

Language Teaching Research Quarterly

2024, Vol. 39, 145–173



Reconciling the Divides: A Dynamic Integrative Analysis of Variability and Commonality in (Pre)primary School English Development in Switzerland

Simone E. Pfenninger^{1*}, Mason A. Wirtz²

¹English Department, University of Zurich, Switzerland

²Department of German Language and Literatures, University of Salzburg, Austria

Received 13 May 2023

Accepted 26 September 2023

Abstract

The main goal of this paper is to suggest a combination of data analyses – notably generalized additive models, time-series clustering methodology, visual methods for significance testing and qualitative analyses – that relate to Complex Dynamic Systems Theory (CDST). To this end, we report on findings from a larger project conducted in a bilingual (pre)primary school in Switzerland, aiming to elucidate the complex ways that L2 English development emerges over time in 45 children who received German/English bilingual instruction over a period of eight years (age 5-12) in combination with emerging extracurricular exposure to English. The results reveal that increased extracurricular activities in English are particularly noticeable during periods of rapid development, but the effects seem temporally limited to the end stages of testing and strongly hinge on the cluster in question (i.e., learners with trajectorial similarities). We relate the findings to the “authenticity gap” between English inside and outside of school, as clusters who perceive a discrepancy between in- and out-of-school encounters with English also show rapid development that is characterized by increased English exposure during extracurricular activities. Methodological implications of adopting non-linear models, which can model complex dynamic relationships in order to better reconcile generalizability, variability, and individuality, are discussed.

Keywords: *Dynamic Integrative Analysis, Variability, Cluster Analysis, Primary School, English Development, Switzerland*

Introduction

The dynamic turn in SLA (e.g. de Bot, 2008; Dörnyei, 2009; Jessner, 2008; Larsen-Freeman, 2006; Larsen-Freeman & Cameron, 2008; Mercer, 2011; Verspoor et al., 2008) has not only afforded new ways of “seeing” but also “doing” research (Benson, 2019; Larsen-Freeman, 2018; Serafini, 2020). For instance, in a complex world, “we lose predictability; the nature of

* Corresponding author.

E-mail address: simone.pfenninger@es.uzh.ch

<https://doi.org/10.32038/ltrq.2024.39.11>

explanation changes; cause and effect work differently” (Larsen-Freeman & Cameron, 2008, p. 72). As a consequence, the conceptions of the causal structures underlying outcomes from the perspective of Complex Dynamic Systems Theory (CDST) are at odds with the assumptions required for standard regression techniques and conventional comparative methods to provide valid causal inferences.

In this paper, we suggest a combination of data analyses – notably generalized additive models, time-series clustering methodology, visual methods for significance testing and qualitative analyses – afforded by dynamic systems frameworks in order to elucidate the complex ways that L2 English development emerges over time in bilingual pre/primary schools through dynamic, reciprocal interactions with context, i.e. extracurricular use of English. We argue that CDST warrants change from a traditional focus on hypothesis testing and generalizability of results (see also Serafini, 2020). Even if the readers do not explicitly adopt a dynamic systems framework, the overall emerging picture is that of a broad shift that has departed in several respects from traditionally established perspectives of causality in applied linguistics (Waninge et al., 2014). Process-oriented approaches, such as the longitudinal micro-development approach adopted in this paper, fundamentally question the feasibility of investigating cause-effect relations, the traditional basis of generalizable theories: since it is highly unlikely that a single cause (or a handful of independent variables) will give rise (in a linear fashion) to such a complex event as L2 learning, some complexivists caution that “differences between individuals cannot and need not be generalized beyond the individual learners we are observing” (Lowie & Verspoor, 2018, p. 20). Others (e.g. de Bot & Larsen-Freeman, 2011, p. 23) suggest focusing on “tendencies, patterns and contingencies” rather than simple cause-effect explanations. Yet others (e.g. Hiver & Al-Hoorie, 2020) make specific suggestions how generalizable statements can be made at the level of system dynamics.

In order to demonstrate how variability and generalizability can be reconciled, we draw on parts of the dataset used in Pfenninger (2020, 2021a, 2021b), where dense longitudinal written and spoken data as well as information on extracurricular use of English was collected from 45 pre/primary school children in Switzerland who received German/English bilingual instruction over a period of eight years (age 5-12). The overall goal is to operationalize the main tenets of the dynamic approach in actual applied linguistics (AL) research; specifically, we attempt to outline ways of identifying L2 developmental patterns that transcend individual L2 developmental trajectories and subsequently explain these clusters quantitatively and qualitatively in terms of individual learner differences in extracurricular L2 use.

Literature Review

Linearity vs. Non-linearity of L2 Processes

Fundamental to the conceptual transformation described above is that the centrality of variation and nonlinearity in understanding the system is key (Larsen-Freeman 2015). For one, it is well-known that the process of L2 development may take a variety of forms and can even be characterized as specific to each individual (Peng et al. 2022). This is particularly important in today’s multilingual and technology-supported culture, which 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), which may lead to divergent learning experiences and developmental processes (Peng et al. 2022). What is more, the individual-level

L2 change over time is a nonlinear *process* rather than *product*. It is characterized by phases of stability, called ‘attractor states’, which alternate with periods of high variability that accompany phase shifts indicative of rapid developmental change, often referred to as ‘repeller states’ (de Bot et al., 2007; Verspoor et al., 2011). Identifying phase shift, however, has proven difficult in the AL literature. For instance, strong peaks or dips in variability have been tested for significance through resampling techniques and Monte Carlo analyses (Penris & Verspoor, 2017; van Dijk et al., 2011; Yu & Lowie, 2020), using Poptools (Hood, 2004). By randomly reshuffling the data 5000 times, a Monte Carlo analysis calculates how often a similar peak occurs in the data set when shuffled. If a peak occurred less than 250 times, it was deemed significant (α at .05). In order to neutralize the variability to some degree, they were supplanted with LOESS curves (Penris & Verspoor, 2017). This is tricky as it has been suggested that “[i]n a complex systems approach, there is no need to smooth away specific variability or to decide some detail is superfluous” (Larsen–Freeman & Cameron, 2008, p. 75).

Individual versus Group-based Designs

The non-linearity of L2 processes and the interconnectedness of variables can only be investigated by including the time dimension, “which at the same time blurs the neat two-dimensional results of *group studies*” (Lowie et al., 2015, p. 79, our emphasis). Or, in Larsen-Freeman & Cameron’s (2008, p. 238) words, “variability in data is not noise to be discarded when averaging across events or individuals, or the result of measurement error” (see also Ellis & Larsen-Freeman, 2006). This stands in stark contrast to the urge in AL to group data/participants into aggregate, social, non-individual entities according to various individual learner differences i.e. to spuriously suggest the presence of cut-offs or threshold effects: English as a native language vs. English as a second language vs. English as a foreign language; monolinguals vs. bilinguals vs. multilinguals; sequential vs. simultaneous bilinguals; 5-7-year-olds vs. 8-9-yr-olds vs. 10-12-yr-olds; etc. In MacIntyre and Mercer (2014, p. 166), for instance, we can read that “[r]esearch that uses group averages and correlations among variables has been a mainstay of SLA theory development over the years”. Furthermore, group observations are analyzed “using statistical procedures that are widely accepted as proof for effects and relationships, and that results in generalizations of sample group scores to population” (Lowie, 2017, p. 125).

Grouping participants comes with several issues. As with many other participant features, proficiency level or nativelikeness is not inherently categorical (see e.g. Birdsong, 2018; Vanhove, 2020). That is, individuals are not simply “native” or “non-native” speakers, or low- or high-proficiency learners, but they are native-like or proficient in English *to varying degrees*. Accordingly, so as to more faithfully capture these constructs, it is advantageous to operationalize and analyze them as a continuous (rather than a categorical) subject factor. As with any other continuous variable, (arbitrarily) assigning participants to categories (e.g. low-, advanced-, and high-proficiency learners) may mask intra-group variability and result in loss of statistical power (see, e.g., Altman, 1998).

Furthermore, studying the aggregate of a group to which the individual belongs raises the important question whether the observed effect is realized in each one of the individuals to whom the treatment condition or personal attribute producing that effect applies (Lowie, 2017). Can aggregated and statistical techniques do any more than “hint at” possible causal patterns

on an aggregated measure? Is the average “grand sweep” (Larsen-Freeman, 2006, p. 594) representative for any of the individual of the population? Several AL researchers have addressed this question from various theoretical angles. Within the CDST framework, in particular, scholars have discussed the issue of aggregation: “A group trend, however significant it may be, cannot accommodate the dynamic multicausality of the emerging language system of an individual, and interpreting a developmental phenomenon on mean trends and variance of group scores at one point in time underestimates the complexity of the developmental process” (Lowie & Verspoor, 2015, p. 78).

Along those lines, Pallotti (2022) cautions that the finding that, oftentimes, individuals do not fit the mean pattern of the group is not an empirical discovery, but a mathematical truism stemming from the very properties of the average; that is, by virtue of being a model, it can only provide a synthetic summary of several cases, without necessarily matching any one of them: “The fact that the average, or a regression line, do not correspond to any particular data point is not a limit of these relatively simple models, nor an issue to wonder about; rather, these models help us to solve concrete problems, like that of making sense of a number of sparse observations” (5). However, in our view, the main problem does not lie with the observation that group measurements may not say much about the individual (which might be obvious indeed) but with the fact that this equivalence (the learner being similar to the group) is an assumption implicitly held by many studies in AL, which is reflected in the widespread use of statistical models that employ analyses comparing means as a default (e.g. ANOVA-based analyses). Presumably this is because CDST proponents have argued that the starting point in research design should not necessarily be the distinction between quantitative and qualitative approaches, or even mixed methods, but rather the distinction between individual versus group-based designs (i.e. idiographic versus nomothetic, see Allport, 1937).

Variability vs. Generalizability

“If we reject simple linear causality, do we have to give up on generalizability?” (Larsen-Freeman, 2017, p. 34) As applied linguists we often want to assess whether a certain effect generalizes beyond the participants sampled to the wider population; i.e. we will want to test if results generalize both to the wider population of learners and the wider population of linguistic materials (Cunnings & Finlayson, 2015). However, in CDST and other process-oriented research agendas, the ways generalizability and prediction are employed are somewhat unconventional, i.e. the problem of the compatibility of CDST research with generalization has been widely discussed. On the one hand, the advantage of longitudinal studies with enough data points, in which various sub-components of the system under study are plotted and traced, is that they show in detail how each individual develops over time; however, they do not lend themselves to generalizations or general trends. On the other hand, Molenaar and Campbell (2009) argue based on the classic ergodic theorem observations across individuals can only be generalized under two strict conditions. The first condition is that the population should be homogenous and the statistical model which describes the group as a whole should apply to all subjects in the population. When the units of a group are strongly similar to each other, then we have an ergodic ensemble, where generalization from the individual to the group is justifiable. However, this condition is more often than not absent in data. The second condition stipulates that the data must remain temporally stable, such that the mean and variance should

not change between the measurements. However, developmental processes are almost always nonstationary and therefore nonergodic – and there is no evidence that the interaction of variables over time is the same for different individuals. In other words, the assumption in Gaussian statistics is that for different individuals within a single category or similar in some demographic way (e.g., intermediate learners of English with the same first language), generalizations can be made from a sample that transfer to the population they represent. Importantly, these generalizations are not warranted for development in the time domain, as the dynamic network of coupled variables impacting on the developmental trajectory of an individual cannot and should not be equated with interindividual variation at one moment in time (e.g., van Geert, 1991).

This is why many CDST scholars agree that the enterprise of generalizing across contexts, time, and systems is not realizable (or indeed necessary): “Instead of generalizable predictions, then, we are content to point to tendencies, patterns, and contingencies. Instead of single causal variables, we have interconnecting parts and subsystems that co-adapt and that may display sudden emergence of new modes of behavior” (de Bot & Larsen-Freeman, 2011, p. 23). Similarly, van Geert (2011) argues that case studies can have a generalizing power, depending on how they are linked to a particular theory: “a truly general theory of development processes is one that can be ‘individualized’ – it can generate theory-based descriptions of individual trajectories in a nontrivial sense” (276). What strikes us as important is that “contingency” does not preclude generalizing (Larsen-Freeman, 2017, p. 34). In other words, CDST is not purely descriptive, and it does not necessarily rule out causal explanations, falsifiability, generalizability, or prediction – consider, for instance, the hypothesis about “high degrees of variability accompanying rapid [L2] development” (Lowie & Verspoor, 2019, p. 2) – see also Hiver and Al-Hoorie’s (2020) description of how generalizable statements can be made at the level of system dynamics. Also, complex systems are not always in a state of unpredictable flux; we can often identify relatively stable phases and recurring patterns within the variation of system behavior, as discussed above. However, it raises the problem of (statistics-based, quantitative) methodology; i.e. it calls for an expansion of available methods. Specifically, it has been suggested (e.g. Hiver & Al-Hoorie, 2020) from a CDST perspective that causal relations – that is, *dynamic causalities* in language development – can only be disclosed in microgenetic, dynamic analyses. Such microgenetic studies do not necessarily have to focus on the individual. We agree with Bulté and Housen (2020) that it is also worthwhile to look for developmental patterns across learners, for example by combining multiple case studies (which does not necessarily involve averaging scores across learners), as demonstrate in this study.

Importantly, CDST supporters do, in fact, also value group studies for identifying general regularities or behaviors that hold for (a majority of) language learners (e.g., Verspoor et al., 2012). Molenaar’s work is not necessarily a plea for a focus on the individual case, but rather strives to build more adequate models that account for individual factors without giving up the search for general laws: “starting with analyses of intra-individual variation does not preclude valid generalization across subjects [...] In this way nomothetic knowledge about idiographic processes can be obtained” (Molenaar, 2015, pp. 37-40) – at least as long as the sample represents an ergodic ensemble. We thus believe, with Pallotti (2022), that rather than renouncing the construction of general, predictive, and thus falsifiable models, we should strive to develop multifactorial, non-linear and probabilistic models with a better fit to the data than

those currently available – such as generalized additive models (GAM), which strike what Pallotti (2022) calls “the right balance between over-simplification distorting reality and under-simplification presenting an overwhelming amount of unmanageable information” (691), i.e. the balance between the theoretical premise that everything is complex and dynamic, and thus irreducible to simple models, and the need to provide accounts of complex dynamic systems that are not limited to individual cases. By being able to (a) reconcile variability and generalizability (by analyzing nonlinear change over time in iterated learning and still detecting a general trend) and (b) disentangle mechanisms that have differing inherent time-courses GAM can contribute to our knowledge of *how* different individual learner differences work and over what time.

Quantitative vs. Qualitative Research

Finally, from a CDST perspective, it is assumed that individuals during L2 development are sensitive to learner-internal (e.g., cognition, socioaffect) and environmental-external (e.g., input, interactions) factors (Larsen-Freeman & Cameron, 2008; Lowie & Verspoor, 2015; Verspoor, 2017): “[W]e need to look at the ‘messy little details’ that make up the ‘here and now’ of real time. We need to take into account learners’ histories, orientations and intentions, thoughts and feelings” (Larsen-Freeman & Cameron, 2008, p. 159).

Particularly appropriate to the longitudinal study of complex systems is the use of “combinations or blends of methodologies” (Larsen-Freeman, 2015, p. 233), which are uniquely equipped to elicit and examine data through both quantitative and qualitative means. We suggest that combining methods to support a single, unified causal inference is also crucial for a more nuanced view of causality. Effects and outcomes are produced through a combination of complex conditions; e.g. more often than not, select phenomena can only retrospectively be interpreted as an effect, rather than being put forth as a prediction that linearly and causally extends into the future and thus also unobserved cases – which Larsen-Freeman and Cameron (2008) refer to as ‘retrocasting’ as opposed to ‘forecasting’.

Summary

On the whole, exploring *processes* as opposed to *products* in L2 development requires an adaption of existing methodologies, the goal being not necessarily to capture acquisitional outcomes at a single point in time, but rather to home in on the temporal specificity and scales of variability as concerns *how*, *when*, and *why* L2 development emerges. While many previous CDST-inspired approaches have focused on non-/linear change in individuals or smaller groups of learners, it has recently been argued (e.g., Kliesch & Pfenninger, 2021; Pfenninger, 2020) that a relatively large number of participants, each with sufficient repeated measures, can be helpful in generating more stable and generalizable results. That said, even group-based micro-longitudinal studies (though comparatively limited to date), which draw their statistical power from repeated measurements rather than from sample size alone, run the risk of falling victim to a ‘grand sweep effect’, i.e., presenting developmental trajectories that may hold at the group level, but which fail to generalize to the individual (e.g., Peng et al., 2022). Adequately addressing this group-to-individual generalizability issue necessitates not only micro-longitudinal designs that can capture temporal effects, but also more nuanced statistical procedures that are capable of reconciling generalizability, variability, and especially

individuality. To this end, Peng et al. (2022) argue in favor of the usefulness of person-centered approaches that *start* at the level of the individual and search for general developmental regularities by aggregating similarly structured processes in a bottom-up manner and, in an additional step, generalize the identified patterned outcomes to a larger group (see also Molenaar, 2015). While initial attempts at applying such person-centered approaches to L2 development do exist (see, e.g., Peng et al., 2022; Wirtz & Pfenninger, 2023), these have not attempted to exploit the strengths of GAMs as multifactorial, non-linear, and probabilistic models for operationalizing individual-level developmental trajectories in combination with time-series cluster analyses to identify ensembles of similarly structured learners. What is more, the present study is particularly noteworthy as concerns the developmental time span (i.e., eight years) and number of data collection points (i.e., 32). Thus, combining a micro-longitudinal study with dense time serial measurements over the course of a significant life period ([pre]primary school) with more nuanced statistical procedures tailored to capturing both the temporal specificity of development and the individuality involved therein position this study as uniquely qualified to answer outstanding questions concerning not only *which* and *how* individual differences variables work in L2 development, but more specifically *for whom* and over what time period?

The Study

Research Questions

The study reports on findings from a larger project conducted in a bilingual (pre)primary school in Switzerland, aiming to elucidate the complex ways that L2 English development emerges over time in 45 children in Switzerland who received German/English bilingual instruction over a period of eight years (age 5-12) in combination with emerging extracurricular exposure to English. Specifically, the following RQs will be addressed:

RQ1: Can we identify L2 supra-developmental patterns that transcend individual L2 developmental trajectories in primary school children?

RQ2: To what extent can periods of rapid development in these supra-developmental patterns be explained by individual learner differences in extracurricular L2 use?

RQ3: How can we qualitatively explain the cluster-related differences found in the quantitative analyses?

Participants

The dataset was taken from the AIM ('Age and Immersion') project (Pfenninger 2020, 2021a, 2021b). This study aimed to explore the factors that contribute to significant L2 (second language) growth in children attending bilingual and regular (pre)primary programs in Switzerland. It investigated how the development in L2 writing and oral language skills can be explained by a complex and dynamic interplay of individual and social factors. For this paper, we used the data obtained from 45 children who had received 50/50 bilingual instruction in German and English (so-called "partial CLIL" programs), 25 of whom were from German-speaking homes (new to English; 15F), while 20 were from English-speaking homes (new to German; 11F). The former were referred to as 'earlyPAC' due to their rather early age of onset of L2 acquisition (age 5), compared to the other groups in the project. The latter comprises

students for whom German was both their L2 and the language spoken locally, and whose parents were native speakers of English. Following Festman's classification in 2018, they were identified as 'international students' (referred to as earlyPAC-int). These were children born outside of Switzerland, raised with another language (English) abroad, and continuously exposed to it within their international families even after relocating to Switzerland (i.e., while living and working there). These international students were integrated into the same classes as the children in the earlyPAC group.

All children were matched for socioeconomic status (and comparable home literacy environments) and were 5 years old during the initial data collection, with the study continuing until they reached the age of 12. Throughout this period, they received between 28 lessons (Grades 1–3) and 30 lessons (Grades 4–6) per week, amounting to approximately 6.5 to 11 hours spent at school each day. Additionally, they were provided with conventional English/German-as-a-second language instruction: 6 hours in Grade 1, 5 hours in Grades 2 and 3, and 4 hours in Grades 4–6 (for further details, refer to Pfenninger, 2020).

Tasks and Procedure

Data collection took place four times a year spanning eight school years, from ages 5 to 12, at consistent intervals of every 3 months. This resulted in a total of 32 data collection points for each participant and task. At each measurement, participants wrote a timed English narrative (topic: the plot of their favorite movie, book or TV series) and completed a re-telling task, which required them to narrate the plot of a silent video they had previously watched at the end of primary school – see Pfenninger (2020) for a more detailed description of these tasks (both 45 minutes). Furthermore, participants were tasked with completing a language awareness questionnaire comprising open-ended questions. Additionally, individual semi-structured interviews, each lasting 10 minutes, were conducted with the students. These interviews aimed to collect information about various aspects, including their use of English in daily activities (extracurricular pursuits), language proficiency, emotional experiences (motivation, anxiety, enjoyable and challenging moments, etc.), the influence of parents, peers, teachers, and language assistants, their English language progress, and their self-reflections on the narratives they created, highlighting both strengths and areas needing improvement. Verbatim transcripts were generated from the recorded interviews. Finally, a questionnaire including 16 closed-ended items on a 10-point scale was administered at each data collection time, tapping into the children's extracurricular use of English:

- using (learning) apps in English (2 items, $\alpha = .880$)
- using English for gaming (4 items, $\alpha = .891$)
- surfing the net/checking pages in English (3 items, $\alpha = .884$)
- using English on social media (2 items, $\alpha = .879$)
- watching movies, series, YouTube in English outside of lessons (3 items, $\alpha = .893$)
- listening to English songs (2 items, $\alpha = .896$)

All materials are available at <http://www.iris-database.org>.

Data Analysis

Children's essays were transcribed using the CLAN program and CHILDES (McWhinney 2000), and, once the transcripts were completed, coders with expertise in linguistics and who

were additionally bilingual English-German speakers coded the student's speech using the koRpus package in R (version 0.11-5). Transcripts were coded for the following aspects:

- morphosyntactic complexity and fluency:
 - mean length of utterance (MLU, i.e. number of morphemes per word)
 - text length in tokens i.e. word count (W)
 - clauses per T-Unit (C/T)
- lexical diversity:
 - Measure of Textual Lexical Diversity (MTLD) (McCarthy & Jarvis, 2010), which is independent of text length;
- accuracy: total number of utterances produced correctly, total number of utterances that contained morphosyntactic and/or semantic errors (token errors and type errors).

To answer the above-mentioned RQs we proceeded in four phases: (1) Generalized additive modeling to quantify individual-level oral and written developmental trajectories; (2) time-series clustering methodology to identify 'supra' developmental patterns (see below); and (3) visual-quantitative analyses alongside (4) qualitative content analyses, both of which were geared towards describing trajectorial differences, homing in specifically on factors relating to extramural exposure to English. The coding procedures and analysis scripts for the quantitative phases outlined below can be found on this article's OSF repository (https://osf.io/purk9/?view_only=13823b26ff744bd593c4c9d404e0ed8c).

Phase 1 (individual-level written and oral GAMs for each participant)

In order to be able to model complex nonlinear L2 trajectories and intra- and inter-individual variation as well as to take account of autocorrelation (nested dependencies, multivariate data, repeated measures), GAM was employed using the *mgcv* R package (Wood, 2006). We included the sole nonlinear predictor *Time*, thus giving us the individual trajectories of each subject for their oral and written GAM. To account for potentially nonlinear differences over time with respect to the general time pattern for each of the five written and oral L2 measures per participant, we included factor smooths for *Time* and *L2 Measure*. Importantly as it concerns this modeling procedure, we did not fit the *Time* effect for groups. Instead, we aimed to assess development at the broader level, taking the respective group into account only after the cluster analysis to determine the extent to which the earlyPAC vs. earlyPAC-int binary was represented at the cluster level (see next phase).

Phase 2 (time-series cluster analysis using individual-level predicted Time 0s, intercepts, and edfs as inputs)

Besides depicting individual-level trajectories of L2 development, we intended to distill developmental patterns that transcend the individual heterogeneity as they emerge from the data. In other words, while learners may indeed evince individually owned developmental trajectories, salient between-person patterns – so-called “supra patterns” (Baba & Nitta, 2014, p. 30) – may yet be hidden in the data; clustering methodologies aid in identifying these. What is more, *time-series* cluster analyses facilitate the additional investigation of the time domain, reconciling inter- and intra-individual variation in order to advance our understanding of “lawful regularities about L2 learners' developmental processes” (Peng et al., 2022, pp. 905–

906). To this end, we extracted several parameters from the individual-level written and oral GAMs that should be representative of critical aspects of their developmental pathways:

- *Predicted Time 0*: In order to capture subjects' respective baseline L2 performance, each learner's performance was estimated at time point 0, using their individual-level written and oral GAMs. This was done also in order to account for intra-individual fluctuations in L2 performance at the onset of the study, which, if neglected, runs the risk of producing over- or underestimations of subjects' actual baseline performance (see also Kliesch et al., 2022).

- *Intercept*: Note that, in non-linear GAMs, the intercept refers to the conditional mean of the response variable (i.e., participants' L2 scores). Simply put, the intercept in this case approximately represents subjects' mean development over time.

- *Edf*: The effective degrees of freedom (*edf*) is an estimate of how many parameters are needed to represent the respective smooth (Wieling, 2018), which essentially indicates the amount of nonlinearity of the smooth (the higher the *edf*, the higher the nonlinearity). Thus, by extracting the *edf* parameter from each learner's individual GAMs, we account for the degree of nonlinearity in participants' respective developmental pathways. Importantly, the *edf* does not indicate the *directionality* of said nonlinearity; that said, as Pfenninger (2020) showed, there were only exceedingly few periods of (significant) decline in learners' development, so the *edf*, in our specific case, can be taken as the amount of nonlinearity in participants' upwards developmental trajectories.

By employing these three parameters as our participant-level inputs for the time-series clustering, we follow Molenaar and Campbell (2009), Peng et al. (2021, 2022) and Wirtz and Pfenninger (2023), who suggested to not only adopt a person-centered approach but also to identify subgroups of similar individuals as ergodic ensembles so that the findings at the subgroup level and those of the individuals composing the subgroup are mutually inferable. At the cluster level, components in each cluster that emerged from the analysis evince similar developmental characteristics (based on the statistical parameters); therefore, the clusters arguably satisfy the condition of being homogeneous: "This improvement in ergodicity is important to the group-to-individual generalizability issue. That is, although it may not be appropriate to directly generalize between an individual and the group as a whole, it is feasible to generalize between an individual and the cluster identified" (Peng et al., 2022, p. 18; see also Wirtz & Pfenninger, 2023). We then conducted a time-series hierarchical cluster analysis (HCA), a multivariate exploratory technique used for identifying new groups or patterns in a bottom-up manner (Staples & Biber, 2015). Within the HCA, we used the Manhattan distance matrix to quantify and reflect the (dis)similarity between individuals' respective Time 0, intercept, and *edf* parameters. We adopted Ward's method as a linkage method so as to minimize within-cluster variance during the clustering process. Overall, the time-series HCA computed subgroups of similar 'objects' in the data, in this case subgroups of similarly behaving learners. After determining a solution to the cluster via the average silhouette width (see Levshina, 2015, p. 311), we plotted (a) a dendrogram and (b) the individual clusters according to the writing and oral data.

Phase 3 (Visual quantitative analysis of earlyPAC vs. earlyPAC-int binary, L2 extracurricular use) and Phase 4 (qualitative content analysis to describe differences in clusters, i.e., learners with trajectorial similarities)

Phases 3 and 4 involved a multidimensional, integrative approach to *explaining* the trajectorial differences as captured and modeled by the clustering procedure, guided by the notion that “[i]t’s not enough to highlight individual variability [...] We still have to explain [it].” (Ellis, 2007, p. 23). In terms of potential signature dynamics—i.e., qualitatively/quantitatively/visually determined underlying dynamic patterns (Dörnyei, 2014; Hiver, 2017)—we place special focus on the role of extramural exposure to English during learner development, and additionally how this may relate to, e.g., changes in strategies, shifts in affective states and motivational flow, cognitive events, and (emergence of new) strategies (i.e., phenomena identified as contributing to periods of rapid development [see Pfenninger, 2020]). More specifically, in addition to whether the earlyPAC vs. earlyPAC-int binary was represented at the cluster level, we (a) addressed in what constellations differences in the extracurricular L2 use was represented in the different clusters and (b) explored which or whether any qualitative aspects set the clusters apart from one another, so implying that the clusters characterized by trajectorial differences were also distinctive of certain qualitative aspects. This qualitative component helps make sense of the quantitative data. Understanding that development is non-linear and varies for different linguistic features or writing abilities in different learners is crucial. However, it remains uncertain how this knowledge can help us interpret a series of elicited oral language and writing episodes (see Norris & Manchón, 2012). The procedure for reconciling quantitative and qualitative approaches to explaining rapid development was as follows:

- In order to identify repeller states (i.e., rapid L2 developmental phases) as opposed to attractor states (i.e. more stable phases), we used visual methods for significance testing by fitting additive models with superimposed periods of significant L2 change for the individual slopes (see Simpson, 2014), which highlight phases of significant growth in each participant’s trajectory (see also Pfenninger, 2021; Pfenninger & Kliesch, 2023; Wirtz & Pfenninger, in press). We then plotted learners’ extracurricular English use to visually determine how such extramural exposure relates to participants’ rapid developmental phases.

- Following this, we extracted the periods of significant change and categorized them binarily (i.e., significant change [repeller state] versus non-significant change [attractor state]). Our goal here was to assess the cluster-specific differences in the scores of subjects’ extracurricular English use during rapid development, focusing specifically on (a) whether extracurricular English use was higher during periods of rapid development (variable: *Significant Increase [TRUE/FALSE]*), (b) for which extracurricular domain (variable: *Extracurricular Activity*), and (c) for which clusters (variable: *Cluster*). To this end, we specified a linear mixed-effects model with a three-way interaction effect:

$$\text{lmer}(\text{Extracurricular English Use} \sim \text{Extracurricular Activity} * \text{Significant Increase} [\text{TRUE/FALSE}] * \text{Cluster} + (1 | \text{id}))$$

We included by-participant random intercepts to account for repeated measures and subject-level idiosyncrasy. Importantly, the time domain was not included in this analysis, as

the primary question was whether extracurricular English use was more prominent during periods of rapid development or more stable periods.

- Finally, we complemented the quantitative results with a qualitative thematic analysis, focusing on moment-to-moment changes, stimuli for change, and phase shifts from repeller to attractor states. Specifically, we follow a method integration approach as proposed by Seawright (2016), which involves applying one method to “produce the final inference” while the other – in this case the qualitative approach – “is used to design, test, refine, or bolster the analysis producing that inference” (8). Thus, instead of ‘mixing and matching’ two methods independently and then comparing the results at the end, “one method is used to overcome the weaknesses of another method *while* that method is being applied” (Hiver et al., 2021, p. 10, italics original). The qualitative data were transcribed and digitalized for analysis using the software MAXQDA (<http://www.maxqda.com/>). Following Mercer (2015), the data were first coded first in an open, grounded manner, coding line-by-line to allow all aspects of the data to be considered and to ensure potentially unexpected features were included in the analysis. In the following rounds of coding, codes were combined or expanded until categories began to form. The codes, categories, memos and matrices were then examined specifically with CDST in mind, with a focus on characteristics of complex dynamic systems, their interrelations and dynamics. Specifically, the categories that emerged included indications of possible changes to contingent stabilities, as revealed by the quantitative analysis. This process was done separately for each learner – focusing on each individual and developing their profile before moving on to the next. Finally, after the individual level of analyses, the data were examined for possible patterns and interactions across each level and for both learners.

Results

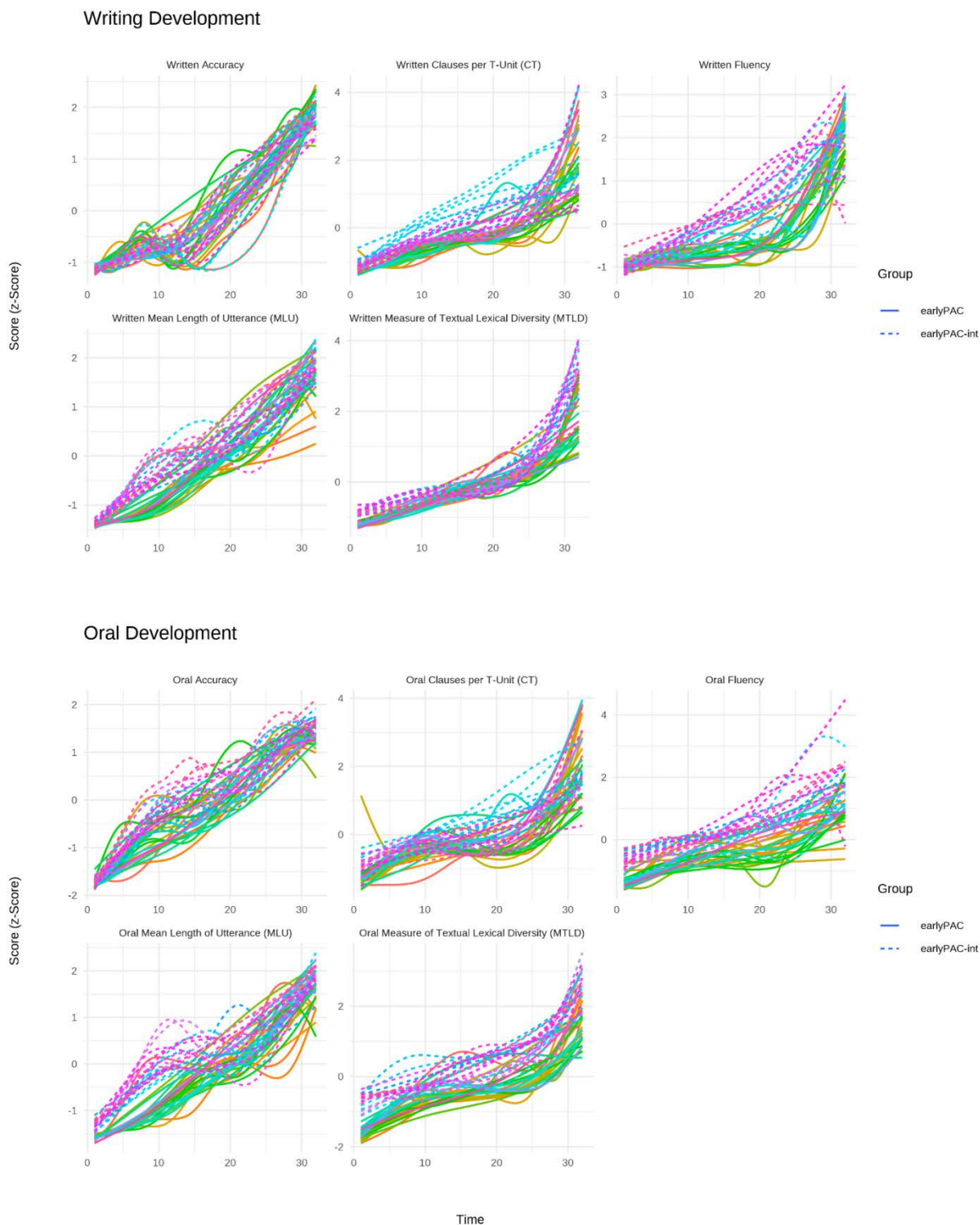
In the following, results will be presented corresponding to the four phases of analysis outlined above. Phases 1 and 2 are geared towards answering RQ1 concerning the identification of supra developmental patterns, whereas Phases 3 and 4 address RQ2 and RQ3 as it relates to *explaining* cluster-related trajectorial differences.

Phase 1: Individual-level written and oral GAMs for each participant

Figure 1 presents the descriptive statistics, that is, the 45 individual developmental pathways of the earlyPAC (n = 25) and earlyPAC-int (n = 20) learners over the course of eight years. It becomes clear that the participants differed in terms of their rate of oral and written development alongside the degree of idiosyncrasy subjects evinced throughout (pre-)primary school. The GAMM analyses in Pfenninger (2020) assessing the effects of time and age of first bilingual exposure (i.e., age of onset, [AO]) also show that learners made significant L2 gains over time across the oral and written measures. Based on these visualized descriptive trajectories, individual GAMs for the oral and written data were run and the three parameters of interest (predicted Time 0, i.e., initial starting-point of development; intercept, i.e., predicted mean development; and the edf value, i.e., the subject-specific degree of nonlinearity in each learner’s trajectory of development) were extracted, in accordance with the first phase of analysis.

Figure 1

Individual L2 Trajectories across Written and Oral Modalities for the Earlypac and Earlypac-int Groups



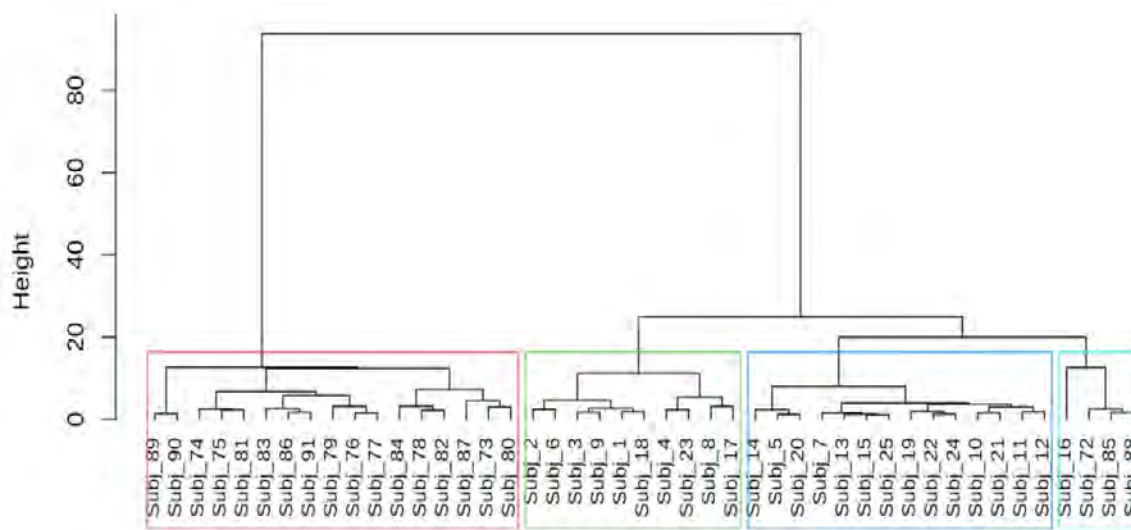
Phase 2: Time-series clustering of learners' developmental pathways

Figure 2 shows the dendrogram for the four-cluster solution from the time-series cluster analysis on subjects' predicted Time 0, intercept, and *edf* values. On the basis of the average

silhouette width, a two-cluster solution emerged as the optimal clustering solution (average silhouette width = 0.452), which strictly – and perhaps unsurprisingly – captured the earlyPAC–earlyPAC-int binary, with the exception of three earlyPAC-int subjects who were found to cluster with the earlyPAC students (subj_72, subj_85, subj_88). That said, in order to better capture more nuanced differences *within* the two-cluster solution and thus better delimit how certain students stand out against the backdrop of the (sub-)group, we proceeded under the second-best cluster solution, a four-cluster solution (average silhouette width = 0.333), which resulted in the original cluster 2 housing the earlyPAC students being separated into three clusters (visualized by the green, dark blue and light blue rectangles in Figure 2).

Figure 2

Dendrogram of the Four-cluster solution from the Time-series Cluster Analysis

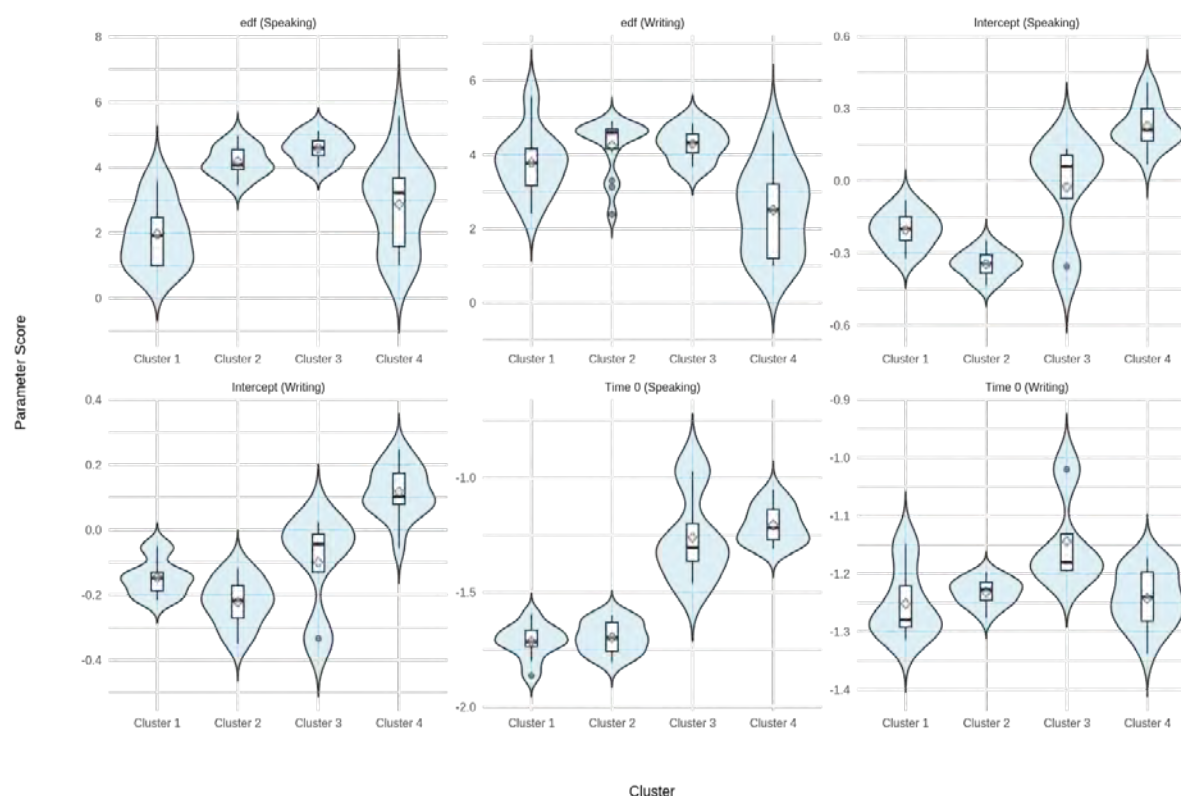


Note. Green = Cluster I; dark blue = Cluster II; light blue = Cluster III; red = Cluster IV

Figure 3 displays the cluster-related differences in learners' *edf*, intercept, and predicted Time 0 values. A non-parametric Kruskal-Wallis rank sum test revealed significant differences between the four clusters across the six measures of interest: Predicted Time 0 (written data: $H_{(3)} = 8.74, p = 0.033$; oral data: $H_{(3)} = 33.30, p < 0.001$), intercept (written data: $H_{(3)} = 32.87, p < 0.001$; oral data: $H_{(3)} = 36.07, p < 0.001$), and *edf* (written data: $H_{(3)} = 19.15, p < 0.001$; oral data: $H_{(3)} = 25.15, p < 0.001$).

Figure 3

Cluster-related Differences in the Predicted Time 0, Intercept, and edf Parameter Inputs



Note. Violin plots visualize the probability density of the data at different values and boxplots indicate the median and the respective quartiles. The rhombus at the center represents the mean, and the grey dots are outliers.

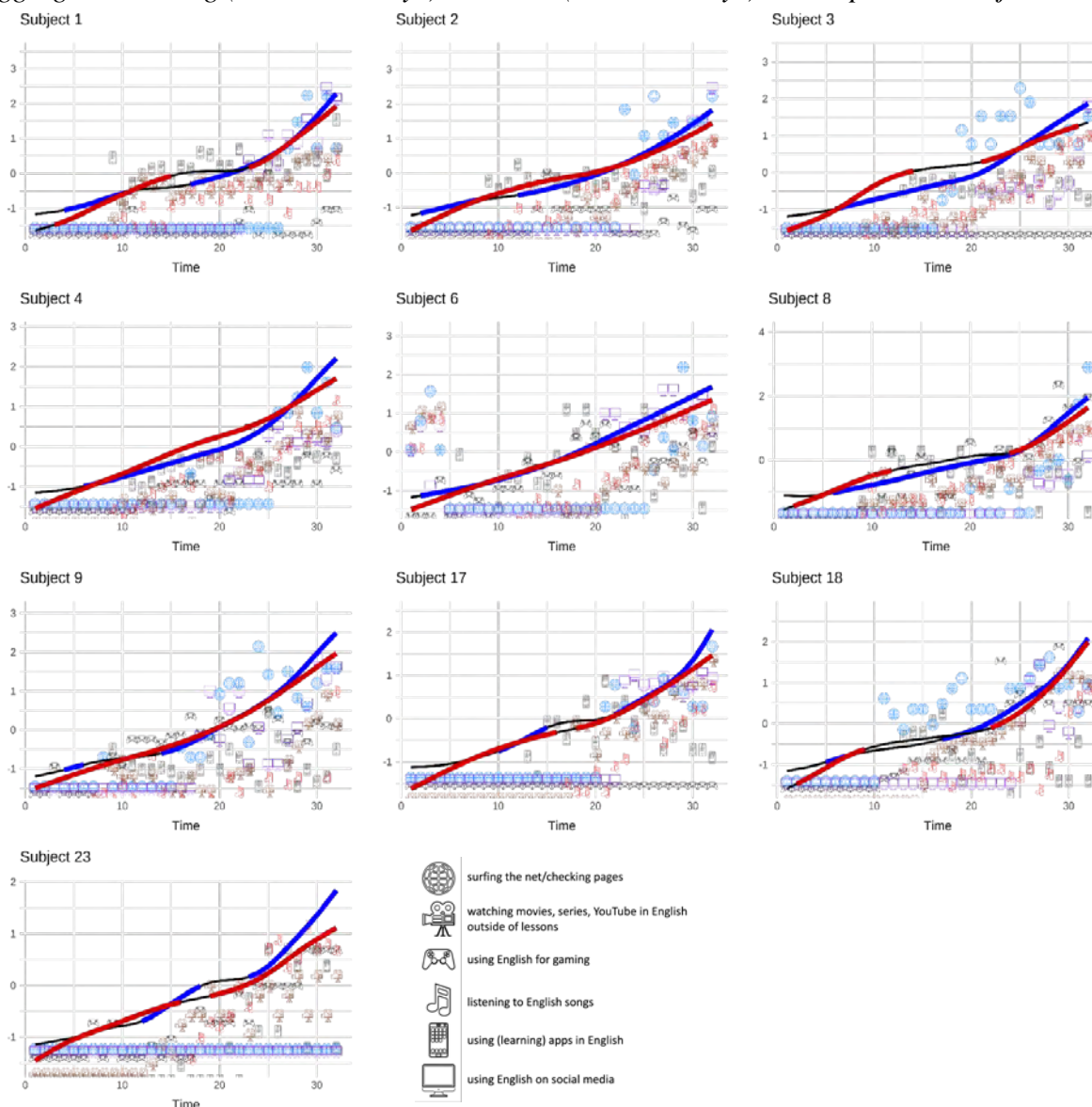
Phase 3: Unraveling the idiosyncrasy and commonality in developmental processes

In order to visually inspect how the clusters developed (differently) over the course of the testing period, we plotted the aggregated written and oral developmental trajectories with superimposed periods of significant change, operationalized as periods during which the 95% confidence interval on the first derivative (i.e., rate of change) of the trend did not include zero (see also Pfenninger, 2020; Kliesch & Pfenninger, 2021; Wirtz & Pfenninger, in press) and the z-scored extracurricular English exposure (see Figures 4, 5, 6, and A1 in the appendix). This aids us in evaluating (a) the extent to which the clusters differed in their (periods of significant) development and (b) whether these rapid developmental states were characterized by higher extracurricular English exposure.

Figures 4 and 5 illustrate the visual methods for significance testing for Clusters 1 and 2. Between these two clusters, there are clear trajectorial differences as captured by the *edf* parameter, i.e., the degree of nonlinearity extracted from the individual GAMs in Phase 1. Specifically, Cluster 1's writing and oral trajectories tend to develop at largely different rates, though comparatively linearly, whereas Cluster 2's writing and oral trajectories follow similar pathways at the inter-individual level, but nonlinearly by comparison. Additionally, across both clusters, trajectorial directionalities differed temporally such that the oral trajectories were steeper in slope during initial periods, but by the end of testing (i.e., typically around Time 25), the writing trajectory became both steeper in slope and overall higher than the oral one.

Figure 4

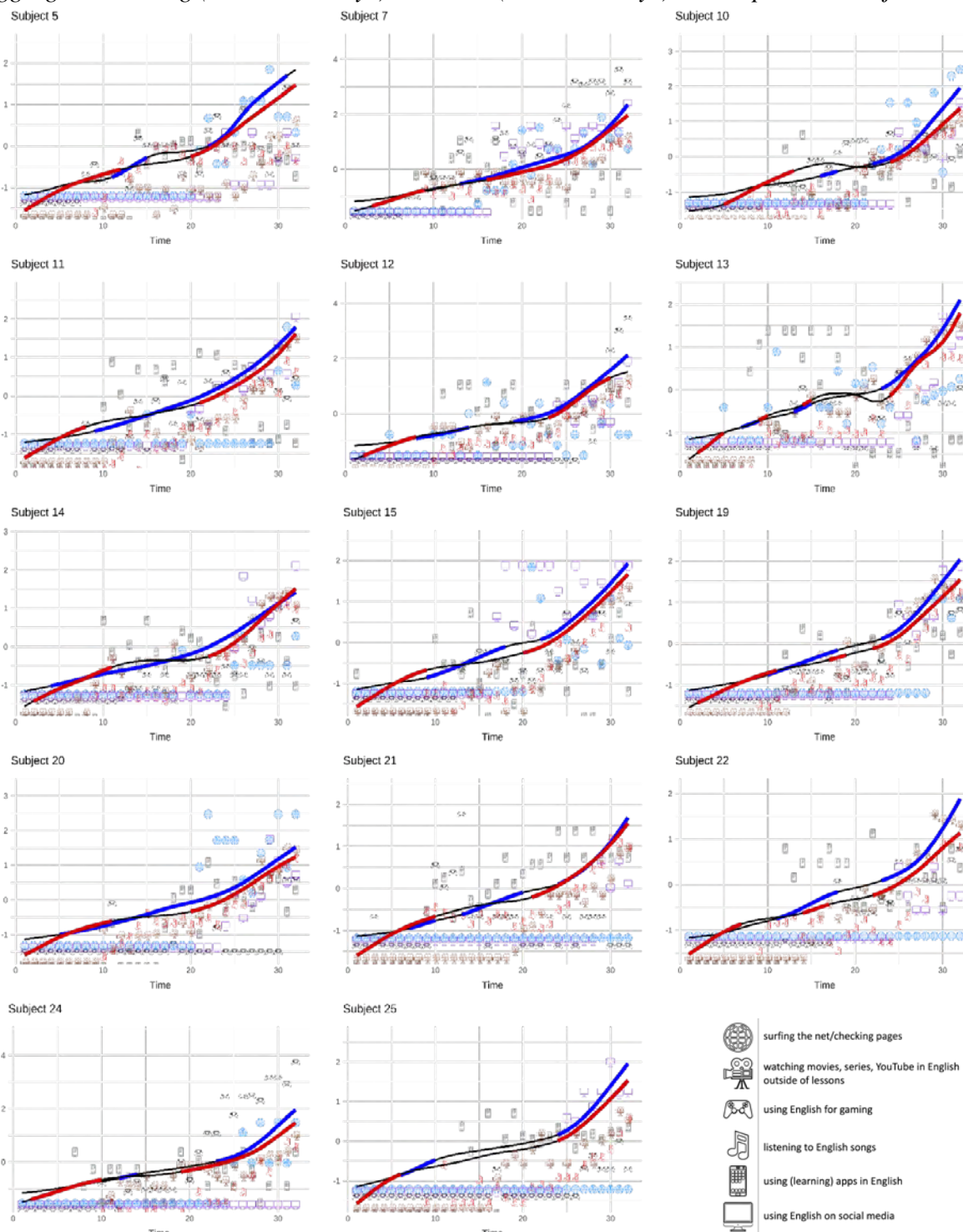
Fitted Additive Model with Superimposed Periods of Significant Change in Cluster 1 Students' Aggregated Writing (BLUE Overlays) and Oral (RED Overlays) Developmental Trajectories



The emojis represent the z-scored extracurricular English exposure values: Cell phone (using [learning] apps in English); controller (using English for gaming); internet (surfing the net/checking pages in English); computer (using English on social media); film (watching movies, series, YouTube in English outside of lessons); music notes (listening to English songs).

Figure 5

Fitted Additive Model with Superimposed Periods of Significant Change in Cluster 2 Students' Aggregated Writing (BLUE Overlays) and Oral (RED Overlays) Developmental Trajectories

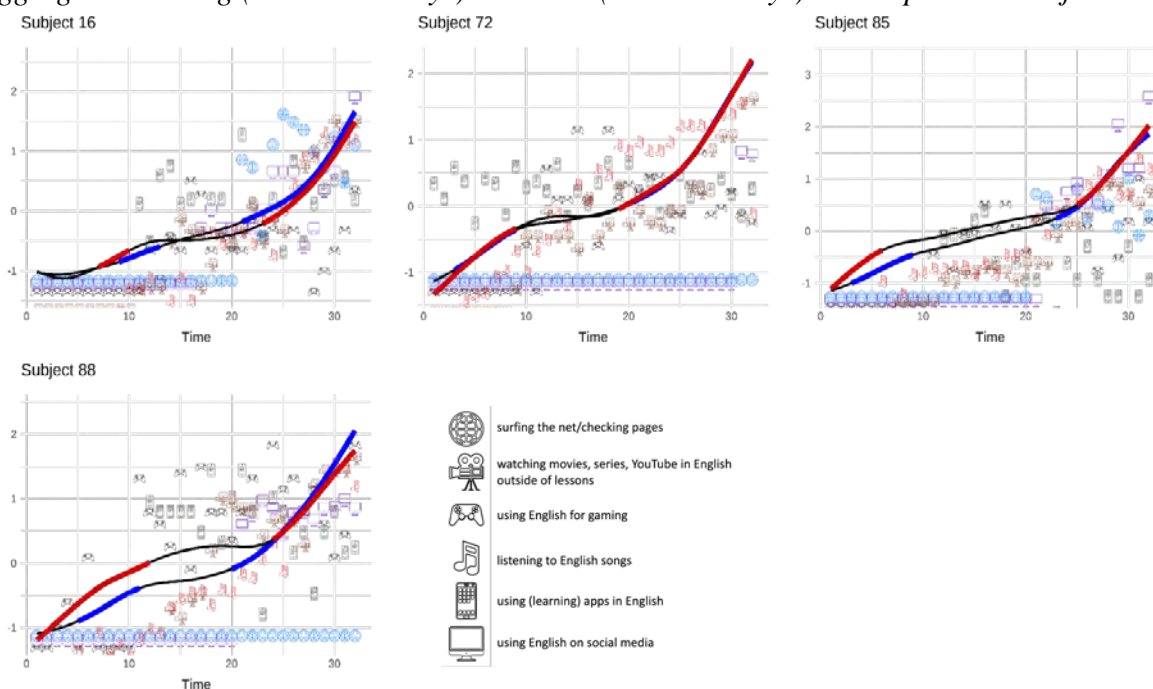


As Figure 6 illustrates, Cluster 3 houses a single earlyPAC (Subject 16) in addition to three other earlyPAC-int students, and Cluster 4 (Figure A1 in the appendix) exclusively includes earlyPAC-int students. As it concerns Cluster 3, we note the similar intra-cluster trajectorial shapes, both for subjects' writing and oral development. Specifically, there appears to be rapid development in the early stages (typically until Time 10), after which periods of comparative

stability can be observed, i.e., what appears to be attractor states. During later stages of development (i.e., at Time 20), subjects again evince phase shifts as captured by the superimposed periods of significant change. Concerning Cluster 4 (exclusively earlyPac-int), the writing and oral developmental trajectories do not develop as closely in tandem as do those of Cluster 3. The attractor states are also less systematic in Cluster 4, and, moreover, do not occur at similar periods inter-individually as was observed in Cluster 3.

Figure 6

Fitted Additive Model with Superimposed Periods of Significant Change in Cluster 3 Students' Aggregated Writing (BLUE Overlays) and Oral (RED Overlays) Developmental Trajectories



Despite idiosyncratic developmental avenues, increased extracurricular activities in English were particularly noticeable during periods of rapid development (as captured by the superimposed periods of significant change), but the effects seem temporally limited to the end stages of testing. That is, in the early stages of the data collection, in which the GAMs also recorded (statistically) significant development, students indicated lower rates of English use during extracurricular activities, though this is likely attributable to subjects' young age, i.e., 5 years old at the beginning of the experimental procedure. In light of this, the increased extracurricular activities in English are likely dynamically intertwined with subjects' increasing age and thus also more avid use of, e.g., apps, social media, etc. That said, we indeed observed individual cases in which extracurricular reports closely coincided with developmental jumps throughout (most of) the data collection, e.g., Subjects 11 and 19 in Cluster 2, Subject 16 in Cluster 3, and Subjects 74, 78, 79, 81, and 91 in Cluster 4, i.e., across clusters and regardless of any earlyPAC–earlyPAC-int binary. On the flip side, there were also cases of disparity in which extracurricular English exposure was particularly high during both the quantitatively captured attractor states as well as during repeller states/phase shifts, e.g., Subjects 13 and 18 in Cluster 2 and Subjects 72 and 88 in Cluster 3.

While the GAM visualizations with superimposed periods of significant change and the plotted extracurricular English exposure do indeed allow for a time-dependent visual analysis concerning whether increased extramural English exposure temporally coincides with rapid development, such visual inspection may not necessarily lend itself to generalizable statements. In order to counterbalance this and so facilitate initial insights concerning the relationship between cluster-related rapid development and increased extracurricular English use, we modeled a three-way interaction taking into account whether increased rates of extramural English exposure were related to significant development for each cluster, the results of which are visualized in Figure 7.

As it concerns the writing data, Figure 7 illustrates that periods of rapid development were characterized by higher extracurricular activities in English, but exclusively for Clusters 1 and 2. Contrariwise, Clusters 3 and 4 (i.e., the earlyPAC-int students and one earlyPAC) evinced more cases of disparity, such that the quantitatively captured phase shifts were not indicative of higher use of extracurricular English. With respect to the oral data for the Clusters 1, 3, and 4, we similarly found no evidence that rapid development was characterized by increased English exposure during extracurricular activities. Exclusively Cluster 2's phases of statistically significant developmental jumps in their oral production temporally coincided with higher reports of extracurricular English use.

Phase 4: Qualitative content analysis to describe differences in clusters, i.e., learners with trajectorial similarities

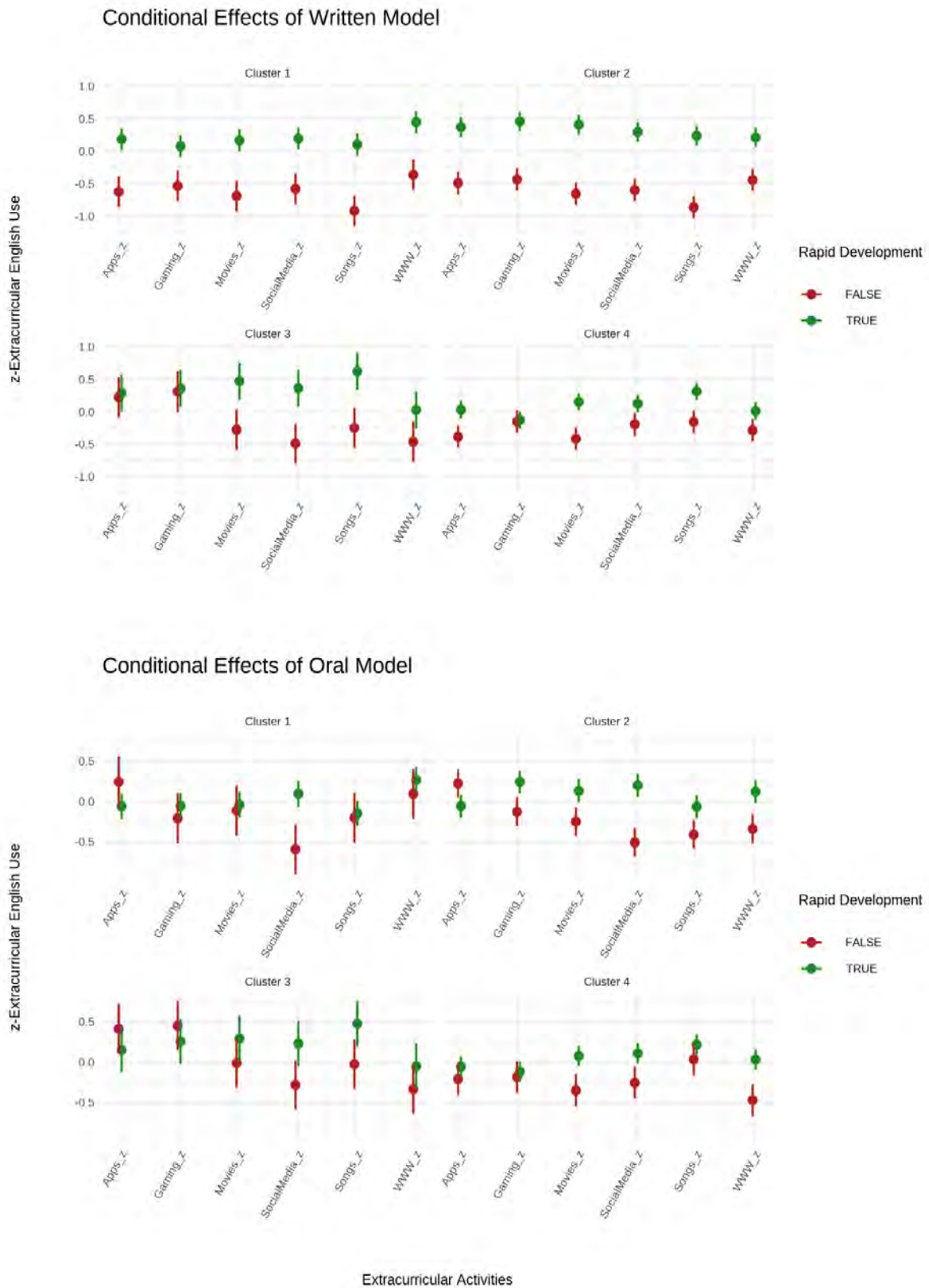
The analysis of the interviews addressed RQ3 and provided additional qualitative support for the meaningfulness of the GAMs and the cluster analysis. For instance, it explained the constellation of Cluster 3 above, which included one student from a German-speaking home (subject 16) and three children from English-speaking homes. The fact that subject 16 showed similar trajectorial shapes as subjects 72, 85 and 88, both for subjects' writing and oral development, could be a reflection of her close ties with subject 85, who she became friends with in Grade 1 of primary school (age 7):

- (1) "Jessica is my new friend. I always wanted to have a best friend." (earlyPAC_16_t11)
- (2) "I try to listen well to Jessica and learn from her." (earlyPAC_16_t18)

The four students in Cluster 3 showed impressive oral and written L2 development (see Figure 6); although they also experienced phases of stability at the beginning of primary school, their learning trajectories included long stretches of significant L2 growth. What is striking about their qualitative data is that they reported positive levels of confidence, high achievement motives, and a sense of progress throughout the 8 years of data collection, which could have contributed positively to her degree of in-class comfort:

Figure 7

Conditional Effect Plot Showing the Fixed Interaction Effects between Rapid Developmental Phases and Clusters on Degree of Extracurricular English Use for each Extracurricular Activity



(3) “I would like to learn to write better because I want to be famous.” (earlyPAC_16_t9)

(4) “I try to use the most difficult words in my vocabulary when writing the essays. I compose sentences using adjectives, verbs and nouns for precision, clarity and impact.” (earlyPAC-int_72_t26)

(5) “My parents and friends also read my essays, and they like them, too. That’s why I really make an effort every single time.” (earlyPAC-int_88_t29)

The content analysis of the qualitative data also revealed that clusters who showed rapid development that was characterized by increased English exposure during extracurricular activities were also the ones who perceived a discrepancy between in- and out-of-school encounters with English. To be more specific, the biggest difference between Clusters 1 and 2 vs. Clusters 3 and 4 was what Henry (2013) called the “authenticity gap” (133) or the “credibility problem” (13) i.e. the perceived gap between the English students learn in school and the English they use outside (see also Sundqvist & Olin-Scheller 2013). To give an example, some of the children interviewed – both in Cluster 2 – described in Grade 5 (age 11) how difficult it was to reconcile English inside and outside of school:

(6) “The English in our free time is cool, but the English in school is a bit monotonous.” (earlyPAC_15_t28)

(7) “I like the English language because I watch YouTube videos in English. However, I don't have good English grades at school, which demotivates me.” (earlyPAC_10_t26)

While participants in Clusters 1 and 2 praised the opportunities for creativity and authentic self-expression they experienced in the bilingual program they attended, English lessons had little to offer in developing language competence, compared to digitally mediated English-language environments outside of school:

(8) “I think my English is very good because I use it while gaming and watching YouTube and Tiktok. At school, the vocabulary is always the same.” (earlyPAC_9_t27)

In stark contrast, the learners in Clusters 3 and 4, who exclusively included PAC-int students except for one child from a German-speaking home, did not experience an authenticity gap, i.e. they did not report that the language classrooms poorly replicated patterns of interaction in real life.

Discussion

Our study corroborated previous findings (e.g. Henry et al., 2018; Henry & Lamb, 2020) that digital technologies can provide unrivaled opportunities to engage with learning in innovative and identity-congruent ways. In line with, for instance, de Graff (2015), we found that contact with English outside of school was a particularly strong predictor for learner outcomes, specifically as it concerns the extension of students’ network of social ties beyond school boundaries. Indeed, Pfenninger (2020, 2021a) also found both from a cross-sectional and longitudinal perspective that increased extracurricular activities were beneficial for L2 outcomes at the end of primary school. The current analysis expanded on these results, taking

a temporal-relational perspective on developmental growth and its quantitative and qualitative relationship to extramural English exposure.

Whereas extracurricular English use seems to temporally coincide with significant developmental jumps in certain clusters (50% of the identified clusters of learners – and only in relation to the last 2-3 years of English acquisition in primary school [ages 10-12]), pointing towards temporally limited effects of extramural English exposure, neither the visual-quantitative GAM analyses nor the linear mixed-effects models could explain cluster-related disparity in terms of why, in select clusters, phase shifts were indicative of increased extracurricular English activities. For this, it was necessary to draw on participant-level qualitative data (see, e.g., Pfenninger & Singleton, 2017; Pfenninger & Kliesch, 2023), considering introspective accounts as to the similarities and differences of cluster-specific drivers for change in rapid developmental phases.

Clusters 1 and 2, who were all earlyPAC students (i.e. children coming from German-speaking families) and who experienced rapid L2 growth characterized by increased English exposure during extracurricular activities, perceived a discrepancy between in- and out-of-school encounters with English. Interestingly, the learners in Clusters 3 and 4 did not perceive difficulties in bridging these two cultures, arguably because the classroom environment was not the only learning context for them (all but one came from English-speaking families). Students tend to expand this gap due to the lack of acknowledgement and utilization of their skills gained by extramural exposure (Sundqvist & Olin-Scheller, 2013, p. 332), which leads to “motivational dissonances between school and extramural English” (Sundqvist & Olin-Scheller, 2013, p. 330). In this context, Henry (2013, p. 133) has argued for enhancing learners’ “self-authenticity.” He defines the latter as engagement with language and with activities congruent “with core self-conceptions” and which, according to him, can have motivating forces (Henry, 2013, p. 141; see also Pinner, 2014).

Criticizing the often “oversimplified” and smooth interactions in textbooks and the classroom, several scholars investigating learner autonomy (e.g. Roberts & Cooke, 2009; Ushioda, 2011) have also highlighted how important it is for learners to find their “authentic voice.” Ushioda (2011), for instance, argues that in order to become autonomous learners, students should “speak as themselves.” She refers more specifically to the often simplistic and pseudo-communicative texts and tasks in textbooks and in the classroom context, which (over-)emphasize language practicing instead of the expression of personal ideas or identities. However, learners should “through the medium of the target language [...] express their own preferred meanings, interests and identities” (Ushioda, 2011, p. 17).

According to Leona et al. (2021, p. 11) countries are still “very behind” in addressing the increased extramural exposure to English of young students and its potentially highly motivating effect on students to learn English. The perceived gap concerning the use of English inside and outside of school calls for methods and techniques that allow teachers to “(1) build scope for agency into activity designs, (2) craft activities in ways that accommodate the interests and concerns of students as unique individuals, and (3) make flexible accommodations that support personal preferences and enable authentic self-expression” (Henry, 2021, p. 229). This demonstrates how the shift in students’ encounters with English has serious implications for teaching, not least in terms of motivation (Henry 2013).

What is more, the cluster analysis appears to be catching trajectorial differences alongside differences in starting conditions, albeit unsurprisingly in light of the cluster inputs (i.e., Time 0 and *edf*). This illustrates that even if the “outcome” is identical under experimental and non-experimental conditions, there is no logical rationale to assume that the “process” is also the same (Hiver et al., 2021), both with respect to L2 development alongside the factors that influence it. In a complementary vein, the qualitative explanations of the cluster analyses underlined the importance of positive social relations at school for L2 learning. Interactions with classroom friends may be an important contributor to L2 development in multilingual primary school classrooms (Bushati et al. 2023), such that bilingual learners must acquire linguistic communication skills in addition to literacy and academic language skills.

Finally, this study showed that there are ways to transcend individual learner variety in order to reconcile an idiodynamic approach with generalizability by capitalizing on commonality and variability in processes of language development. Although statistical robustness and generalizable results are not always the goal of CDST-informed studies, arguably there is value in identifying developmental regularities, i.e., interindividual recurring phenomena (Ellis, 2007; Pfenninger, 2021a). In this study, we advocate for the use of generalized additive modeling procedures not only in order to account for autocorrelation in nonlinear patterns and capture periods of rapid development (repeller states) versus periods of more stability (attractor states), but more importantly to better operationalize individual-level developmental pathways, e.g., through the extraction of GAM parameters such as subject-level intercepts, predicted Time 0, and *edfs*. These take into account mean development, starting levels, and the degree of nonlinearity respectively; by combining these parameters with time-series clustering methodologies, we can operationalize individuality in order to transcend the individual heterogeneity and make inferences that are more strongly predictive of the trends of certain subgroups of individuals (e.g., Molenaar & Campbell, 2009; Peng et al., 2022; Wirtz & Pfenninger, 2023). In so doing, our goal is “not only to do justice to the state of affairs in the population at large, but also to the *individuals* therein” (Wirtz, 2023, p. 263, italics original).

Conclusion


Content-wise, this study addressed outstanding issues regarding both the quantitative and qualitative role of extracurricular English exposure and its temporal coincidence with language development during (pre)primary school. We found that increased extracurricular activities in English were particularly noticeable during periods of rapid development (as captured by the superimposed periods of significant change), but the effects seem temporally limited to the end stages of testing and strongly hinged on the cluster in question (i.e., learners with trajectorial similarities). These findings appear to be related to the so-called “authenticity gap” between English inside and outside of school, as clusters who perceived a discrepancy between in- and out-of-school encounters with English also showed rapid development that was characterized by increased English exposure during extracurricular activities.


Methodologically speaking, the design of this study is noteworthy among the growing body of CDST-inspired studies because of its longitudinal design (i.e., 8 years), dense data collection points (i.e., 32 time points per participant and task), and its number of learners (i.e., 45 children). As such, the study strikes a balance between more ‘traditional’ longitudinal group studies with a classical pre-post-test design (2–4 data collection points) and more recently

emerged CDST-inspired longitudinal microgenetic case studies. Additionally, instead of readily assuming a binary distinction between the groups earlyPAC and earlyPAC-int, we proceeded in a bottom-up investigative manner, taking the earlyPAC / earlyPAC-int binary into account only after the cluster analysis. This does justice to recent calls not to obscure within-group variation in lieu of focusing exclusively on between-group variance (e.g., Siegelman et al., 2023; Özsoy & Blum, 2023). Finally, we believe the dynamic integrative approaches blending quantitative and qualitative analyses – notably generalized additive modeling, time-series clustering methodology, visual methods for significance testing and qualitative content analysis – are particularly novel for CDST-guided studies and studies in SLA more generally. Combining quantitative scales with qualitative, time-dependent data allowed us to qualitatively trace introspective development across the testing period and juxtapose this with quantitative tools, resulting in a more extensive analysis of change during the experiment and thus giving us “a better chance of observing complex, emergent and qualitative changes over time” (Hiver et al., 2021, p. 11).

There are likely to be further factors that contribute to more wholly explaining the observed processes and products found in the present study. Potential variables include, e.g., cognitive resources such as working memory and explicit/implicit language learning aptitude; further data-rich studies may attempt to explore how within-person fluctuations in cognitive functioning temporally relate to periods of significant development as well as comparative stability. Moreover, while we homed in on a comparatively large sample of learners, future work may limit the number of focal learners in order to zoom in on how processes relating to, e.g., extramural English exposure impacts individual-level processes and outcomes in language development. Finally, on a more computational note, we attempted to operationalize supra-developmental patterns using three parameters from GAMs (intercept, predicted Time 0, and *edf*), which may yet underestimate individual differences in students’ learning trajectories. Additional research should thus attempt to extract a larger number of parameters from the individual-level generalized additive models (e.g., length of attractor states versus length of repeller states, intensity of phase shifts, etc.) in order to, with a more fine-tuned lens, capture learners’ individual developmental pathways. In so doing, it should be possible to more heartily address (supra-developmental) patterned outcomes often overshadowed by learner variety and learning heterogeneity and thus better reconcile generalizability, variability, and individuality.

ORCID

 <https://orcid.org/0000-0002-0433-4812>

 <https://orcid.org/0000-0002-9408-1993>

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access

This article is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which grants permission to use, share, adapt, distribute and reproduce in any medium or format provided that proper credit is given to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if any changes were made.

References

- Allport, G. W. (1937). *Personality: A psychological interpretation*. Holt, Rinehart & Winston.
- Altman, D. G. (1998). Categorizing continuous variables. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (pp. 563–567). Wiley.
- Baba, K., & Nitta, R. (2014). Phase transitions in development of writing fluency from a complex dynamic systems perspective. *Language Learning*, 64(1), 1–35. <https://doi.org/10.1111/lang.12033>
- Benson, P. (2019). Ways of seeing: The individual and the social in applied linguistics research methodologies. *Language Teaching*, 52(1), 60–70. <https://doi.org/10.1017/S0261444817000234>
- Birdsong, D. (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology*, 9(1), 1–17. <https://doi.org/10.3389/fpsyg.2018.00081>
- Birdsong, D., & Vanhove, J. (2016). Age of second language acquisition: Critical periods and social concerns. In E. Nicoladis & S. Montanari (Eds.), *Bilingualism across the lifespan: Factors moderating language proficiency* (pp. 163–181). American Psychological Association. <https://doi.org/10.1037/14939-010>
- Bulté, B., & Housen, A. (2020). A DUB-inspired case study of multidimensional L2 complexity development: Competing or connecting growers. In W. Lowie, M. Michel, M. Keijzer & R. Steinkrauss (Eds.), *Usage-based dynamics in second language development* (pp. 50–86). Multilingual Matters. <http://dx.doi.org/10.21832/9781788925259-006>
- Bushati, B., Kedia, G., Rotter, D., Christencen, A. P., Krammer, G., Corcoran, K., & Schmörlzer-Eibinger, S. (2023). Friends as a language learning resource in multilingual primary school classrooms. *Social Psychology of Education*. <https://doi.org/10.1007/s11218-023-09770-6>
- Cummings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonski (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 159–181). Routledge.
- de Bot, K. (2008). Introduction: Second language development as a dynamic process. *The Modern Language Journal*, 92(2), 166–178. <https://doi.org/10.1111/j.1540-4781.2008.00712.x>
- de Bot, K., & Larsen-Freeman, D. (2011). Researching second language development from a dynamic systems theory perspective. In M. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development: Methods and techniques* (pp. 5–23). John Benjamins.
- de Bot, K., Lowie, W. M., & Verspoor, M. H. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, 10(1), 7–21.
- de Graaff, R. (2015). Vroeg of laat Engels in het basisonderwijs; Wat levert het op? *Levende Talen Tijdschrift*, 16(2), 3–15.
- Dewaele, J. M. (2009). Individual differences in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 623–646). Emerald.
- Dörnyei, Z. (2009). Individual differences: Interplay of learner characteristics and learning environment. *Language Learning*, 59(1), 230–248.
- Dörnyei, Z. (2014). Researching complex dynamic systems: ‘Retrodictive qualitative modeling’ in the language classroom. *Language Teaching*, 47(1), 80–91. <https://doi.org/10.1017/S0261444811000516>
- Douglas Fir Group. (2016). A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal*, 100(s1), 19–47.
- Ellis, N. C. (2007). Dynamic systems and SLA: The wood and the trees. *Bilingualism: Language and Cognition*, 10(1), 23–25. <https://doi.org/10.1017/S1366728906002744>
- Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics-introduction to the special issue. *Applied Linguistics*, 27(4), 558–589. <https://doi.org/10.1093/applin/aml028>
- Festman, J. (2018). Vocabulary gains of mono- and multilingual learners in a linguistically diverse setting: Results from a German-English intervention with inclusion of home languages. *Frontiers in Communication*, 3(26), 1–15. <https://doi.org/10.3389/fcomm.2018.00026>
- Henry, A. (2013). Digital games in ELT: bridging the authenticity gap. In E. Ushioda (Ed.), *International perspectives on motivation: Language learning and professional challenges* (pp. 133–155). Palgrave MacMillan.
- Henry, A. (2021). Motivational connections in language classrooms: A research agenda. *Language Teaching*, 54(2), 221–235. <https://doi.org/10.1017/S0261444820000026>

- Henry, A., & Lamb, M. (2020). L2 motivation and digital technologies. In M. Lamb, K. Csizér, A. Henry, & S. Ryan (Eds.), *The Palgrave handbook of language learning motivation* (pp. 599–619). Palgrave Macmillan.
- Henry, A., Korp, H., Sundqvist, P., & Thorsen, C. (2018). Motivational strategies and the reframing of English: Activity design and challenges for teachers in contexts of extensive extramural encounters. *TESOL Quarterly*, 52(2), 247–273.
- Hiver, P. (2017). Tracing the signature dynamics of language teacher immunity: A retrodictive qualitative modeling study. *The Modern Language Journal*, 101(4), 669–690. <https://doi.org/10.1111/modl.12433>
- Hiver, P., & Al-Hoorie, A. H. (2020). *Research methods for complexity theory in applied linguistics*. Multilingual Matters.
- Hiver, P., Al-Hoorie, A. H., & Larsen-Freeman, D. (2021). Toward a transdisciplinary integration of research purposes and methods for complex dynamic systems theory: beyond the quantitative–qualitative divide. *International Review of Applied Linguistics in Language Teaching*, 60(1), 7–22. <https://doi.org/10.1515/iral-2021-0022>
- Hood, G. M. (2004). PopTools (Version 2.6.2.) [CSIRO].
- Jessner, U. (2008). A DST Model of multilingualism and the role of metalinguistic awareness. *The Modern Language Journal*, 92(2), 270–283. <https://doi.org/10.1111/j.1540-4781.2008.00718.x>
- Kliesch, M., & Pfenninger, S. E. (2021). Cognitive and socioaffective predictors of L2 microdevelopment in late adulthood: A longitudinal intervention study. *The Modern Language Journal*, 105(1), 237–266.
- Kliesch, M., Pfenninger, S. E., Wieling, M., Stark, E., & Meyer, M. (2022). Cognitive benefits of learning additional languages in old adulthood? Insights from an intensive longitudinal intervention study. *Applied Linguistics*, 4(4), 653–676.
- Larsen-Freeman, D. (2006). The Emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4), 590–619. <https://doi.org/10.1093/applin/aml029>
- Larsen-Freeman, D. (2012). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics? *Language Teaching*, 45(2), 202–214. <https://doi.org/10.1017/S0261444811000061>
- Larsen-Freeman, D. (2015). Complexity theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (2nd ed., pp. 227–244). Routledge.
- Larsen-Freeman, D. (2017). Complexity theory: The lessons continue. In Lourdes Ortega & Zhao Hong Han (Eds.), *Complexity theory and language development: In celebration of Diane Larsen-Freeman* (pp. 11–50) John Benjamins.
- Larsen-Freeman, D. (2018). Looking ahead: Future directions in, and future research into, second language acquisition. *Foreign Language Annals*, 51(1), 55–72. <https://doi.org/10.1111/flan.12314>
- Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford University Press.
- Leona, N. L., van Koert, M. J., van der Molen, M. W., Rispen, J. E., Tijms, J., & Snellings, P. (2021). Explaining individual differences in young English language learners' vocabulary knowledge: The role of extramural English exposure and motivation. *System*, 96(1), 1–23. <https://doi.org/10.1016/j.system.2020.102402>
- Levshina, N. (2015). *How to do linguistics with R. Data exploration and statistical analysis*. Benjamins.
- Lowie, W. (2017). Lost in space? Methodological considerations in complex dynamic theory approaches to second language development research. In L. Ortega & Z. Han (Eds.), *Complexity theory and language development in celebration of Diane Larsen-Freeman* (pp. 123–142). John Benjamins.
- Lowie, W. M., & Verspoor, M. H. (2015). Variability and variation in second language acquisition orders: A dynamic reevaluation. *Language Learning*, 65(1), 63–88.
- Lowie, W. M., & Verspoor, M. H. (2019). Individual differences and the ergodicity problem. *Language Learning*, 69(51), 184–206. <https://doi.org/10.1111/lang.12324>
- MacIntyre, P., & Mercer, S. (2014). Introducing positive psychology to SLA. *Studies in Second Language Learning and Teaching*, 4(2), 153–172. <https://doi.org/10.3758/BRM.42.2.381>
- McCarthy, P. M., & Jarvis, S. (2010). MTL-D, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2), 381–392.
- McWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum.
- Mercer, S. (2011). Understanding learner agency as a complex dynamic system. *System*, 39(4), 427–436. <https://doi.org/10.1016/j.system.2011.08.001>
- Mercer, S. (2015). Dynamics of the self: A multilevel nested systems approach. In Z. Dörnyei, P. D. MacIntyre, & A. Henry (Eds.), *Motivational dynamics in language learning* (pp. 139–163). Multilingual Matters.
- Molenaar, P. C. M. (2015). On the relation between person-oriented and subject-specific approaches. *Journal for Person-Oriented Research*, 1(1–2), 34–41.
- Molenaar, P. C. M., and C. G. Campbell. 2009. The new person-specific paradigm in psychology. *Current Directions in Psychological Science* 18(2), 112–117.
- Norris, J. M., & Manchón, R. M. (2012). Investigating L2 writing development from multiple perspectives: Issues in theory and research. In R. M. Manchón (Ed.), *L2 Writing Development: Multiple Perspectives* (pp. 221–244). de Gruyter.

- Özsoy, O., & Blum, F. (2023). Exploring individual variation in Turkish heritage speakers' complex linguistic productions: Evidence from discourse markers. *Applied Psycholinguistics*, 44(4), 534–564. <https://doi.org/10.1017/S0142716423000267>
- Pallotti, G. (2022). Cratylus' silence: On the philosophy and methodology of Complex Dynamic Systems Theory in SLA. *Second Language Research*, 38(3), 689–701. <https://doi.org/10.1177/0267658321992451>
- Peng, H., S. Jager, and W. Lowie. 2021. A person-centred approach to L2 learners' informal mobile language learning. *Computer Assisted Language Learning*, 35(9), 2148–2169. doi:10.1080/09588221.2020.1868532
- Peng, H., W. Lowie, and S. Jager (2022). Unravelling the idiosyncrasy and commonality in L2 developmental processes: A time-series clustering methodology. *Applied Linguistics*, 43(5), 891–911, <https://doi.org/10.1093/applin/amac011>
- Penris, W., & Verspoor, M. (2017). Academic writing development: A complex, dynamic process. In S. Pfenninger & J. Navracics (Eds.), *Future research directions for applied linguistics* (pp. 215–242). Multilingual Matters. <https://doi.org/10.21832/9781783097135-012>
- Pfenninger, S. E. (2020). 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. <https://doi.org/10.1111/modl.12666>
- Pfenninger, S. E. (2022). Emergent bilinguals in a digital world: A dynamic analysis of long-term L2 development in (pre)primary school children. *International Review of Applied Linguistics*, 60(1), 41–66. <https://doi.org/10.1515/iral-2021-0025>
- Pfenninger, S. E. (2021). About the INTER and the INTRA in age-related research: Evidence from a longitudinal CLIL study with dense time serial measurements. *Linguistics Vanguard*, 7(s2), 20200028. <https://doi.org/10.1515/lingvan-2020-0028>
- Pfenninger, S. E., & Kliesch, M. (2023). Variability as a functional marker of second language development in older adult learners. *Studies in Second Language Acquisition*, 45(4), 1004–1030. doi:10.1017/S0272263123000013
- Pfenninger, S. E., & Singleton, D. (2017). *Beyond age effects in instructional L2 learning. Revisiting the age factor*. Multilingual Matters.
- Pinner, R.S. (2014). The authenticity continuum: towards a definition incorporating international voices. *English Today*, 30(4), 22–27
- Roberts, C., & Cooke, M. (2009). Authenticity in the adult ESOL classroom and beyond. *TESOL Quarterly*, 43(4), 620–642.
- Seawright, J. (2016). *Multi-method social science: Combining qualitative and quantitative tools (Strategies for Social Inquiry)*. Cambridge University Press. <https://doi.org/10.1017/CBO9781316160831>
- Serafini, E. J. (2020). Further exploring the dynamicity, situatedness, and emergence of the self: The key role of context. *Studies in Second Language Learning and Teaching*, 10(1), 133-152.
- Siegelman, N., Elgort, I., Brysbaert, M., Agrawal, N., Amenta, S., Arsenijević Mijalković, J., Chang, C. S., Chernova, D., Chetail, F., Clarke, A. J. B., Content, A., Crepaldi, D., Davaabold, N., Delgersuren, S., Deutsch, A., Dibrova, V., Drieghe, D., Filipović Đurđević, D., Finch, B., ... Kuperman, V. (2023). Rethinking first language–second language similarities and differences in English proficiency: Insights from the ENGLISH Reading Online (ENRO) Project. *Language Learning*. <https://doi.org/10.1111/lang.12586>
- Simpson, G. L. (2014). *Identifying periods of change in times series with GAMs*. Retrieved October 20, 2020, from <https://fromthebottomoftheheap.net/2014/05/15/identifying-periods-of-change-with-gams/>
- Staples, S., & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 243–274). Routledge.
- Sundqvist, P., & Olin-Scheller, C. (2013). Classroom vs. extramural English: Teachers dealing with demotivation. *Language and Linguistics Compass*, 7(6), 329-338. <https://doi.org/10.1111/lnc3.12031>
- Ushioda, E. (2011). Motivating learners to speak as themselves. In G. Murray, X. Gao & T. Lamb (Eds.), *Identity, motivation and autonomy in language learning* (pp. 11–24). Multilingual Matters.
- van Dijk, M., Verspoor, M. H., & Lowie, W. M. (2011). Variability and DST. In M. H. Verspoor, K. de Bot, & W. M. Lowie (Eds.), *A dynamic approach to second language development: Methods and techniques* (pp. 55–84). John Benjamins.
- van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, 98(1), 3–53. <https://doi.org/10.1037/0033-295x.98.1.3>
- van Geert, P. (2011). The contribution of complex dynamic systems to development. *Child Development Perspectives*, 5(4), 273–278. <https://doi.org/10.1111/j.1750-8606.2011.00197.x>
- Vanhove, J. (2020). When labeling L2 users as nativelike or not, consider classification errors. *Second Language Research*, 36(4), 709–724. <https://doi.org/10.1177/0267658319827055>
- Verspoor, M. H. (2017). Complex dynamic systems theory and L2 pedagogy. In L. Ortega & Z. Han (Eds.), *Complexity Theory and Language Development* (pp. 143–162). Benjamins.

- Verspoor, M. H., de Bot, K., & Lowie, W. M. (eds.) (2011). *A dynamic approach to second language development*. Benjamins.
- Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal*, 92(2), 214–231. <https://doi.org/10.1111/j.1540-4781.2008.00715.x>
- Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21(3), 239–263. <https://doi.org/10.1016/j.jslw.2012.03.007>
- Waninge, F., Dörnyei, Z., & de Bot, K. (2014). Motivational dynamics in language learning: Change, stability, and context. *The Modern Language Journal*, 98(3), 704–723.
- Wirtz, M. A. (2023). *Inter- and intra-individual variation in adult L2 sociolinguistic repertoires. dynamics of linguistic, socioaffective and cognitive factors* [Doctoral dissertation, University of Salzburg].
- Wirtz, M., & Pfenninger, S. (2023). Variability and individual differences in L2 sociolinguistic evaluations: The GROUP, the INDIVIDUAL and the HOMOGENEOUS ENSEMBLE. *Studies in Second Language Acquisition*, 45(5), 1186-1209. <https://doi.org/10.1017/S0272263123000177>
- Wirtz, M. A., & Pfenninger, S. E. (in press). Signature dynamics of development in L2 sociolinguistic competence: evidence from an intensive micro-longitudinal study. *Language Learning*.
- Wood, S. (2006). *Generalized additive models: An introduction with R*. CRC Press.
- Yu, H., & Lowie, W. (2020). Dynamic paths of complexity and accuracy in second language speech: a longitudinal case study of Chinese learners. *Applied Linguistics*, 41(6), 855–877. <https://doi.org/10.1093/applin/amz040>

Appendix

Figure A1

Fitted Additive Model with Superimposed Periods of Significant Change in Cluster 4 Students' Aggregated Writing (BLUE Overlays) and Oral (RED Overlays) Developmental Trajectories

