantity and quality have to be taken into account when examining L2 acquisition (see also Norris & Ortega, 2000).

Concerning learner-related factors, several studies have revealed positive correlations between L1 verbal short-term memory and the acquisition of L2 vocabulary (Cheung, 1996; Gathercole & Baddeley, 1989; Martin & Ellis,
Effects of Language Immersion in Kindergarten

2012; Service & Kohonen, 1995) and grammar (French & O’Brien, 2008; Kormos & Safar, 2008). In addition, L2 competencies vary as a function of the learner’s general cognitive abilities, especially reasoning (Andringa, Olsthoorn, van Beuningen, Schoonen, & Hulstijn, 2012; Genesee & Hamayan, 1980). Furthermore, L2 outcomes are related to the learner’s socioeconomic status (SES). Children with lower parental SES typically achieve lower L2 proficiency compared to children with higher parental SES (Dale, Harlaar, & Plomin, 2012; Gathercole, 2002; Huaqing Qi, Kaiser, Milan, & Hancock, 2006). Hence, it is necessary to control for these learner-related factors when studying the efficacy of L2 instruction (for further learner-related factors, including L1 proficiency, L2 aptitude, and age of onset, see Birdsong, 2005; Sparks, Patton, Ganschow, & Humbach, 2009). In addition, both method- and learner-related factors are crucial for L1 development; examples of such factors include quantity and quality of child-directed speech (e.g., Rowe, 2012), verbal short-term memory (e.g., Gathercole & Baddeley, 1989; Gathercole, Service, Hitch, Adams, & Martin, 1999), and SES (e.g., Hoff, 2003, 2006). Therefore, the control of these factors is also important when examining L1 development.

Effects of School Immersion Programs on Language Development

Since the seminal study of Lambert and Tucker (1972), academic achievement and language development of majority-language children participating in immersion programs at school have been studied extensively (Genesee, 1983, 1987, 2004; Swain & Johnson, 1997). These studies have revealed that majority-language students in immersion classes achieve higher L2 proficiency compared to students with conventional language-driven L2 instruction (e.g., Barik & Swain, 1978; Genesee, 2004). The advantage of immersion has been reported for both comprehension and production skills. While immersion students often achieve L2 comprehension skills comparable to age-matched native speakers of the respective language, the productive language and grammar skills do not reach nativelike levels (e.g., Harley, Allen, Cummins, & Swain, 1990; Umbel, Pearson, Fernández, & Oller, 1992; Yan & Nicoladis, 2009). In Germany, a recent longitudinal study reported similar results with regard to English (L2) vocabulary development of German children (Zaunbauer et al., 2012). After controlling for cognitive abilities and SES, English vocabulary was found to increase significantly from Grade 1 to Grade 4 in both immersion students and students with conventional English instruction, though immersion students outperformed conventionally educated children.

Concerning L1 development, majority-language immersion students have been shown to develop comparable or even higher L1 competencies compared
to monolingually educated students (e.g., Bamford & Mizokawa, 1991; Cheng, Li, Kirby, Qiang, & Wade-Woolley, 2010; Harley, Hart, & Lapkin, 1986; Swain & Lapkin, 1982). The fact that L1 development is unimpaired by immersion is presumably due to students’ daily contact with their L1 outside school (e.g., at home, through media, and in the community), which provides sufficient L1 input for age-appropriate L1 development. In summary, the available studies confirm that immersion is an effective method for L2 acquisition at school and that immersion is not detrimental to students’ L1 development.

Effects of Immersion on Language Development in Kindergarten

Although many immersion programs start in kindergarten, evaluations of these programs usually target learning outcomes in school-age children (e.g., Barik & Swain, 1978; Genesee, 1981; Harley et al., 1986; Lambert & Tucker, 1972; Turnbull, Lapkin, & Hart, 2001). Out of the few published studies on language development in children attending bilingual kindergartens, most have addressed potential impairments in L1 development in minority-language children, with inconsistent results (e.g., for negative effects, see Leseman, 2000; Wong Fillmore, 1991; for null effects, see Rodríguez, Díaz, Duran, & Espinosa, 1995; Winsler, Díaz, Espinosa, & Rodríguez, 1999).

With respect to majority-language children, research has focused on L2 acquisition. In an early evaluation study by Barik and Swain (1975), children in a total French immersion group achieved higher receptive language competency in French (L2) when compared to control children receiving conventional (language-driven) French instruction. Both groups were comparable regarding cognitive ability, English (L1) proficiency, and SES. In a review of partial immersion preschool programs in Singapore, Eng, Gan, and Sharpe (1997) emphasized the importance of immersion quality for successful L2 acquisition. The authors compared L2 skills in 4- to 6-year-old children from private and nonprivate immersion preschools (with more than 50% of L2 English instruction). In both groups, the children’s English language skills were lower than in native English-speaking children. Furthermore, the children from nonprivate preschools performed worse in oral English fluency tasks than the children from private preschools. The poor L2 performance in the children from nonprivate preschools was explained through lower quality of L2 input, namely, inappropriate language learning materials and methods, poor classroom activities, and minimal verbal interaction between teachers and children. However, as SES was not controlled, the possibility that SES differences between the groups contributed to the differences in L2 proficiency cannot be ruled out. Such a confound is likely because groups with lower SES are
excluded from private preschools because of high tuition fees (see Eng et al., 1997, p. 197).

**Effects of Immersion in German Kindergartens**

In contrast to the North American kindergarten model, German kindergarten is not integrated into the school system. Accordingly, no formal reading and writing instruction is provided. Children typically attend kindergarten from age 2 or 3 until school entry at the age of 6 years. Kindergarten groups are typically mixed in age, usually with several simultaneous educators in a group of 10 to 25 children. Therefore, partial immersion in German kindergartens differs from that in North America. In North America, only one language per lesson is used by the teacher, meaning that some content is delivered in the L1 while other content is provided in the L2. In contrast, German preschool immersion programs typically follow the “one person–one language” principle (Döpke, 1992; Ronjat, 1913). That is, one native L2-speaking educator communicates exclusively in the children’s L2, while other educators always use the children’s L1. Both L1 and L2 are used during the whole kindergarten time. Thus, children continue to communicate in their L1 with peers and with the L1 educators. This means that a child can avoid immersion to a great extent if, for instance, the child is not motivated to use the L2 and resorts to using only the L1.

To our knowledge, there is only one published document focusing on L1 and L2 acquisition in children attending German immersion kindergartens—Early Language and Intercultural Acquisition Study (ELIAS), a collection of research reports edited by Kersten, Rohde, Schelletter, and Steinlen (2010); see Kersten (2010) for the public part of the final report. ELIAS revealed some essential results concerning children’s L2 (English) and L1 (German) vocabulary development. First, the children’s L2 receptive vocabulary increased significantly over a period of one year, but did not reach the level of a native English comparison group (Schelletter & Ramsey, 2010). Second, the receptive vocabulary of a group of children with higher input quality, assessed via an observational tool, was superior to that of a group of children with lower levels of input quality (Weitz, Pahl, Flyman Mattsson, Buyl, & Kalbe, 2010). Third, similar results were shown for input quantity, operationalized as L2 input per week (i.e., estimated time of L2 exposure in different preschools), where L2 receptive vocabulary was larger in the highest quantity group compared to three lower quantity groups, which did not differ from one another. Finally, the children reached average scores in a German test of language development, leading to the conclusion that L1 acquisition was unimpaired by English immersion (Steinlen, Neils, Piske, & Trumpp, 2010).
Although ELIAS yielded important findings concerning the efficacy of L2 immersion programs in German kindergartens, there are some methodological limitations to consider. First, the participation in the bilingual preschool program was voluntary, as is typically the case in immersion programs. This may result in a selection bias, because well-educated, highly motivated parents are more likely to choose immersion education programs for their children. This clearly limits generalizability of research findings (see Genesee, 2004). Second, parents’ SES was not controlled for, although SES has been shown to be an important predictor of L2 development (e.g., Dale et al., 2012; Gathercole, 2002; Huaqing Qi et al., 2006). Third, there was no pretest assessing either children’s level of L1 and L2 proficiency or any control variables (e.g., cognitive abilities) prior to the onset of the immersion program. Thus, a priori differences between the groups (e.g., high vs. low contact quality group) cannot be ruled out.

**The Present Study**

The objective of the present study was to investigate L1 and L2 acquisition in children attending a German-English immersion kindergarten compared to age- and gender-matched control children receiving conventional foreign language instruction in another kindergarten over 2.5 years. The conventional foreign language instruction was a weekly L2 English course with a focus on explicit instruction, which is very common in German kindergartens. Most importantly, in order to avoid selection bias, two kindergartens that did not have any foreign language instruction prior to the study were chosen in close cooperation with the local municipal administration, and English instruction was implemented as a standard curriculum for all children. That is, the immersion program was introduced in one kindergarten, and the conventional foreign language instruction was introduced in another kindergarten specifically for this study. When enrolling their children in kindergarten, the parents had no knowledge about the planned implementation of the English instruction. Furthermore, they did not know in which kindergarten the immersion or the conventional programs were to be implemented. In order to ensure comparability of the groups at the onset of L2 instruction, a pretest was administered to assess L1 and L2 development and control variables, which included cognitive abilities, verbal short-term memory, and SES. L1 and L2 measures were again assessed at three later time points (once per year).

These two groups, which only differed in a L2 learning method, thus allowed us to carry out a comparison of children’s L1 development over a 2.5-year
Effects of Language Immersion in Kindergarten

period. Furthermore, with respect to children’s L2 development, the efficacy of the immersion program could be examined in comparison to the conventional instruction group, which exemplified a common form of early foreign language education in Germany (a “traditional” course with explicit instruction). In addition, quantity and quality of L2 input was estimated for each child in the immersion group, enabling an examination of the impact of input quantity and quality on L2 learning at the individual level within this kindergarten group. Prior studies on the efficacy of early immersion programs focused on receptive L2 skills, as it is a generally accepted assumption that language comprehension precedes language production, particularly for children (e.g., Clark, 1993; Melka, 1997). L2 production is a difficult task for young L2 learners, and they often refuse to produce L2 utterances in testing situations. Thus, in this study, assessment of L2 production was confined to the later measurement points, that is, after 20 and 31 months of immersion.

In view of the special form of early partial immersion in German kindergartens, our predictions were as follows. First, we expected no detrimental effects of immersion on the development of L1 (German), as reflected by similar levels of L1 achievement in the immersion and conventional instruction groups (Hypothesis 1). Second, we expected a higher level of L2 (English) skills in the immersion group compared to the conventional instruction group in each of the three posttests (Hypothesis 2). Third, we expected both contact quantity and contact quality to be associated with L2 skills in the immersion group (Hypothesis 3).

Method

Participants

There were two learner groups with different English learning programs: the immersion group and the conventional instruction group. As is typical in German kindergartens, both groups were age-mixed, with ages ranging from 2 to 6 years. Furthermore, in each year, older children left the groups for school entry, and new children entered the groups. Data were collected at four measurement points over 2.5 years: Time 1 (T1) pretest, T2 after 8 months of English instruction, T3 after 20 months, and T4 after 31 months. The conventional instruction and immersion programs were implemented in two separate kindergartens located in the same catchment area in Kaiserslautern. Because the programs were implemented as standard curricula in the kindergartens, all kindergarten children participated in the intervention. For all children participating in the study, parents gave written consent. All participating children spoke only German at home and were unfamiliar with any other languages.
Cross-Sectional Sample
Because the control kindergarten (conventional instruction in three groups with 25 children in each) was larger than the immersion kindergarten (one group of 20–24 children), a control child matched for age and gender was chosen for each child in the immersion group. The resulting matched sample consisted of a total of 27 children per group across the study period of 2.5 years (for seven children attending the immersion program, no child of same age and gender was available in the conventional instruction group). However, the sample sizes varied across the measurement time points due to school entry and new children entering the groups (see Figure 1). The resulting matched cross-sectional samples and their demographic characteristics are summarized in Table 1.

Longitudinal Sample
Out of the final sample, 20 children were tested across all four time points and thus provided complete longitudinal data. Because only matched pairs (a child in the immersion group and the matched child in the conventional instruction group) were lost over the study period of 2.5 years because of school entry, the two groups remained matched at all measurement time points (T1, T2, T3, T4) with respect to SES, gender, and age. The data of these children (immersion group: \( n = 10 \), 5 female/5 male, \( M_{age} = 3;0 \), range = 2;1–3;9; conventional instruction group: \( n = 10 \), 5 female/5 male, \( M_{age} = 3;1 \), range = 2;1–3;9) were used for the longitudinal analyses.

Sample Used in Analyses of Contact Quantity and Quality
For the analyses addressing the impact of contact quality and quantity on L2 learning in the immersion group, no comparison with the conventional instruction group was required. Thus, the data from all children in the immersion group (including those for whom no matched control children in the conventional instruction group were available) were included in the correlation analyses (T2: \( n = 22 \), 10 female/12 male, \( M_{age} = 4;10 \), range = 2;10–6;5; T3: \( n = 24 \), 14 female/10 male, \( M_{age} = 4;7 \), range = 3;0–6;5; T4: \( n = 24 \), 15 female/9 male; \( M_{age} = 5;1 \), range = 2;11–6;5). The same correlation analyses were conducted for the conventional instruction group data (T2: \( n = 25 \), 13 female/12 male, \( M_{age} = 4;11 \), range = 3;0–6;5; T3: \( n = 21 \), 12 female/9 male, \( M_{age} = 4;10 \), range = 2;9–6;5; T4: \( n = 20 \), 12 female/8 male; \( M_{age} = 5;0 \), range = 3;0–6;4).
Figure 1 Sample sizes varying across measurement time points (T1 pretest, T2 after 8 months of English instruction, T3 after 20 months, T4 after 31 months). Vertical borders designate cross-sectional samples from T2, T3, and T4 (N = total sample size; nIG = sample size of the immersion group; nCG = sample size of the conventional instruction group).
Table 1  Characteristics of cross-sectional samples

<table>
<thead>
<tr>
<th>Testing time</th>
<th>Immersion</th>
<th>Conventional instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$ (f/m)</td>
<td>Mean age (range)</td>
</tr>
<tr>
<td>Time 1 (T1)</td>
<td>20 (11/9)</td>
<td>4;0 (2;1–5;8)</td>
</tr>
<tr>
<td>Time 2 (T2)</td>
<td>20 (11/9)</td>
<td>4;9 (2;10–6;8)</td>
</tr>
<tr>
<td>Time 3 (T3)</td>
<td>18 (10/8)</td>
<td>4;8 (3;0–6;5)</td>
</tr>
<tr>
<td>Time 4 (T4)</td>
<td>17 (10/7)</td>
<td>5;0 (2;10–6;4)</td>
</tr>
</tbody>
</table>

**English Learning Programs**

Following the principle “one person–one language” in the immersion group, one native British English-speaking educator interacted solely in English with the children, while the other educators in the team used German. The English-speaking educator was recruited especially for this study and was schooled in immersion principles (contextualization, recasts, feedback rules, etc.) prior to the intervention using ELIAS teacher training materials (Burmeister & Steinlen, 2009). The schooling emphasized the importance of contextualizing and visualizing English utterances to enable children to infer meaning, avoidance of negative feedback, translation of children’s German utterances into English, and constantly encouraging, but not forcing, children to speak English. Special importance was attached to corrective feedback and recasts; children’s utterances should be taken up, paraphrased, and expanded to provide a correct model (e.g., child: “Sarah will water haben, nicht tea”; educator: “Ah, Sarah wants to have water, but not tea”). The children in the immersion group spent 4 to 7 hours per day in the kindergarten.

In the conventional instruction group, the same English educator held a playful L2 English course with explicit instruction (English language games, songs, finger plays, etc.). This weekly course lasted 30 minutes, as is customary in those kindergartens where English instruction is provided; it dealt with a variety of topics (e.g., colors, numbers, animals, body parts, furniture, and forms). A substantial portion of the L2 vocabulary tested through the test battery (see below) was provided explicitly in this course. Various course materials, such as books with lesson plans, songs, and finger plays (e.g., Fiedler, 2009), were provided and used by the English educator. Due to the large size of the conventional instruction kindergarten, the children were divided into four groups with 8 to 14 children of similar age. The courses for the groups were held consecutively in a separate room in the kindergarten.
Questionnaires

Parent Questionnaire

SES was estimated by means of a parent questionnaire addressing (a) the mother’s and father’s level of schooling (no graduation to A-levels/Abitur), (b) level of professional education (no qualification to Ph.D.), and (c) total monthly family income. The scores for level of educational achievement and level of professional education were averaged to an educational score (2–11 levels) for the mother and father separately, which were then combined with income scores, yielding a composite SES score for both parents (3–19).

Contact Questionnaire

For the immersion group, weekly ratings of contact quality and quantity were provided by the English educator for each individual child. For contact quantity, the total time (in hours) a child spent interacting with the English educator was estimated. This total contact time included both group and one-on-one interactions. Contact quality (i.e., the intensity of the contact) was assessed by means of a rating scale with three categories: low (little interaction, child avoids contact), medium (child does not actively initiate the contact, but is willing to participate on request), and high (child responds verbally or nonverbally in communicative situations; child actively initiates English contact almost daily). The weekly contact quantity was summed to obtain a total contact quantity score. The weekly contact quality scores were averaged to a mean contact quality score. For the conventional instruction group, the English educator kept an attendance list to get the number of attended courses as a measure of contact quantity.

Test Battery

The test battery consisted of two L1 German measures, four L2 English measures, two measures of nonverbal intelligence, and one test of verbal short-term memory.

German Language Measures

The Aktiver Wortschatztest für 3- bis 5-jährige Kinder (AWST-R [Productive Vocabulary Test for 3- to 5-Year-Old Children]; Kiese-Himmel, 2005) is a standardized picture-naming test assessing productive German vocabulary in 3- to 5-year-old children; it requires children to name pictures representing objects and actions. Two parallel versions of the test, with 37 items in each, were constructed by assigning even- and odd-numbered items from the original test to different sets. The resulting test versions were alternated at the four measurement time points. Test-retest reliabilities were: $r = .89$ (T1-T2), $r = .83$ (T2-T3), and $r = .87$ (T3-T4).
The Test for Reception of German Grammar (TROG-D; Fox, 2006) is a standardized test of comprehension of German grammar for children aged 3 to 10 years old; it requires children to select, out of four alternatives, the picture that corresponds to a target sentence. The syntactic complexity of the target sentences increased in the course of the test. Two parallel versions of the test were constructed by dividing the total set of sentences into 42 pairs with equal syntactic structure and difficulty. From each pair, one sentence was assigned to version A and one to version B. The resulting versions were alternated at the four time points. Test-retest reliabilities of the two versions were: \( r = .82 \) (T1-T2), \( r = .69 \) (T2-T3), and \( r = .70 \) (T3-T4).

**English Language Measures**

An English Receptive Vocabulary Task was constructed for the purpose of this study and was confined to the pretest (T1). This task was included to confirm that the children had virtually no knowledge of English prior to the implementation of the learning programs. The test consisted of 10 English words representing common and concrete nouns (e.g., tree). In each trial, a word was presented in combination with a corresponding picture and three distractors. The children were asked to point to the picture corresponding to the target word. Test reliability (Cronbach’s alpha) was .65.

The British Picture Vocabulary Scale–Third Edition (BPVS III; Dunn, Dunn, & National Foundation for Educational Research, 2009) is a standardized test of English receptive vocabulary in 2- to 16-year-old children. In this task, children were asked to select, out of four alternatives, the picture corresponding to a spoken English target word. The test consists of 14 sets of 12 increasingly difficult words (e.g., ball, happy, binoculars). Testing was terminated when eight or more errors were produced in a single set. Test-retest reliabilities were: \( r = .71 \) (T2-T3) and \( r = .73 \) (T3-T4).

An English Act-Out Task was constructed by the researchers in order to assess L2 sentence comprehension in interactive communication situations. In the context of an interactive playing situation, the children were supposed to follow instructions of an English-speaking hand puppet. The instructions were spoken by the experimenter operating the puppet. Three kinds of instructions were used: (a) movement commands (e.g., “Clap your hands”), (b) instructions embedded in a short story played with Playmobil® figures (e.g., “The boy is playing with the train”), and (c) instructions concerning actions with different objects (e.g., “Throw the tissue into the rubbish bin, please”). Using a predetermined scoring manual, the reactions of the children were scored according to the number of correctly solved elements in each instruction (maximum 112
Response consistency ($T_2 \alpha = .96; T_3 \alpha = .97; T_4 \alpha = .97$) and test-retest reliabilities ($T_2-T_3 r = .85; T_3-T_4 r = .84$) for this task were high.

The Expressive Vocabulary Test–Second Edition (EVT-2; Williams, 2007) is a standardized productive English vocabulary test for persons aged between 2 years 6 months and 90 years. As language production lags behind reception (Clark, 1993), the test was only conducted at $T_3$ and $T_4$. The test consists of colored pictures representing objects or actions. The children were asked to say the English word corresponding to a given picture (e.g., cat, thumb). When the children named the picture in German, they were reminded to use English. The difficulty of the items increased in the course of the test. Testing was terminated when errors were produced in five consecutive trials. Test-retest reliability was $r = .59$ ($T_3-T_4$). According to the manual (Williams, 2007), the range of test-retest reliability (interval of 2 to 8 weeks) of this test is $r = .94-.96$.

**Control Variables**

The Snijders-Oomen non-verbaler Intelligenztest für 2,5- bis 7-jährige Kinder (SON-R 2.5-7; Intelligence Test for 2.5- to 7-Year-Old Children; Tellegen, Laros, & Petermann, 2007) is a standardized nonverbal intelligence test for children aged between 2.5 and 7 years. Two subtests (Categories and Analogies) were administered. In both subtests, the children were instructed to detect sorting rules and to sort test cards or stones according to the respective rules. Item difficulty increased during the course of the subtests. Test reliability for both subtests was high (Categories $\alpha = .90$; Analogies $\alpha = .88$).

Children’s verbal short-term memory was assessed by means of the Nonword Repetition subtest from the Sprachentwicklungstest für drei- bis fünfhjährige Kinder–Subtest Phonologisches Arbeitsgedächtnis für Nichtwörter (SETK 3-5; Grimm, 2001), a standardized test for language development in German children. The children were instructed to repeat 18 nonwords increasing in length from two to five syllables. One point was given for each nonword correctly repeated. Cronbach’s alpha for this task was .85.

**Procedure and Analyses**

To increase standardization and objectivity of the tests, audio recordings of the verbal test items (nonwords, words, sentences) were created. The items were produced by a trained female speaker in a sound-attenuated laboratory and recorded with a 44,100 Hz sampling rate and 16-bit resolution. In the testing sessions, the recordings of the verbal items were presented via loudspeakers, with the respective pictures shown on a notebook display. At all four testing points, the children were tested individually in a quiet room in their
kindergarten. The tests were conducted by members of the research team over two to three sessions lasting 20–30 minutes each.

A combined cross-sectional and longitudinal design was used. In the cross-sectional design, all available data at a given time point were used. The groups were compared by means of *t* tests using bootstrapping (Efron & Tibshirani, 1993), because the data of some tests at several measurement points in one or both groups were not normally distributed (10 out of 38 Kolmogorov-Smirnov tests, *p* < .05). Bootstrapping is robust to violations of distribution normality because this test does not rely on the form of the underlying probability distribution of the variable. Bias-corrected and accelerated (BCa) bootstrap (asymmetrical) confidence intervals are reported. In the longitudinal design, analyses were confined to the data from children who attended kindergarten during the complete study period, thus allowing us to examine individual developmental curves. The longitudinal data were analyzed using two-way factorial mixed analyses of variance (ANOVAs) with L2 (English) learning method as a between- and time as a within-subjects factor. In addition, the impact of contact quality and quantity on L2 acquisition in the immersion group was assessed by means of correlation analyses.

L1 and L2 competencies were assessed at each of the four time points. Assessment of L2 production was confined to T3 and T4. Variables affecting L2 acquisition, including SES, general cognitive abilities, and verbal short-term memory, were assessed only at T1 (pretest) in order to ensure the comparability of the groups. Contact quality and quantity in the immersion group were measured through weekly ratings by the English educator. The dependent variable in all analyses was the number of correct responses.

**Results**

Analyses of the pretest data for 54 children (27 per group) in the final sample confirmed that the groups did not differ with respect to SES, intelligence, verbal short-term memory, or language abilities prior to the implementation of the English learning programs (all *p*s > .10; see Table 2). The same analyses were performed on the data from the children in the longitudinal sample (*n* = 20), yielding no differences between the groups (all *p*s > .50).

**Immersion and L1 and L2 Development**

*Cross-Sectional Analyses*

To examine the effect of immersion on L1 German development, separate *t* tests were conducted for each posttest (T2, T3, T4). At each time point, the *t* tests confirmed comparable performance in the immersion and conventional
### Table 2 Summary of control variables and language measures at pretest and results of independent-samples *t*-tests (based on 1,000 bootstrap samples)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Immersion</th>
<th>Conventional instruction</th>
<th>Bootstrap <em>t</em>-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>M</em></td>
<td><em>SD</em></td>
<td><em>n</em></td>
</tr>
<tr>
<td>Mother’s education</td>
<td>8.15</td>
<td>2.03</td>
<td>26</td>
</tr>
<tr>
<td>Father’s education</td>
<td>9.42</td>
<td>1.60</td>
<td>26</td>
</tr>
<tr>
<td>SES parents&lt;sup&gt;a&lt;/sup&gt;</td>
<td>15.21</td>
<td>2.88</td>
<td>21</td>
</tr>
<tr>
<td>SON-R analogies</td>
<td>7.26</td>
<td>4.04</td>
<td>27</td>
</tr>
<tr>
<td>SON-R categories</td>
<td>6.56</td>
<td>4.00</td>
<td>27</td>
</tr>
<tr>
<td>SETK 3–5–PGN&lt;sup&gt;b&lt;/sup&gt;</td>
<td>7.39</td>
<td>5.11</td>
<td>23</td>
</tr>
<tr>
<td>TROG-D</td>
<td>18.56</td>
<td>11.68</td>
<td>27</td>
</tr>
<tr>
<td>AWST-R</td>
<td>17.89</td>
<td>9.83</td>
<td>27</td>
</tr>
<tr>
<td>L2 receptive vocabulary</td>
<td>3.41</td>
<td>2.85</td>
<td>27</td>
</tr>
</tbody>
</table>

*Note.* <sup>a</sup>*n* is reduced because of missing income information of the parents. <sup>b</sup>*n* is reduced because four children refused to repeat nonwords. *SE* = standard error; CI = confidence interval; SON-R = intelligence test; SETK 3–5–PGN = verbal short-term memory test; TROG-D = German receptive grammar test; AWST-R = German productive vocabulary test.

The conventional instruction groups for both L1 receptive grammar (TROG-D) and L1 receptive vocabulary (AWST-R, all ps > .38; see Table 3). Thus, no detrimental effect of immersion on L1 achievement was found, supporting Hypothesis 1. As shown in Table 3, there was no visible increase in the scores, because the samples from T2 to T4 differed. The mean age of the samples from T2 to T4 did not change (see Table 1), so an improvement in the cross-sectional analysis was not expected.

Concerning the effect of immersion on L2 English receptive skills (BPVS III, English Act-Out Task), *t*-tests confirmed better performance in the immersion group when compared to the conventional instruction group in each of the three posttests (Hypothesis 2). The effect sizes of the group differences (Cohen’s *d*) were strong (see Table 4). With respect to L2 productive vocabulary (EVT-2), the analysis showed significant group differences with strong effect sizes in favor of the immersion group at T3. However, this advantage was no longer evident at T4.

### Longitudinal Analyses

A two-way (group × time) ANOVA comparing L1 receptive grammar (TROG-D) scores revealed a significant main effect of time, *F*(3, 54) = 37.28,
Table 3 Summary of German language measures for the cross-sectional samples at T2, T3, and T4 and results of independent-samples t tests (based on 1,000 bootstrap samples)

<table>
<thead>
<tr>
<th>Time</th>
<th>Measure</th>
<th>Immersion</th>
<th>Conventional instruction</th>
<th>Bootstrap t test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>SD</td>
<td>n</td>
</tr>
<tr>
<td>T2</td>
<td>TROG-D</td>
<td>26.15</td>
<td>10.36</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>AWST-R</td>
<td>24.95</td>
<td>6.19</td>
<td>20</td>
</tr>
<tr>
<td>T3</td>
<td>TROG-D</td>
<td>25.78</td>
<td>11.03</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>AWST-R</td>
<td>27.39</td>
<td>7.35</td>
<td>18</td>
</tr>
<tr>
<td>T4</td>
<td>TROG-D</td>
<td>28.24</td>
<td>7.92</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>AWST-R</td>
<td>24.12</td>
<td>6.71</td>
<td>17</td>
</tr>
</tbody>
</table>

Note. No improvement in the average values is visible because the samples from T2 to T4 differ (see Table 1). SE = standard error; CI = confidence interval; TROG-D = German receptive grammar test; AWST-R = German productive vocabulary test.

\( p < .001, \eta_p^2 = .67 \), but no significant main effect of group, \( F(1, 18) = 0.12, p = .74 \), \( \eta_p^2 = .01 \), and no significant group × time interaction, \( F(3, 54) = 0.21, p = .89 \), \( \eta_p^2 = .01 \). Similar results were obtained for L1 productive (AWST-R) vocabulary, with a significant main effect of time, \( F(3, 54) = 52.10, p < .001, \eta_p^2 = .74 \), and no significant group, \( F(1, 18) = 0.19, p = .67, \eta_p^2 = .01 \), or interaction, \( F(3, 54) = 0.15, p = .93, \eta_p^2 = .01 \), effects. Figure 2 illustrates the equivalence of developmental courses in the immersion and conventional instruction groups for both L1 measures (receptive grammar, productive vocabulary), with age-dependent performance improvements over time, which did not differ between the two groups (Hypothesis 1). Thus, as in the cross-sectional analyses, no detrimental effect of immersion on L1 achievement was found.

A two-way ANOVA targeting the development of L2 receptive vocabulary (BPVS III) revealed significant main effects of time, \( F(2, 36) = 6.28, p < .01, \eta_p^2 = .26 \), and group, \( F(1, 18) = 7.77, p < .05, \eta_p^2 = .30 \), but no significant two-way interaction, \( F(2, 36) = 0.78, p = .47, \eta_p^2 = .04 \). As indicated in Figure 3 (left panel), the English receptive vocabulary improved over time, and the children in the immersion group outperformed the conventional instruction children at T2, T3, and T4 (Hypothesis 2). For the English Act-Out Task, a similar two-way ANOVA again revealed significant main effects of time, \( F(2, 36) = 70.57, p < .001, \eta_p^2 = .80 \), and group, \( F(1, 18) = 10.15, p < .01, \eta_p^2 = .36 \), as well as a significant group × time interaction, \( F(2, 36) = 5.53, p < .01, \eta_p^2 = .24 \), confirming stronger performance improvements in
Table 4  Summary of English language measures for the cross-sectional samples at T2, T3, and T4 and results of independent-samples $t$ tests (based on 1,000 bootstrap samples)

<table>
<thead>
<tr>
<th>Time</th>
<th>Measure</th>
<th>Immersion</th>
<th>Conventional instruction</th>
<th>Bootstrap $t$ test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$n$</td>
</tr>
<tr>
<td>T2</td>
<td>BPVS III</td>
<td>25.65</td>
<td>13.06</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Act-out task</td>
<td>46.30</td>
<td>34.51</td>
<td>20</td>
</tr>
<tr>
<td>T3</td>
<td>BPVS III</td>
<td>24.83</td>
<td>14.94</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Act-out task</td>
<td>48.88</td>
<td>34.18</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>EVT-2</td>
<td>10.44</td>
<td>10.89</td>
<td>18</td>
</tr>
<tr>
<td>T4</td>
<td>BPVS III</td>
<td>23.41</td>
<td>17.95</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Act-out task</td>
<td>57.12</td>
<td>36.50</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>EVT-2</td>
<td>6.00</td>
<td>8.42</td>
<td>17</td>
</tr>
</tbody>
</table>

Figure 2 L1 grammar (TROG-D, left) and vocabulary (AWST-R, right) development of the children with full longitudinal data ($n = 20$) for the immersion and conventional instruction groups (there is no performance improvement in vocabulary from T3 to T4 due to ceiling effects). Error bars enclose ± 1 SE.
Figure 3 Development of L2 language comprehension (BPVS III, left; English Act-Out Task, right) of the children with full longitudinal data ($n = 20$) for the immersion and conventional instruction groups (both groups showed comparable performance for English receptive vocabulary at T1). Error bars enclose $\pm 1 \text{SE}$.
Table 5 Correlations between children’s performance in the English language tests and contact quantity (in hours) and quality at T2, T3, and T4

<table>
<thead>
<tr>
<th>Time</th>
<th>Measure</th>
<th>Immersion</th>
<th>Conventional instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>quantity</td>
</tr>
<tr>
<td>T2</td>
<td>BPVS III</td>
<td>22</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>Act-out task</td>
<td>22</td>
<td>.16</td>
</tr>
<tr>
<td>T3</td>
<td>BPVS III</td>
<td>24</td>
<td>.64**</td>
</tr>
<tr>
<td></td>
<td>Act-out task</td>
<td>24</td>
<td>.61**</td>
</tr>
<tr>
<td></td>
<td>EVT-2</td>
<td>24</td>
<td>.46*</td>
</tr>
<tr>
<td>T4</td>
<td>BPVS III</td>
<td>24</td>
<td>.61**</td>
</tr>
<tr>
<td></td>
<td>Act-out task</td>
<td>24</td>
<td>.72***</td>
</tr>
<tr>
<td></td>
<td>EVT-2</td>
<td>24</td>
<td>.43*</td>
</tr>
</tbody>
</table>


the immersion compared to the conventional instruction group (Figure 3, right panel).

Finally, the ANOVA comparing the development of L2 productive vocabulary between T3 and T4 revealed a significant main effect of group, $F(1, 18) = 5.79, p < .05, \eta^2_p = .24$, but no significant effect of time, $F(1, 18) = 1.58, p = .23, \eta^2_p = .08$, or a significant interaction effect, $F(1, 18) = 1.94, p = .18, \eta^2_p = .10$. As the $F$ value over 1 for the interaction effect was considerably larger than in the other analyses, Bonferroni-corrected tests of simple main effects showed better productive knowledge in the immersion group ($M = 9.40, SD = 9.92$) than in the conventional instruction group ($M = 0.20, SD = 0.42$) at T3, $t(18) = 2.93, p < .01, d = 1.18$, but no group differences between the immersion group ($M = 9.20, SD = 9.85$) and the conventional instruction group ($M = 4.10, SD = 4.93$) at T4, $t(18) = 1.46, p = .16, d = 0.69$. Performance in the conventional instruction group improved from T3 to T4, $t(9) = 2.68, p < .05, d = 1.38$, but no such improvement was evident in the immersion group, $t(9) = 0.08, p = .94, d = 0.02$.

Contact Quantity and Quality and L2 Development

The impact of contact quantity and quality on L2 acquisition (BPVS III, English Act-Out Task, EVT 2) in the immersion group was assessed by means of correlation analyses (Hypothesis 3), summarized in Table 5. At T2, none of the correlations between the contact measures and English language measures reached significance. At T3 and T4, each of the three English measures was
significantly correlated with contact quantity. Contact quality was significantly associated with both English receptive tests (BPVS III, English Act-Out Task) at T3 and T4. Furthermore, English production (EVT-2) was significantly associated with contact quality at T3, but no correlation was found at T4. The correlations between input quantity (number of attended courses) and English language measures in the conventional instruction group (also shown in Table 5) were similar to those of the immersion group, with lower correlations for the BPVS III but comparable correlations for the English Act-Out Task at T4.

**Discussion**

The present study examined the effects of immersion on L1 and L2 acquisition in German kindergarten children. Prior work has shown that immersion is an effective L2 learning method in school settings (Barik & Swain, 1978; Genesee, 1983, 1987, 2004; Lambert & Tucker, 1972), but little is known about the effects of immersion in kindergartens. Moreover, most studies are cross-sectional and thus do not allow for investigating L2 developmental pathways. In Germany, there has been only one study examining L1 and L2 development of children in immersion kindergartens (ELIAS; Kersten, 2010), but its results are difficult to interpret because a priori group differences cannot be ruled out. To overcome these limitations, this study examined the effects of immersion on L1 and L2 language development longitudinally over the course of 2.5 years, using two comparable groups of kindergarten children. We discuss our findings with respect to (a) the effects of immersion on L1 development, (b) the effects of immersion on L2 acquisition, and (c) the impact of contact quantity and quality on L2 development in immersion kindergartens.

**Immersion and L1 Development**

Both cross-sectional and longitudinal analyses confirmed comparable performance in the L1 measures (productive vocabulary, receptive grammar) in the immersion and conventional instruction groups after 8, 20, and 31 months of instruction in English. Thus, no detrimental effect of immersion on L1 development was found, supporting Hypothesis 1. This result is consistent with the findings of the ELIAS study (Kersten, 2010; Steinlen et al., 2010) and with prior research addressing the effects of immersion in majority-language children with non-German L1 in school contexts (Cheng et al., 2010; Cromdal, 1999; Harley et al., 1986; Lambert & Tucker, 1972; Swain & Lapkin, 1982; Yan & Nicoladis, 2009). However, in prior studies, L1 competence was not assessed prior to the onset of immersion, and thus the conclusion of no detrimental effects might not have been justified. This is because average performance after immersion could
be based on above-average results before the start of the immersion program and could thus reflect a deterioration in L1 competence. This argument is not farfetched in view of the potential selection bias, as well-educated and highly motivated parents are especially willing to opt for an immersion kindergarten.

In the current study, a pretest ensured the comparability of the groups with respect to L1 development and with respect to other variables known to affect L2 acquisition, such as SES, cognitive abilities, and verbal short-term memory. Thus, the conclusion that early partial immersion has no detrimental effects on L1 development is justified, to the best of our knowledge, for the first time.

The null effect of immersion on L1 development was expected. Children enrolled in a partial immersion program receive sufficient L1 input at home and by L1 educators and children in the kindergarten group, making a detrimental effect on L1 development unlikely. However, the conclusion that L1 is unaffected by immersion has to be restricted to the language components considered in our study (productive vocabulary, receptive grammar), because other components (e.g., phonology, pragmatics) were not examined. In addition, only majority-language children were tested; therefore, this finding cannot be generalized to minority-language children. When children learn the majority language as their L2 at school, their L1 is a minority language that is usually not used in public life. In addition, the status of L1 and SES possibly differ in minority- and majority-language children, making it more likely for differences in L1 acquisition to emerge during L2 learning.

**Immersion and L2 Development**

The children from the immersion group outperformed the children in the conventional instruction group in all L2 measures, supporting Hypothesis 2. The group differences were more pronounced for L2 receptive skills, that is, receptive vocabulary and sentence comprehension in interactive situations, than for L2 production measures. Concerning the receptive measures, the cross-sectional analyses revealed group differences with strong effect sizes, and the longitudinal analyses confirmed significant gains in L2 receptive vocabulary and sentence comprehension in both groups, and stronger improvements in sentence comprehension for the immersion group. These findings indicate that immersion is an effective method of early foreign language learning, specifically with respect to L2 receptive skills.

This study attributes group differences in L2 receptive competence to the different L2 learning methods, because the groups were highly comparable with respect to other variables that typically affect L2 acquisition (age, gender, parental SES, cognitive abilities, verbal short-term memory, L1 development,
and prior L2 knowledge). However, the learning programs differed not only with respect to the kind of instruction used (immersion vs. explicit instruction), but also with respect to L2 contact time, reaching several hours per day in the immersion group compared to about 30 minutes per week in the conventional instruction group. This difference in contact time could have mediated group differences in L2 performance. In fact, studies on L2 development show a positive relation between contact time and L2 outcomes (Kersten, 2010; Moyer, 2011; Vermeer, 2001; Weitz et al., 2010). However, in view of young children’s needs and their limited attention span, it is impossible to provide conventional early L2 instruction for several hours per day. With immersion, on the other hand, L2 learning proceeds largely incidentally, as a by-product of the children’s regular kindergarten activities (e.g., playing, doing handicrafts, eating). As the current study aimed to compare the efficacy of two L2 learning methods suited for kindergarten children, controlling for contact time was both unreasonable and impractical.

Furthermore, the two L2 learning methods differ with respect to the amount of variation of on-task time within each group. In immersion settings in German kindergartens, some children can avoid or actively search for communication with a L2-speaking educator, whereas in the conventional instruction group every child participating in a course gets the same input. This difference is reflected through significantly higher standard deviations in the immersion group compared to the conventional instruction group. The finding that the immersion group outperformed the conventional instruction group, as reflected through higher mean scores of the immersion group in the L2 English tests, does not mean that the immersion method is preferred for every child, because the on-task time depends not only on the English educator but also on the motivation of the child.

Both L2 learning methods examined here are age appropriate. While L2 learning in early immersion was incidental in contextual, meaningful activities, similar to the way L1 is acquired, in the conventional L2 course, the learning was more explicit yet also age appropriate (carried out through playful instruction with finger plays, songs, etc.). Even though older learners, compared to younger ones, are more efficient L2 learners because of higher cognitive maturity (see Genesee, 2004), an early start of L2 learning should be preferred assuming that a higher level of L2 proficiency (e.g., nativelike pronunciation) is more likely to be achieved when the L2 is introduced early in life (Birdsong, 2005; DeKeyser & Larson-Hall, 2005; Johnson & Newport, 1989). However, the role of cognitive maturity or readiness, as it applies to specific L2 learning methods, was not directly examined here.
Despite continuous improvements in L2 receptive vocabulary, the scores achieved by the children fell far below the monolingual norms (conversion into norms of the youngest norm group was even not possible). This finding mirrors the results from the ELIAS study, in which children in German immersion kindergartens did not reach nativelike levels in L2 reception (Kersten, 2010; Schelletter & Ramsey, 2010). Nativelike L2 competence levels reported in prior studies with older children (e.g., Barik & Swain, 1978; Genesee, 2004) may be explained through differences in the immersion method between kindergartens and schools. In the latter, children have to produce L2 output, which might enhance L2 competence levels. Furthermore, in school settings, children often have no possibility to communicate in their L1 with their educators.

Both English learning methods (immersion, conventional course) seemed to promote L2 production only in a limited way. As language comprehension typically precedes production (e.g., Clark, 1993), L2 production was assessed for the first time after 20 months of L2 learning. In both groups, children produced only a few words (maximum 10), which was far below the norms for 2;6–2;7-year-old native English-speaking children (Williams, 2007). While the children of the immersion group produced more words than the children in the conventional instruction group after 20 months (T3), no group difference was found after 31 months (T4). This is likely due to a significant increase in the conventional instruction group’s production performance from T3 to T4, whereas no improvement occurred in the immersion group.

There are different explanations for the limited improvements in L2 productive vocabulary. First, L2 production of a young child in test situations may vary substantially with the child’s mental state and motivation, resulting in a loss of reliability. In fact, the L2 productive test was much less reliable when compared to the L2 receptive tests. Second, the absence of L2 productive vocabulary improvements from T3 to T4 in the immersion group may be explained by a fossilization process resulting from children’s fulfilling their L2 communicative needs with relatively limited vocabulary (Kersten, 2010). According to this account, there was no need for children to improve L2 production beyond the level achieved at T3 because this level was sufficient to communicate with the English kindergarten educator.

Children’s limited L2 utterances in everyday situations could explain why their L2 performance level remained far below the level of age-matched English children. According to Swain (1985), the production of comprehensible output in the process of negotiation for meaning is crucial for L2 learners to achieve nativelike L2 proficiency. Just like adult learners, primary-school L2 learners seem to benefit from negotiation for meaning, as it provides
opportunities to receive comprehensible input, to produce comprehensible output, and to modify L2 output through feedback (Oliver, 1998). The special form of early partial immersion in German kindergartens offers few of these opportunities. Moreover, whereas form-focused instruction is sometimes provided in immersion in school contexts (Genesee, 2004), there was no L2 grammar instruction in our study because of learners’ young age. This is important because instruction with an explicit focus on form, in which learners are made aware of specific language features, is more effective when compared to uninstructed language learning (e.g., Long & Crookes, 1992).

**Language Contact and L2 Development in Immersion**

Contact quantity and quality were significantly associated with L2 outcomes in the immersion group after 20 months of L2 learning, with better English comprehension and production linked to higher contact quantity and quality, which supported Hypothesis 3. This result is in line with prior studies on L2 learning (Moyer, 2011; Vermeer, 2001) and confirms that, beyond the mere provision of L2 input, an adequate intensity of interaction with the learner is important for efficient L2 development. Thus, for partial immersion in German kindergartens, following the “one person—one language” principle, the present study shows—to the best of our knowledge for the first time—the importance of both contact quality and quantity for L2 learning at the individual level within one kindergarten group, complementing similar results from the ELIAS study using group comparisons (Weitz et al., 2010). However, the precise nature of the relationships between the contact measures and L2 outcomes remains unclear. Contact quantity and quality may foster L2 learning, but an effect in the opposite direction or a bidirectional relationship is also possible, because better L2 proficiency enables the children to communicate in the L2 more intensively. In addition, both contact quantity and quality were found to be associated with age ($r = .53$ and $r = .41$, respectively). Concerning quantity, the correlation is due to the fact that older children were, on average, enrolled in the immersion program for longer time periods. Concerning contact quality, older children tended to interact more intensively with the English educator. Thus, the impact of age on the associations between contact measures and L2 proficiency remains to be explored. This is especially important because a relationship between input quantity and L2 performance was also present for the conventional instruction group. The magnitude of this association was lower for the standardized BPVS III but comparable for the English Act-Out after 2.5 years of English learning compared to those of the immersion group.
Limitations and Future Research
As in many studies on L2 learning (see Norris & Ortega, 2000; Plonsky, 2013), the sample size in the current study was relatively small. However, it was sufficient to reveal large differences in L2 receptive skills between the immersion and conventional instruction groups and to reveal the crucial role of contact quality for L2 acquisition in the immersion kindergarten group. In addition, the results of longitudinal analyses were supported by cross-sectional analyses, which showed comparable findings. A large-scale longitudinal study with an adequate number of participants in different age groups (e.g., 2–3-, 4–5-, and 6-year-old children) could reveal different L2 development patterns in immersion kindergarten settings as a function of age. It is possible that older children benefit more from immersion than younger children due to their further developed short-term memory, cognitive abilities, and L1 skills (see Genesee, 2004). On the other hand, younger children may outperform older ones in the long run because they have the chance to spend more time in the immersion setting.

Furthermore, we assessed only quantitative data for a limited set of language aspects (L2 sentence comprehension, L2 receptive and productive vocabulary), which may not have revealed the full picture of the children’s language skills. Future research on L2 instruction methods should include other targets, such as grammar use, pronunciation, and qualitative measures of learners’ spontaneous L2 production. In-depth analyses of qualitative data may also reveal some interesting aspects of children’s learning in immersion conditions. For this, it might be necessary to build a corpus of child–educator interactions over an extended period of immersion in order to analyze children’s L2 input (e.g., in terms of vocabulary size and grammatical structures, frequency and quality of child-directed and child’s L2 speech, and type of corrective feedback), as well as children’s L2 output and the relationship between the two. Such corpus-based analyses would enable comparisons of findings across immersion and monolingual English or German child corpora (e.g., through CHILDES; MacWhinney, 2000).

Conclusion
The present longitudinal study showed that the special form of partial immersion in German kindergartens is an effective method to promote L2 receptive skills. This finding suggested that immersion should be preferred over conventional English courses, although both methods were found to promote L2 receptive skills. Importantly, there was no detrimental effect of immersion on L1 development. In addition, the study revealed an important role of contact quantity and quality for L2 development in kindergarten immersion programs.
However, L2 achievement remained far below nativelike levels, with L2 production remaining rudimentary; thus, neither of the L2 learning methods fostered L2 production on a large scale. However, it is reasonable to assume that early L2 contact could enhance later L2 learning at school, not only for receptive skills but (in the long run) for production as well.

Final revised version accepted 23 May 2015

Notes
1 The present study focused on immersion provided to majority-language children; L1 and L2 development can often differ in minority- and majority-language children because of very different living and learning conditions (Genesee, 2004).
2 The ELIAS project consists of several substudies published as different chapters in a book edited by Kersten et al. (2010).

References


E. Eilers (Eds.), Language and literacy in bilingual children (pp. 175–206). Clevedon, UK: Multilingual Matters.


