Research Article

Simone E. Pfenninger*

Emergent bilinguals in a digital world: a dynamic analysis of long-term L2 development in (pre)primary school children

https://doi.org/10.1515/iral-2021-0025 Received February 3, 2021; accepted February 3, 2021; published online March 3, 2021

Abstract: In this study, I present dense, longitudinal data exploring the insights that a Complex Dynamic Systems Theory (CDST) perspective can bring to bear on patterns of relationships found between learner individual differences - notably age of onset (AO) and extracurricular L2 English use - in children in (pre)primary programs in Switzerland. We studied 71 children who had received 50/50 bilingual instruction in German and English (so-called "partial CLIL" programs) as well as 105 children in "minimal CLIL" programs with almost uniquely monolingual German instruction (90% German, 10% English). In the data analysis, (1) generalized additive mixed modeling (GAMM) was combined with (2) mixed-effects regression modeling. The findings show that AO may exert an effect on L2 performance in bilingual but not traditional instructional settings. Furthermore, contact with English outside school is a strong predictor for learner outcome, regardless of the intensity of instruction and an early or late start respectively. We conclude that the traditional view of the age factor in instructional settings needs to give way to a new understanding of L2 development in intensive exposure conditions, in which age of acquisition is seen as a major determinant.

Keywords: age; CLIL; complex dynamic systems theory (CDST); digital technologies; individual differences

^{*}Corresponding author: Simone Pfenninger, Department of English and American Studies, University of Salzburg, Erzabt-Klotz-Straße 1, 5020 Salzburg, Austria, E-mail: Austria.simone.pfenninger@sbg.ac.at. https://orcid.org/0000-0002-0433-4812

Open Access. © 2021 Simone E. Pfenninger, published by De Gruyter. © BY This work is licensed under the Creative Commons Attribution 4.0 International License.

1 Introduction

SLA research has identified three recent trends that are particularly pervasive in language education of young learners. First, very young second language (L2) learners have recently become a new population of learners, which opens challenging questions regarding not just expectable outcomes but also the very nature and aims of L2 learning at this age (Muñoz 2019). Second, as a "major educational initiative" (Heras and Lasagabaster 2015: 72) in Europe, content and language integrated learning (CLIL) has been taking root in recent decades, with the promise that early bilingual instruction will result in higher levels of L2 proficiency, while simultaneously equipping students with other key skills such as intercultural awareness (e.g., Wode 2004). Third, today's multilingual and technologysupported culture is redefining when, why, and how languages – in particular English as a foreign language (EFL) – are learned and used (Douglas Fir Group 2016; Larsen-Freeman 2017). Never before has L2 learners' childhood been so intertwined with the use of technology, which presents learners with unprecedented opportunities for exposure to and use of English as a target language regardless of their physical location (e.g., Motteram 2013).

Each of these three topics has received a lot of attention in its own right, albeit not in interaction with the others. For instance, the nature and time course of L2 development in children with varying ages of first bilingual language exposure has been scarcely investigated (but see Kovelman et al. 2008). Similarly, Muñoz (2019) argues that an age-related research agenda should include more longitudinal studies that cast light on the L2 learning trajectories of early and late learners in all conditions – including immersive programs, which constitute a hybrid form between naturalistic L2 acquisition (where a general "earlier = better" trend emerges) and instructional L2 contexts (where often no age effects favoring early starters are manifest). Also, most current longitudinal studies on CLIL and age have aimed at investigating age effects in pseudo-longitudinal/cross-sectional studies with standard pretest-posttest-delayed posttest designs. However, as Peng et al. (2020) rightly point out, it is important to take an ecological approach that simultaneously examines individual learners and their interdependence with spatial-temporal context so as to cope with the increasing complexity and diversity of language use in the 21st century.

The current study takes cross-sectional and longitudinal mixed-methods approaches to CLIL and age in a somewhat different direction, both in terms of the aims and type of modeling. The main goal is to present a combination of snapshot data collected at one point in time with dense quantitative and qualitative data, exploring the impact of a range of individual differences (ID) – including age of

onset (AO) – and extracurricular L2 use on the L2 English development of children attending bilingual and regular (pre)primary L2 programs respectively in Switzerland. According to Hiver and Al-Hoorie (2020) collecting time-dense data can genuinely be one of the ways of studying the temporal and phenomenological aspects of human functioning and behavior that is most compatible with Complex Dynamic Systems Theory (CDST) – arguably "the most widely used and powerful explanatory framework in science" (van Gelder 1998: 622).

In the quantitative analysis, we use generalized additive mixed modeling (GAMM) to look for generalizations that apply to a larger number of individuals. As a second step, GAMM is used to identify the data points where a significant change in L2 development occurs, indicating L2 growth. Qualitative data, then, give a richer insight of the feelings, emotions, cognitive processes etc that the L2 tasks in the GAMM analysis miss. Finally, based on qualitative data, linear mixed-effects regression modeling is used to confirm the perceived effects retrospectively. Within this framework, we can describe changes in L2 development over a longer period of time and the learner's own rationale for the changes.

Results presented here should be considered in conjunction with the analyses of the same learners and dataset presented in Pfenninger (2020a, 2020b), in which the goal was to elucidate what *causes* significant L2 growth, and how L2 writing and oral language development are mediated by a complex, dynamic constellation of individual and social factors.

2 Literature review

2.1 Age-within-CLIL

The current view offered on the age factor in SLA is that "success" in additional languages is a function of the quantity and quality of language experience rather than simply a matter of maturational or general age effects – irrespective of the age of the learner (e.g., Singleton and Pfenninger 2018). For instance, it seems more and more likely that the reason why younger L2 starters in a naturalistic environment tend to be more proficient in the L2 in the long run than older starters is attributable to a range of socio-affective factors, e.g., to how they experience the L2, rather than to age specifically (e.g., Blom and Paradis 2016). In instructional contexts, too, a growing body of evidence from research in education, psycholinguistics, cognitive science and neurolinguistics challenges the conventional view of the age factor as the non-plus-ultra predictor of L2 learning outcome (e.g., Jaekel et al. 2017). Accordingly, interest of policy makers in many European countries has shifted from regular, low-input L2 programs to different types of intensive language programs. In general, a course or program is deemed intensive when the hours available for instruction are concentrated in blocks of time, giving students exposure to the L2 for several hours a day. The length of the intensive experience varies widely across countries and programs, however, as does the corresponding terminology (see Cenoz et al. 2014). Importantly, intensive programs may have the potential to break through the attainment ceiling typically associated with early L2 learning without necessitating changes in the time allotted to the different school subjects. This is the promise of CLIL, whose main idea is that proficiency will be developed in both the non-language subject and the language in which it is taught (Cenoz et al. 2014; Coyle et al. 2010). Such beneficial effects were found across different versions of intensive L2 instruction (e.g., Collins and White 2011), which is why CLIL has been described as "a major contribution to make to the [European] Union's language learning goals" (European Commission 2003: 8).

2.2 Online contexts are a major driving force in L2 acquisition

Outside the classroom, online contexts are a major driving force in today's globalized and technologized world, as learners have available a multiplicity of diversified and inexhaustible online resources - including L2 learning resources that could serve to provide authentic language input - through which to explore personal goals, learning interests and preferences, and which potentially expand upon prior knowledge, language abilities, and digital competencies (de Graaff 2015; Peng et al. 2020; Traxler and Kukulska-Hulme 2016). Recent studies (e.g., Henry and Thorsen 2019; Kuppens 2010; Motteram 2013; Sockett 2014) indicate that informal online language practices not only become more and more common – they are an effective way to learn, in part because of their influence on affective aspects of language learning. According to Richards (2015), the core features of learning beyond the classroom are agency, motivation and interaction. The learner has the capacity to act and engage with the material or other learners in a collaborative interaction, thereby benefitting from feedback and clarification requests. When used in the L2, language learning is linked to these activities with authentic input; the learner can build on everyday experiences.

3 This study

This is a two-stage study, consisting of a cross-sectional component with traditional frequentist statistics and a longitudinal "idiodynamic" component with CDST

methods and techniques, in order to show different aspects of L2 development throughout (pre)primary school as well as L2 attainment at the end of primary.

To relate the method explicitly to CDST, we need to take several factors into account: first, that the individual experience is structurally dynamic, undergoing relentless (non-linear) change; second, that the individual experience is complex and idiosyncratic. Third, as a dynamic system is never isolated, other systems that might interact with the focal system need to be identified, "with a particular need to be alert to the ways in which the focal system might adapt as a response to the interaction" (Dörnyei et al. 2015: 424). This also includes the dynamic interaction with the social context in which a particular ID variable – such as the age factor – is situated. A minor change in social interaction may lead to subsequent changes in motivation, achievement, and learning behavior (Lowie and Verspoor 2019).

3.1 Research questions

Two exploratory research questions guided the current study:

- (1) What insights can a CDST perspective through the use of GAMM bring to bear on patterns of relationships found between learner individual differences in children who are educated in bilingual schools?
- (2) Can the L2 English development and attainment in bilingual (pre)primary school be explained by certain learner individual differences (i.e., age of first CLIL onset and extracurricular L2 activities)?

The notion of bilingualism is used here for three main reasons. First, bilingual education (BE) "can also refer to 'immersion', in which a foreign language... is the medium of instruction" (Admiraal et al. 2006: 75). Second, CLIL is a form of bilingual education, since it promotes higher levels of oral and written language proficiency in an additional language than would be found in more traditional taught programs (Murphy and Evangelou 2016). Finally, many CLIL programs - including the one under investigation here – are expected to produce ideal balanced bilinguals (Cenoz et al. 2014).

It is also important to bear in mind that this study does not focus on betweengroup comparisons across CLIL programs (comparing, e.g., partial CLIL with minimal CLIL) but on within-group and between-learner analyses of age effects. The different implementations of the schooling system in Switzerland, the relatively small number of schools involved, the different number of instructional hours, and bias attributable to selection/(self)-selection would compromise such comparability in CLIL studies (see also Aguilar and Muñoz 2014; Bruton 2011 on these issues).

3.2 Participants

While many CDST-related studies consist of single case designs (e.g., Spoelman and Verspoor 2010) or involve a small number of learners only (e.g., Chan et al. 2015), dynamic and usage-based proponents do value group studies, e.g., for identifying general patterns or behaviors that hold for (a majority of) language learners (Hiver and Al-Hoorie 2020). I agree with Bulté and Housen (2020) in that one important aim of (quantitative) L2 research should be to somehow arrive at conclusions that extend beyond one single, specific (and preferably randomly selected) learner. What is more, looking for developmental patterns across learners does not necessarily involve averaging scores across learners, as will be shown in this paper.

In the cross-sectional part of this study, 176 students (L1 Swiss German) who varied in their age of first CLIL instruction onset (5, 7, or 9) were recruited at the end of primary education (age 12). Seventy-one of them were in partial CLIL (PAC) classes (see description below), while 105 of them came from six minimal CLIL (MIC) classes in two different cantons in Switzerland (Zurich and Basel), where students had started learning English at different ages. Fifty-four of them were early starters (AO 8; henceforth earlyMIC), while 51 were late starters (AO 11; henceforth lateMIC). The MIC participants' mean age at testing was 12;6 (range 11–14), and students received 2 h of English instruction per week.

Three groups of learners in the PAC program formed the focal group (all from the same community in Zurich), i.e., the longitudinal part of this study: a group of 25 Swiss learners from monolingual German-speaking homes with age of first CLIL exposure 5 (earlyPAC; length of German/English PAC until the end of primary: 8 years), a group of 24 Swiss learners of English with starting age 7 (midPAC; length of PAC: 6 years), and a group of 22 Swiss learners with starting age 9 (latePAC). Table 1 displays information about these subjects:

Table 1: Focal subje	ects participating in	n the longi	itudinal part.
----------------------	-----------------------	-------------	----------------

Group	Number of subjects (sex)	•	L2	Age of first CLIL exposure	No. of trajectories	Total no. of measurements (4 times/year)
1 earlyPAC	25 (15 F)	German	English	5	32	800
2 midPAC	24 (12 F)	German	English	7	24	576
3 latePAC	22 (14 F)	German	English	9	16	352

Children in the PAC program were drawn from a private (pre)primary bilingual school in Switzerland. These schools reserve one-half of school time for teaching/learning only in the L2. Learners received exposure to English via subject-matter instruction and communicative activities that did not focus on the grade-level curriculum. In addition, they received traditional English/ German-as-a-second language instruction: 6 h each in Grade 1, 5 h each in Grades 2 and 3, and 4 h each in Grades 4-6. Finally, all the children were matched for SES (and similar home literacy environments), considering that home literacy is a significant factor in early literacy development (Kovelman et al. 2008).

3.3 Tasks and procedure

The 176 students in the cross-sectional design were asked to write a timed English narrative (topic: the plot of their favorite movie, book or TV series) and complete a re-telling task requiring them to narrate the plot of a silent video they had previously watched at the end of primary school – see Pfenninger (under review) for a more detailed description of these tasks. The 71 students in the focal group of the longitudinal design each wrote one such English narratives and did one oral re-telling task per term (i.e., four times a year, see Table 1).

In order to identify which contextual and socio-affective elements may be relevant to significant L2 growth, it is necessary to explore the participants' own perspectives through various forms of introspection (Ushioda 2015). Thus, participants were asked to complete a language awareness questionnaire with open-ended questions. In addition, semi-structured interviews lasting 10 min were carried out individually with the students. During the interview and in the questionnaire, information about their use of English in everyday life (extracurricular activities), knowledge of languages, emotions (motivation, anxiety, more and less enjoyable moments, etc.), the role played by their parents, peers, the teacher and the language assistant, their progress in English, and their reflections on the narratives produced (strengths and weaknesses) was gathered. Verbatim transcripts were produced from the recordings of the interviews.

In the second step, i.e., the cross-sectional component of the study, five predictors in addition to AO were chosen, informed by the themes that emerged in the qualitative part of the longitudinal component (see Supporting Information S1 for reliability information). The questionnaire scale items were fine-tuned with a pilot study.

3.4 Coding

Coders with expertise in linguistics, who were also bilingual German-English speakers, coded transcripts of the children's speech, using the koRpus package in R (version 0.11–5). Table 2 shows how the measures were calculated.2

From the perspective of CDST, complexity, fluency, and accuracy (CAF) comprises the three subsystems of the language system (Yu and Lowie 2020), although CAF is of course not exclusive to CDST (e.g., Bulté and Housen 2014). In this study, oral language and writing development are operationally defined as language development – rather than control over one's own textual output – measured by indices of CAF and lexical richness as displayed in speech and writing. As such, speech and writing in this study are more of a "medium for eliciting insights about L2 acquisition" (Norris and Manchón 2012: 224) than a developmental target in its own right. What is more, it is clear that CAF cannot be observed on the basis of a small number of measures; ideally, they should be realized as a constellation of multiple features each, considering that, e.g., different complexity dimensions do not necessarily develop in parallel, and that the relationship between different dimensions of complexity can be both supportive and competitive.

3.5 Statistical analysis

In the longitudinal analyses, generalized additive mixed modeling (GAMM) was performed using the mgcv R package (Wood 2006), and results were plotted using the packages ggplot2 and itsadug (van Rij et al. 2015). We fitted separate smooths to

Table 2:	Measures	and their	calculations.	
----------	----------	-----------	---------------	--

Morphosyntactic complexity and fluency	Mean length of utterance (MLU)	number of morphemes per word
	Fluency (word count)	Written text length in tokens for written data; Pruned syllables per minute for oral data
	Clause ratio (Bulté and Housen 2014; Polat and Kim 2014)	Clauses/T-unit for writing; Clauses/analysis of speech units (AS-unit) for oral language
Lexical richness	Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis 2010)	Total number of words in the text is divided by the total factor count
Accuracy	Error-free units (Polio and Shea 2014)	Total number of error-free T-units/AS-units

the trajectories of the subjects and used model comparison and difference smooths to see whether the three AO groups were different. GAMM have several advantages for a study that aims to relate the method explicitly to CDST:

- GAMM describes the iterative nature of the processes involved, which is central to the notion of development: the next "state" of development is a function of the preceding state and a condition for the next state.
- GAMM accounts for interdependency in learner's internal subsystems (e.g., perceptual-motor, cognitive, and psychological systems of the learner) and external subsystems (such as other language users within the speech community), i.e., connected components cannot be treated as independent variables or components.
- GAMM takes account of autocorrelation (nested dependencies, multivariate data, repeated measures). Autocorrelation happens, e.g., when the data are collected over time, and the data can no longer be treated as random. Just as with the CDST principles of interdependence and complex causality, each data point becomes very similar to the one just before it and the one just after it.
- GAMM can model complex nonlinear trajectories: in both L1 and L2 development, some subsystems may take off slowly at first, then all of a sudden jump, and level off at the end.
- In GAMM, the time series data is the target of inferential statistics (data is not automatically aggregated as in ANOVA-type analyses); thus, the learner becomes "a representative of himself or herself, rather than a representative of the larger class group" (Lasagabaster 2017: 109).
- GAMM is able to deal with missing data and creates strategies to summarize meaningful events within the data stream.
- When we look at the individual raw data, it is sometimes difficult to find out whether there is any general improvement or change in the data. We can use smoothing techniques, just as in more traditional approaches to statistics, to see whether there is a general trend or not. The purpose of a smoother is to "sketch" the general trend of the data, and leave out many of the irregularities of the actual data. Smoothers are therefore well suited to representing a direction because they give an impression of the general pattern of development.

From a cross-sectional perspective, we had to account for multiple layers of influence in different L2 learning contexts. Mixed-effects modeling is an effective method to examine such nested systems and to model and partition the variance attributable to each of these levels (Hiver and Al-Hoorie 2020). An important feature of mixed modeling is its ability to (1) explicitly take context into account (one of the main lessons of CDST), and (2) model not only the mean of the attribute in question, but also its variability.

The mixed-effects models in this study included hierarchical random effects for classes (n=12) and groups (earlyPAC, midPAC, latePAC, midMIC, lateMIC), and crossed random effects for subjects and items respectively, using the lme4 package (version 1.1–21) in R (Version 3.6.0; R Development Core Team). While continuous fixed effects and dependent variables were centered, categorical fixed effects were recoded to use contrast coding. Models were fitted using a maximum likelihood technique. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. All models reported were fitted using Laplace estimation with the R software.

4 Results

In the first step, the goal was to gain insight into the actual L2 developmental process of the individual participants in the longitudinal dataset (RQ1). Figure 1 shows the oral and written L2 development of the three AO groups in the PAC program. At first glance, these figures reveal clear differences between the three groups in terms of height and shape of the trajectories. However, what they all have in common is stronger slopes – i.e., faster learning rates – in the last three years of primary school compared to earlier stages of the L2 development.

In order to compare AO effects across multiple measures – which running a separate GAMM for each L2 measure does not allow – the *z*-transformed scores of the five oral and five written measures were combined in two models (one for written data and oral data respectively) with the structure presented in Supporting Information S2. A model comparison suggested that the inclusion of the difference smooths improved the model fit significantly (see Tables A and B in the Supporting Information for the model output). Accuracy in the earlyPAC was taken as the reference level.

Since GAM(M) does not have straightforward interpretable coefficients (Winter and Wieling 2016), visualization of model fits is essential. For instance, visual methods for significance testing can show where and in what way the trajectories differ by plotting the difference smooth itself along with a confidence interval at different points (using the plot_diff function from the itsadug package). Figures 2–6 illustrate the levels of "group" that the difference smooth is based on using corresponding pointwise confidence intervals (minus random effects) for the oral measures (for further data visualization, see Pfenninger 2020b). When the shaded confidence band does not overlap with the *x*-axis (i.e., the value is significantly different from zero), this is indicated by a red line on the *x*-axis (and vertical dotted lines)

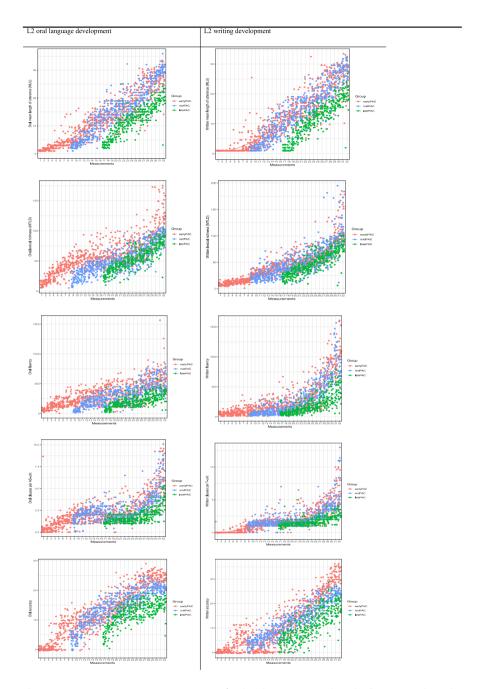


Figure 1: AO groups across 10 L2 measures (oral and written mean length of utterance, lexical richness, fluency, complexity and accuracy).

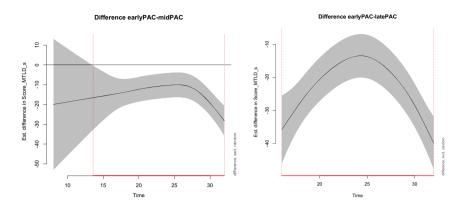


Figure 2: Difference smooths for oral MTLD.

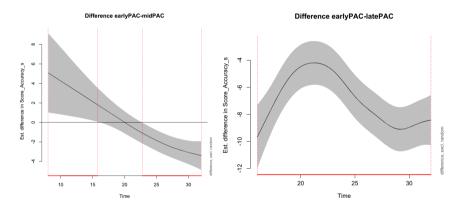


Figure 3: Difference smooths for oral accuracy.

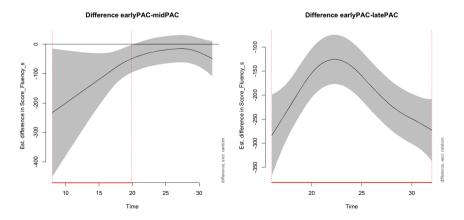


Figure 4: Difference smooths for oral fluency.

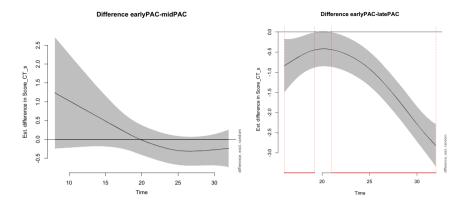


Figure 5: Difference smooths for oral complexity.

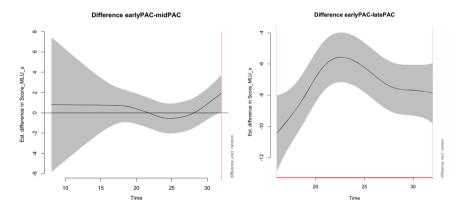


Figure 6: Difference smooths for oral MLU.

According to Figures 2–6, latePAC showed significantly different trajectories compared to earlyPAC across all oral and written measures. The L2 development of the earlyPAC and midPAC overlapped to some degree, i.e., they often differed at the beginning of their L2 development when the midPAC lagged behind, as well as at the end when the midPAC not only caught up with, but outperformed, the earlyPAC. For oral fluency and oral complexity there were no differences between these two AO groups.

The subject-specific random effects were significant in both models across all measures (see Tables A and B in the Supporting Information), which suggests that the trajectories for the subjects were indeed different (inter-individual variation). Furthermore, visualization of the data illustrates how the individual curves showed a crisscross change in the period under investigation (intra-individual

variation). This indicates that the changes in L2 speech and writing skills differed between children and that each child developed more or less at their own pace and produced their unique L2 trajectory, as Figures 7-9 illustrate for spoken complexity.

Measurements before or after a certain focal point (e.g., scores at the end of primary school) give a different picture (see also Pfenninger 2020b). Thus, the findings of a snapshot analysis may not be representative of a longer period of time and cross-sectional results, which follow below, must be treated with caution.

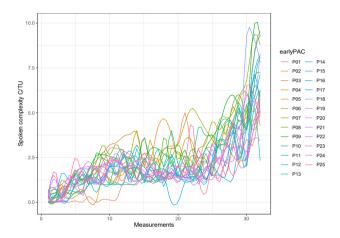


Figure 7: Individual growth curves for the development of the earlyPAC for spoken complexity.

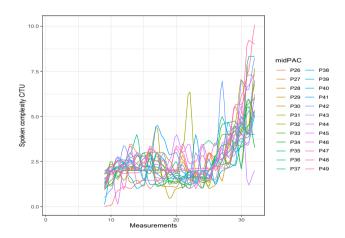


Figure 8: Individual growth curves for the development of the midPAC for spoken complexity.

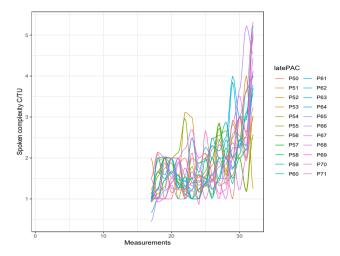


Figure 9: Individual growth curves for the development of the latePAC for spoken complexity.

Next, we consulted the results of the qualitative analyses in the same dataset (Pfenninger under review) to determine when and why L2 development is statistically significantly increasing (or decreasing) as indicated by GAMM; here, we focused on participants' self-reports during periods of significant L2 growth. Examples of the interconnected systems influencing L2 growth included the following: socio-affective states; students' encounters with English outside of school; the "people factor" (Lasagabaster 2017); cognitive events; and strategies. In particular, it became clear that starting in fourth grade, at the age of 10, children spent large amounts of time in English-language - predominantly online – environments outside the classroom. Besides affective states the use of digital technologies came out on top of all the factors hypothesized to interact with students' motivational flows. In general, the following themes emerged: amount of time spent in a (real/online) situation involving native speaker contact; surfing the net/checking pages in English; video games; social media; (learning) apps; movies, series, YouTube; and songs.

In a last step, i.e., the cross-sectional analysis of this study, we then used the regression results from the mixed-effects models to test the learners' perceptions, e.g., how the use of digital technologies modifies motivated learning behavior and influenced L2 outcomes at the end of different types of EFL schooling at primary level in Switzerland (RQ2). Tables D and E in the online Supporting Information show full results of the two mixed models that were specified including the z-transformed scores of the five oral and five written measures respectively (see also Table C in the Supporting Information for the descriptive statistics and reliability information).

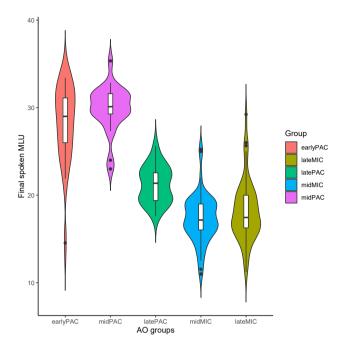


Figure 10: Spoken MLU for the five CLIL groups.

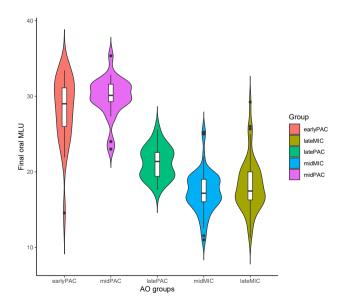


Figure 11: Written MLU for the five CLIL groups.

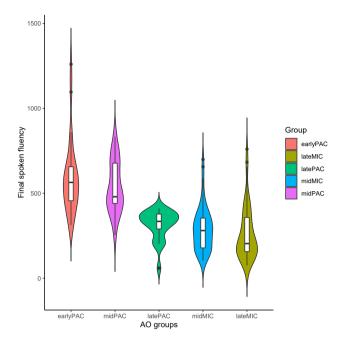


Figure 12: Spoken fluency for the five CLIL groups.

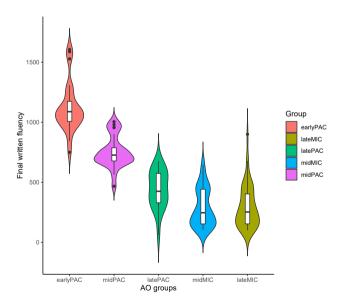


Figure 13: Written fluency for the five CLIL groups.

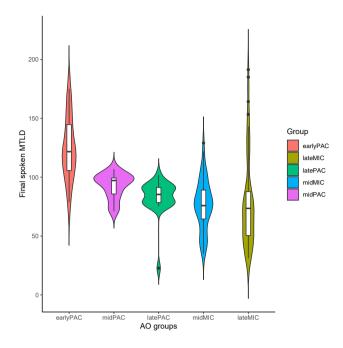


Figure 14: Spoken lexical richness for the five CLIL groups.

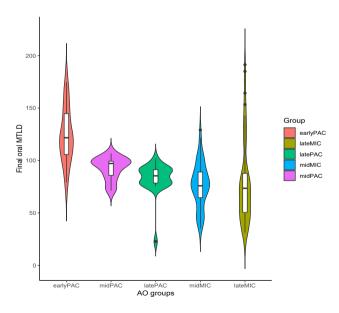


Figure 15: Written lexical richness for the five CLIL groups.

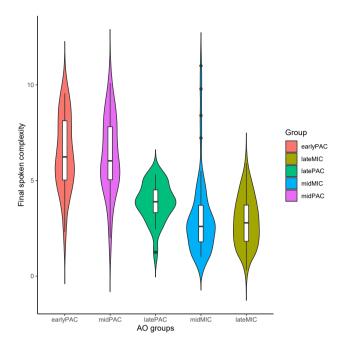


Figure 16: Spoken complexity for the five CLIL groups.

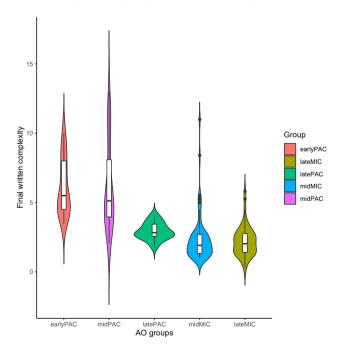


Figure 17: Written complexity for the five CLIL groups.

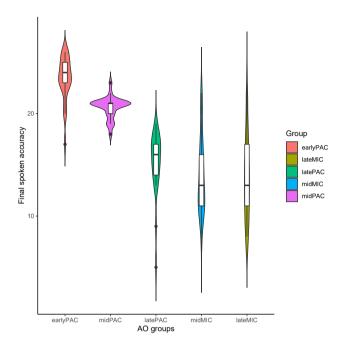


Figure 18: Spoken accuracy for the five CLIL groups.

In a nutshell, out of the five phenomena pointed out by the participants only one reached significance (the people factor in the oral data), but there were significant interactions in the oral model between (a) motivation and extracurricular L2 use, and (b) motivation and the people factor. In the written data, we found significant interactions between (c) motivation and cognitive events and (d) motivation and strategies. As Figures 10–19 show, an earlier AO was a significant predictor of all of the tested L2 skills in the PAC program except for spoken fluency, whereas there were no age-related differences whatsoever in the MIC program.

5 Discussion

The purpose of the current study was to evaluate the developmental patterns of the L2 oral and written production of Swiss English learners over a period of four to eight years. Both product- and process-oriented methods were adopted. The findings further the results from other ongoing research (Pfenninger 2020b) that age of first CLIL instruction onset may exert an effect on performance in an instructional setting, as there were significant differences between children with

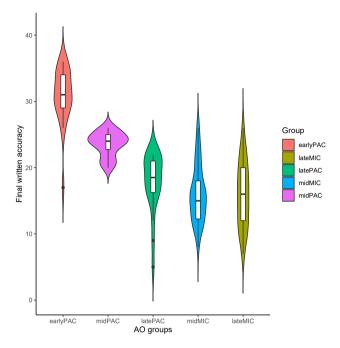


Figure 19: Written accuracy for the five CLIL groups.

AO 5 and children with AO 9 in terms of both height and shape of the L2 trajectories. By contrast, AO 5 and AO 7 overlapped to a great extent. However, the difference between AO groups in L2 development scores seemed to decrease over time, which tempts us to speculate that the children in the latePAC group might also catch up with the other two AO groups at a later point in their school career. The study also showed that GAMM – and in particular its visualizations – has advantages over traditional snapshot analyses of being able to reveal when different age groups catch up with each other. There is, consequently, a pressing need for widening the scope of age-related investigation into bilingual school contexts and to bring in the appropriate methodological tools for explaining when and why later starters catch up with earlier starters.

The results of the qualitative analysis of the longitudinal data and the quantitative analysis of the cross-sectional data corroborated previous findings that extracurricular or curricular engagement with particular types of technology such as digital gaming and watching movies is an important source for learning (Henry and Thorsen 2019; Reinders 2017) – which arguably explains the fast learning rates in the last three years of primary revealed by the GAMM. De Graaff (2015), for

instance, found that contact with English outside school is a strong predictor for learner outcome, regardless of an early or late start – a picture which also emerged in this study. Starting from the age of 10, students' network of social ties extended beyond school boundaries; the effect of these extracurricular activities on L2 outcomes was confirmed by the mixed models run at the end of primary school.

The mixed-methods approach used in this study turned out to be particularly useful for our purposes because (1) it does justice to the complexity of the phenomenon under investigation, and (2) it allows for surprising *inductive* discoveries of possible effects. More often than not certain phenomena can be interpreted only retrospectively as an effect, rather than being firmly stated as a prediction that follows an unbroken linear causality into the future (Byrnes 2017). Instead of relying exclusively on quantitative measures to assess the variables in the study, in-depth interviews were also applied. These interviews give a richer insight of the feelings, emotions, cognitive processes etc. that the language tasks miss – and they informed the mixed models in the cross-sectional part of the study. According to Dörnyei et al. (2015), mixed methods research lends itself well to CDST studies, "especially if it allows unanticipated factors into the mix" (425) (see also Hiver and Al-Hoorie 2020).

Finally, the study showed that there are ways of reconciling an idiodynamic approach with generalizability. While statistical robustness is not a goal from a CDST perspective, arguably there is value in identifying existing regularities in L2 phenomena (Ellis 2007). Longitudinal studies with dense data such as this one are required to test such expectations, as mere snapshots of states are insufficient. In consequence, statistical methods capable of tracking change and accounting for variability and auto-correlation in non-linear patterns are needed. In this study, I advocated the use of generalized (mixed-effects) regression framework, including generalized additive (mixed) models (GAM(M)s), which represent an important statistical development and provide a valuable set of tools for analyzing L2 data.

6 Conclusion

Content-wise, the current study is an answer to Muñoz's (2015) and Dalton-Puffer and Smit's (2013) call for research on the issue of an optimum initial proficiency level for CLIL at primary level and more longitudinal CLIL studies. The lesson to draw from our research findings is that that late starters (AO 9) in partial CLIL programs not only attain lower proficiency levels than earlier starters (AOs 5 and 7) by the end of primary school, they also show markedly different L2 trajectories. Furthermore, learners' intensive contact with English both inside and outside the

classroom positively benefits their L2 development and enhances their engagement with the L2.

Methodologically speaking, we believe that the design of this study is noteworthy among the growing body of CDST-inspired studies of L2 development because of (1) its combination of cross-sectional analysis and longitudinal design with fairly dense data collection points, (2) the integration of quantitative and qualitative analyses, and (3) its sample size, which is relatively large for a micro-development study. Blending research methods is a genuinely productive way to produce a more multidimensional understanding of an issue, and this underscores the value of methodological diversity for CDST research (Hiver and Al-Hoorie 2020).

There may be other possible explanations from other theoretical perspectives about these observed effects. Potential under-theorized factors in this study are, for instance, cognitive factors such as aptitude or intelligence. Future studies might limit the number of focal learners in order to zoom in on each individual and identify specific parameters of CDST, such as attractor states (system outcome states), phase shifts, emergence and co-adaptation. However, the integrated mixed methods approach employed in this study demonstrates that the application of an ecological and person-centered approach means not rejecting but rather complementing the L2 frameworks developed in recent decades "so as to optimally respond to the realities of our highly mobile, globalized, and digitalized world, in which millions of people endeavor to learn new languages, in different instructional settings and for different reasons" (Peng et al. 2020).

References

Admiraal, Wilfried, Gerard Westhoff & Kees de Bot. 2006. Evaluation of bilingual secondary education in the Netherlands: Students' language proficiency in English. Educational Research and Evaluation 12(1). 75-93.

Aguilar, Marta & Carmen Muñoz. 2014. The effect of proficiency on CLIL benefits in engineering students in Spain. International Journal of Applied Linguistics 24(1). 1-18.

Blom, Elma & Johanne Paradis. 2016. Introduction: Special issue on age effects in child language acquisition. Journal of Child Language 43(3). 473-478.

Bruton, Anthony, Miguel García López & Mesa Raquel Esquiliche. 2011. Incidental L2 vocabulary learning: An impracticable term? TESOL Quarterly 45(4). 759-768.

Bulté, Bram & Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. Journal of Second Language Writing 26. 42-65.

Bulté, Bram & Alex Housen. 2020. A DUB-inspired case study of multidimensional L2 complexity development: Competing or connecting growers? In Lowie Wander, Michel Marije, Rousse-Malpat Audrey, Keijzer Merel & Steinkrauss Rasmus (eds.), Usage-based dynamics

- in second language development: In celebration of Marjolijn Verspoor, 87-131. Bristol: Multilingual Matters.
- Byrnes, Heidi. 2017. Adopting a complexity theory perspective in language studies: From pedagogy to curriculum as a "simplex" system. In Plenary talk given at the 7th International Conference on Classroom-oriented Research, Konin (Poland), 9–11 October.
- Cenoz, Jasone, Fred Genesee & Dirk Gorter. 2014. Critical analysis of CLIL: Taking stock and looking forward. Applied Linguistics 35(3). 243-262.
- Chan, Huiping P., M. Verspoor & Louisa Vahtrick. 2015. Dynamic development in speaking versus writing in identical twins. Language Learning 65(2). 298-325.
- Coyle, Do, Philip Hood & David Marsh. 2010. CLIL: Content and language integrated learning. Cambridge: Cambridge University Press.
- Dalton-Puffer, Christiane & Ute Smit. 2013. Content and language integrated learning: A research agenda. Language Teaching 46. 545-559.
- de Graaff, Rick. 2015. Vroeg of laat Engels in het basisonderwijs; Wat levert het op? Levende Talen Tijdschrift 16(2). 3-15.
- Dörnyei, Zoltán, Peter D. MacIntyre & Alastair Henry (eds.). 2015. Motivational dynamics in language learning. Bristol: Multilingual Matters.
- Douglas Fir Group. 2016. A transdisciplinary framework for SLA in a multilingual world. The Modern Language Journal 100(s1). 19-47.
- Ellis, Nick C. 2007. Dynamic systems and SLA: the wood and the trees. Bilingualism: Language and Cognition 10. 23-25.
- Henry, Alastair & Cecilia Thorsen. 2019. Engagement with technology: Gaming, immersion and sub-optimal experiences. Technology in Language Teaching & Learning 1(2). 52-67.
- Heras, Arantxa & David Lasagabaster. 2015. The impact of CLIL on affective factors and vocabulary learning. Language Teaching and Research 19(1). 70-88.
- Hiver, Phil & Ali H. Al-Hoorie. 2020. Research methods for complexity theory in applied linguistics. Bristol: Multilingual Matters.
- Jaekel, Nils, Michael Schurig, Merle Florian & Markus Ritter. 2017. From early starters to late finishers? A longitudinal study of early foreign language learning in school. Language *Learning* 67(3). 631–664.
- Kovelman, Ioulia, Stephanie A. Baker & Laura-Ann Petitto. 2008. Age of first bilingual language exposure as a new window into bilingual reading development. Bilingualism: Language and Cognition 11. 203-223.
- Kuppens, An H. 2010. Incidental foreign language acquisition from media exposure. Learning, Media and Technology 35(1). 65–85.
- Larsen-Freeman, Diane. 2017. Complexity theory: The lessons continue. In Lourdes Ortega & ZhaoHong Han (eds.), Complexity theory and language development: In celebration of Diane Larsen-Freeman, 11-50. Amsterdam: John Benjamins.
- Lasagabaster, David. 2017. Pondering motivational ups and downs throughout a two-month period: A complex dynamic system perspective. Innovation in Language Learning and Teaching 11(2). 109-127.
- Lowie, Wander & Marjolijn Verspoor. 2019. Individual differences and the ergodicity problem. Language Learning 69(s1). 184–206.
- McCarthy, Philip M. & Jarvis. Scott. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods 42. 381-392.
- Motteram, Gary. 2013. Innovations in learning technologies for English language teaching. London: British Council.

- Muñoz, Carmen. 2015. Time and timing in CLIL: A comparative approach to language gains. In Maria Juan-Garau & Joana Salazar-Noguera (eds.), Content-based language learning in multilingual educational environments, 87-104. Berlin: Springer.
- Muñoz, Carmen. 2019. A new look at age: young and old L2 learners. In John W. Schwieter & Alessandro Benati (eds.), The Cambridge handbook of language learning, 430-450. Cambridge: Cambridge University Press.
- Murphy, Victoria A. & Maria Evangelou (eds.). 2016. Early childhood education in English for speakers of other languages. London: British Council.
- Norris, John M. & Rosa Manchón. 2012. Investigating L2 writing development from multiple perspectives: Issues in theory and research. In Rosa Manchón (ed.), L2 writing development: Multiple perspectives, 221-244. Berlin: de Gruyter.
- Peng, Hongying, Sake Jager, Steven Thorne & Wander Lowie. 2020. A holistic person-centred approach to Mobile-Assisted Language Learning. In Lowie Wander, Michel Marije, Rousse-Malpat Audrey, Keijzer Merel & Steinkrauss Rasmus (eds.), Usage-based dynamics in second language development: In celebration of Marjoliin Verspoor, 132–155. Bristol: Multilingual Matters.
- Pfenninger Simone E. 2020a. The dynamic multicausality of age of first bilingual language exposure: Evidence from a longitudinal CLIL study with dense time serial measurement. The Modern Language Journal 104(3). 662-686.
- Pfenninger Simone E. 2020b. About the INTER and the INTRA in age-related research: Evidence from a longitudinal CLIL study with dense time serial measurements. Linquistics Vanquard. https://doi.org/10.1515/lingvan-2020-0028.
- Polat, Brittany & Youjin Kim. 2014. Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. Applied Linguistics 35(2). 184-207.
- Polio, Charlene & Mark Shea. 2014. An investigation into current measures of linguistic accuracy in second language writing research. Journal of Second Language Writing 26. 10-27.
- Reinders, Hayo. 2017. Digital games and second language learning. In Steven L. Thorne & Stephen May (eds.), Language, education and technology, 3rd edn., 329-344. New York: Springer.
- Richards, Jack C. 2015. The changing face of language learning: Learning beyond the classroom. RELC Journal 46(1). 5-22.
- Singleton, David & Simone E. Pfenninger. 2018. L2 acquisition in childhood, adulthood and old age: Misreported and under-researched dimensions of the age factor. Journal of Second Language Studies 1(2). 254-275.
- Sockett, Geoffrey. 2014. The online informal learning of English. London: Palgrave.
- Spoelman, Marianne & Marjolijn Verspoor. 2010. Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. Applied Linquistics 31(4). 532-553.
- Traxler, John & Agnes Kukulska-Hulme. 2016. Mobile learning: The next generation. New York: Routledge.
- Ushioda, Ema. 2015. Context and complex dynamic systems theory. In Zoltán Dörnyei, Peter D. MacIntyre & Alastair Henry (eds.), Motivational dynamics in language learning, 9-42. Bristol: Multilingual Matters.
- van Gelder, Tim. 1998. The dynamical hypothesis in cognitive science. Behavioral and Brain Sciences 21. 615-628.
- van, Rij, Bart Hollebrandse Jacolien & Petra Hendriks. 2015. itsadug: Interpreting time series and autocorrelated data using GAMMs. R package version 1.0.1.

- Winter, Bodo & Martijn Wieling. 2016. How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling, Journal of Language Evolution 1(1). 7-18.
- Wode, Henning, 2004. Frühes Fremdsprachenlernen. Englisch ab Kita und Grundschule: Warum? Wie? Was bringt es? Kiel: Verein für frühe Mehrsprachigkeit an Kindertageseinrichtungen und Schulen FMKS e.V.
- Wood, Simon. 2006. Generalized additive models: An introduction with R. Boca Raton: CRC Press. Yu, Hanjing & Wander Lowie. 2020. Dynamic paths of complexity and accuracy in second language speech: A longitudinal case study of Chinese learners. Applied Linguistics 41(6). 855-877.

Supplementary Material: The online version of this article offers supplementary material (https:// doi.org/10.1515/iral-2021.0025).