Journal of Second Language Pronunciation

Assessing the state of the art in longitudinal L2 pronunciation research: Trends and future directions

--Manuscript Draft--

Manuscript Number:	JSLP-20059R2
Full Title:	Assessing the state of the art in longitudinal L2 pronunciation research: Trends and future directions
Short Title:	Assessing longitudinal L2 pronunciation research
Article Type:	Article
First Author:	Charles Nagle
Corresponding Author:	Charles Nagle Iowa State University Ames, IA UNITED STATES
Funding Information:	
Section/Category:	Review Articles
Keywords:	L2 pronunciation; longitudinal research; research synthesis; research methods
Abstract:	Longitudinal research methods often call to mind studies of various lengths. However, longitudinal research involves complex decisions related to study length, number of sessions, and session spacing, and these longitudinal choices must be coordinated with other aspects of research methodology. In this synthesis, I assess the state of the art in longitudinal L2 pronunciation research by analyzing 39 longitudinal L2 pronunciation studies published between 2006 and 2021 that did not include a pronunciation-specific intervention. I examine longitudinal design choices in light of participant sample characteristics such as age and context of learning, and measurement framework characteristics, which include choices related to target structures and tasks. Among other findings, results point to a lack of longer-term, multiwave studies dealing with pronunciation development. I offer suggestions for future work that can enhance the scope of L2 pronunciation research as well as recommendations for conducting and reporting longitudinal research.
Author Comments:	Dear Dr. Levis (Dear John), thank you for the positive feedback on the article. Kazuya's additional feedback was helpful. As before, I'll send the coding book to you via email because I can't upload it to the system. My hope is that the coding book with the studies can be published as supplementary online material accompanying the article.
Order of Authors Secondary Information:	

Assessing the state of the art in longitudinal L2 pronunciation research: Trends and future directions

Charles L. Nagle

Iowa State University

Acknowledgement:

I am indebted to Tracey Derwing and Murray Munro for their constant support and encouragement. Their longitudinal studies have served as a model for my own work on many occasions. I would also like to thank Shelby Bruun for her help on this project and Pavel Trofimovich, Tracey Derwing, and Kazuya Saito for their insightful comments on earlier versions of this manuscript. Last but not least, I would like to thank the researchers whose work I have syntehsized in this article for their contribution to the field.

Abstract

Longitudinal research methods often call to mind studies of various lengths. However, longitudinal research involves complex decisions related to study length, number of sessions, and session spacing, and these longitudinal choices must be coordinated with other aspects of research methodology. In this synthesis, I assess the state of the art in longitudinal L2 pronunciation research by analyzing 39 longitudinal L2 pronunciation studies published between 2006 and 2021 that did not include a pronunciation-specific intervention. I examine longitudinal design choices in light of participant sample characteristics such as age and context of learning, and measurement framework characteristics, which include choices related to target structures and tasks. Among other findings, results point to a lack of longer-term, multiwave studies dealing with pronunciation development. I offer suggestions for future work that can enhance the scope of L2 pronunciation research as well as recommendations for conducting and reporting longitudinal research.

Keywords: L2 pronunciation; longitudinal research; research synthesis; research methods

1. Introduction

Fifteen years ago, Ortega and Iberri-Shea observed that “many questions concerning second language learning are fundamentally questions of time and timing” (2005, p. 27). In their review of longitudinal research in second language acquisition (SLA), they called for greater consideration of the timescale and level of granularity of longitudinal research. Since their publication, “questions of time and timing” have pervaded nearly every aspect of SLA. The current state of the art, underpinned by dynamic approaches to second language (L2) development (de Bot et al., 2007), recognizes the emergent, self-organizing, and adaptive nature of L2 systems. One important aspect of this new research agenda is understanding both how L2 development unfolds over time and how different developmental processes overlap and influence one another: “We need an additional understanding of how the pieces fit together, interacting in space and time over many different levels of granularity and timescale” (Ellis, 2014, p. 398).

L2 pronunciation researchers have consistently underscored the value of a longitudinal perspective. For example, in her survey of research on exceptional pronunciation learners, Moyer highlighted the need to consider pronunciation learning (or, more precisely, accent) from a dynamic, interactive perspective, according to which “the affective and the cognitive go hand-in-hand as one seeks, and consciously utilizes, the input available” (2014a, p. 18). And at the first Pronunciation in Second Language Learning and Teaching conference, Derwing called attention to the ways in which longitudinal research can inform pedagogy: “Wouldn’t it be helpful to have some longitudinal studies to know what aspects of pronunciation will likely take care of themselves over time? Such information would allow teachers to focus on intransigent problems” (2010, p. 27).

In response to this theoretical shift, longitudinal research methods have evolved, moving away from simpler designs, in which performance is sampled over two or three data points, toward denser multi-wave studies that provide a higher-resolution view of development. As a result, there is increasing diversity in longitudinal SLA research and an increasing awareness among SLA researchers that longitudinal methods must be coordinated with the developmental processes they are intended to measure (Hiver & Al-Hoorie, 2019). Yet, longitudinal research can be daunting because although most pronunciation researchers have received training in research methods (e.g., through a research methods course in graduate school, working as a postdoc), they may not have received any specific training in longitudinal methods (Derwing & Munro, 2017). Thus, despite consistent calls for more longitudinal work, pronunciation researchers may be reticent to carry out longitudinal studies. What's more, without a firm understanding of the state of the art in longitudinal pronunciation research, researchers may not be sure where to begin both in terms of the conceptual topics that should be addressed from a longitudinal perspective and the methods needed to address them.

With this in mind, in this study I set out to synthesize longitudinal pronunciation research published after Ortega and Iberri-Shea (2005), focusing on the longitudinal characteristics of this body of work in relation to other elements of research methodology. Because pronunciation instruction has already received substantial attention in the literature (Lee et al., 2015; McAndrews, 2019; Saito & Plonsky, 2019; Thomson & Derwing, 2015), I limited the scope of this synthesis to longitudinal L2 pronunciation studies that did not include a pronunciation-specific intervention.

2. Background

2.1 Longitudinal Design Choices in Longitudinal L2 Pronunciation Research

At its simplest, longitudinal research involves collecting data from the same group of participants over time. However, this definition masks considerable complexity in longitudinal design choices: Over what observation period will data be collected? How many data points will be included, and how will they be spaced throughout the data collection interval? Will all participants begin the study at the same time, or will multiple cohorts be recruited? Should any higher-order groupings (e.g., classes) be taken into consideration? These questions make evident that many different longitudinal designs are possible, all of which lead to distinct, but complementary, views of development. Many of these designs have been attested in the L2 pronunciation literature, including long-range studies that examine L2 learners over the course of many years (e.g., Derwing & Munro, 2013) and dense multi-wave studies that measure L2 speech over many sessions within a shorter developmental window (e.g., Casillas, 2020). Such longitudinal design choices correspond to different perspectives of L2 pronunciation learning.

In their 7-year study on Mandarin and Slavic language speakers who had relocated to Canada, Derwing & Munro (2013) analyzed development over successively longer periods. Their study provided evidence of a window of maximal development during the first year of L2 residence when marked changes in pronunciation were observed. At the same time, individual differences in willingness to communicate seemed to affect L2 pronunciation learning within both learner groups. By analyzing the same groups of learners over intervals of varying length, Derwing and Munro captured both non-linear trends at the group level and individual variation in learning. Overall, then, their study underscores the unique potential of longitudinal research to render a portrait of the variables that influence rate and shape of development over an extended time frame (i.e., over years of L2 study or experience).

Longitudinal design choices, particularly the number of data points that studies include, allow for a different level of granularity (i.e., resolution) in analyzing and interpreting patterns of pronunciation development. For example, both Nagle (2019a) and Casillas (2020) investigated

English speakers' production of L2 Spanish stops. Nagle examined learners over five data points distributed at equal half-semester intervals over learners' second, third, and fourth semesters of instruction (i.e., two during the second and third semesters, and a final data point in the fourth semester), and Casillas examined learners over eight data points, one per week over an 8-week domestic immersion program. Both studies showed that learners progressed toward more accurate production of Spanish stops, but a few key differences were observed (e.g., different growth functions for voiced and voiceless stops) due to different longitudinal approaches. Casillas summarized the importance of data density: "Having more experimental sessions over a shorter period of time allowed the present study, in a sense, to zoom in on the learning process" (2020, p. 24). Indeed, if pronunciation development is conceptualized in terms of nested developmental periods that influence one another, then manipulating longitudinal characteristics while holding other methodological features constant can shed light on prototypical developmental cycles. Perhaps initial conditions in rate and shape of change set the stage for long-term pronunciation learning.

Longitudinal research is also ideally positioned to reveal the interplay between learning context and individual differences in L2 pronunciation development. Here too, longitudinal design choices matter. For instance, Saito et al. (2018) investigated the relationship between motivation, emotion, and experience and L2 comprehensibility over two data points encompassing 108 L2 English learners' third term of instruction. Nagle (2018) also investigated the relationship between motivation and L2 comprehensibility (and accentedness) in a sample of 25 L2 Spanish learners, but his study included five data points during learners' second, third, and fourth semesters of instruction. Saito et al. addressed the relationship between individual differences and comprehensibility gain scores, whereas Nagle tracked individual differences in motivation dynamically, investigating changes in comprehensibility and accentedness in relation to changes in motivation. Thus, two longitudinal studies that seem similar may in fact yield different insights depending on their longitudinal characteristics. Saito et al.’s study involves questions related to predictors of pronunciation achievement over a semester of instruction, whereas Nagle’s study affords insight into the extent to which changes in learner psychology predict changes in pronunciation over time. These two studies also demonstrate the tension that sometimes arises when coordinating longitudinal data collection decisions with other elements of methodology. Saito et al. achieved a much larger sample size than Nagle, enhancing the generalizability of findings, but Nagle achieved a more time-sensitive view of development by studying a smaller number of learners over more data points.

One last study that deserves attention is Saito et al. (2019), which examined both how learners interacted over time and the extent to which they benefitted from video-based interaction. In this way, the authors scrutinized learning process (the types of interactional moves that learners made over time) and product (gains resulting from the interaction) from a longitudinal perspective. Overall, then, this body of longitudinal work highlights the diverse designs that can be leveraged to provide insight into many different aspects of pronunciation learning.

2.2 Methodological Choices in Longitudinal L2 Pronunciation Research

There is no doubt that all aspects of quantitative research methodology have received increased attention in recent years (e.g., Gass et al., 2020). Two areas are particularly germane to L2 pronunciation research: participant sampling practices and measurement frameworks (Saito & Plonsky, 2019). There is a growing concern over sampling bias in SLA research. As Andringa and Godfroid (2020) have pointed out, SLA researchers routinely generalize their findings to the general population despite the fact that study samples are not representative. Researchers often recruit university students, a form of convenience sampling, and there is a tendency to conduct research in Western, Educated, Industrialized, Rich, and Democratic, or WEIRD, contexts (Henrich et al., 2010). These trends are also reflected in longitudinal L2 pronunciation research, where college-aged samples predominate. Moreover, there seems to be a noticeable bias toward studies involving English (Nagle et al., 2019), precluding a panoramic understanding of how pronunciation develops in linguistic, social, and cultural contexts where English is not the native or target language. Beyond demographic variables and language pairings, participant sampling practices also encompass the learning context, which typically refers to the distinction between instructed or foreign language (FL) learning and naturalistic or second language (SL) learning. By understanding the intersection between sample characteristics and longitudinal design choices, we can design more representative studies that address pronunciation learning throughout the lifespan in diverse contexts of learning.

A second area of special interest for longitudinal L2 pronunciation research is the measurement framework, which encompasses the target structure, data collection tasks, and coding and measurement schemes. Saito and Plonsky (2019) have differentiated between specific pronunciation constructs, such as individual segmental and suprasegmental features, and global pronunciation constructs, such as intelligibility and comprehensibility, which together provide complementary information on L2 speakers' pronunciation proficiency (see also Munro & Derwing, 2015, for discussion of local vs. global intelligibility). They have also drawn a distinction between controlled and spontaneous production tasks (e.g., word reading and picture description, respectively), arguing that performance on these tasks reflects controlled versus spontaneous pronunciation knowledge. By using this framework to understand the focus of current longitudinal L2 pronunciation research, future work will be well-positioned to address questions related to how controlled and spontaneous pronunciation knowledge develop, leading to a richer view of L2 pronunciation learning.

2.3 The Current Research Synthesis

There has been considerable growth in longitudinal L2 pronunciation research in recent years. As a field, we are rapidly moving toward greater variety in longitudinal designs that yield insight into pronunciation development as a dynamic process, instantiated over many different time frames and levels of granularity. Beyond variation in longitudinal methods, there is also substantial variation in other aspects of methodology, including participant samples, measurement frameworks, and analytical approaches. Even though these dimensions of research methodology can be evaluated independently, the reality is that they do not exist in a vacuum. Rather, they influence one another, and their collective implementation dictates perspectives on the nature of L2 pronunciation learning. Thus, by examining the intersection between longitudinal choices and other aspects of research methodology, we can arrive at an understanding of the current state of the art in longitudinal L2 pronunciation research. In this research synthesis, I therefore set out to address the following questions:

1. What general trends in longitudinal characteristics, participant samples, measurement practices, and statistical approaches are evident in longitudinal L2 pronunciation research published between 2006 and 2021?
2. What conceptual and methodological gaps in the literature exist when longitudinal characteristics are considered in light of participant samples and measurement practices?

3. Method

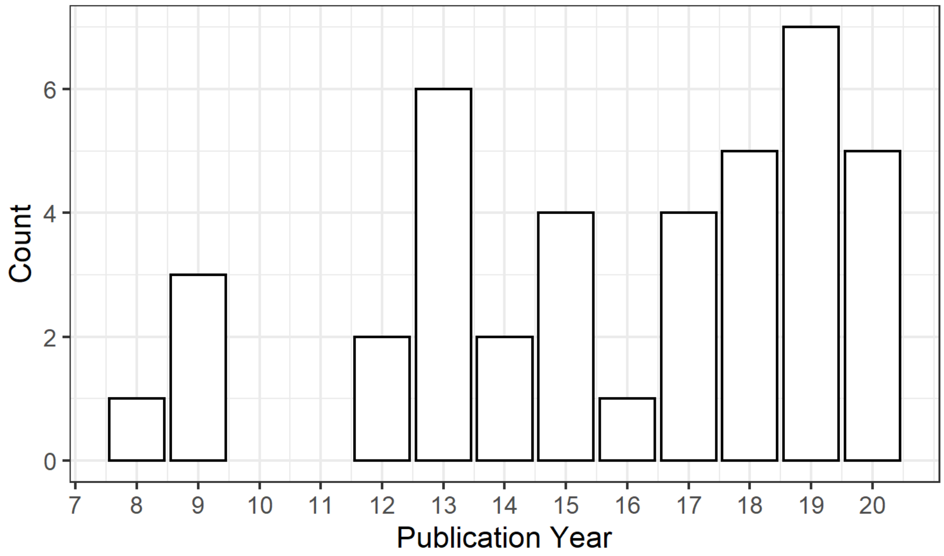
3.1 Study Retrieval and Publication Trends

With the help of a research assistant, I searched major databases such as Linguistics and Language and Behavior abstracts, 17 prominent journals known to publish L2 pronunciation scholarship (Appendix A), and the proceedings of the Pronunciation in Second Language Learning and Teaching conference using a combination of longitudinal keywords (e.g., *longitudinal*, *time*, *development*), pronunciation keywords targeting specific features (e.g., *segmentals*, *vowels*, *intonation*), pronunciation keywords targeting global features (e.g., *comprehensibility*, *accentedness*), and general pronunciation terminology (e.g., *pronunciation*, *speech*, *phonetics*). To be included in the review, studies had to focus on any aspect of L2 pronunciation and had to examine data collected from the same group of L2 participants over two or more data points. Following these criteria, we identified and retrieved over 100 studies published between 2006 and May 2021. In this synthesis, I focus on 39 studies published in peer-reviewed scholarly journals and conference proceedings that did not involve a pronunciation-specific intervention.[1]

To provide some context on this body of literature, I first examined publication trends. As shown in Figure 1, longitudinal L2 pronunciation studies have increased since Ortega & Iberri-Shea (2005), especially over the past few years. Publication topics represent a range of disciplinary foci: L2 learning and its relationship to L1 phonetic drift (e.g., Chang, 2012, 2019), the perception-production link in L2 speech learning (e.g., Casillas, 2019; Nagle, 2018a, 2020), the psychosocial and cognitive factors that shape L2 pronunciation development (e.g., Nagle, 2018b; Saito et al., 2018; Saito et al., 2019), naturalistic L2 pronunciation development (e.g., Derwing & Munro, 2013; Munro & Derwing, 2008; Zielinski & Pryor, 2020), and instructional approaches that can help learners achieve more comprehensible L2 speech (e.g., Akiyama & Saito, 2016; Saito & Akiyama, 2017; Trofimovich et al., 2009).

Figure 1

Longitudinal L2 Pronunciation Publications by Year



3.2 Coding

I coded studies for the following longitudinal features: total length in months, number of sessions (i.e., number of data points), and data spacing, a binary variable that reflects whether data was collected at equal or irregular intervals. I also created a categorical length variable to allow for meaningful aggregation and comparison of studies of similar length. In practice, this was challenging because many studies report length on somewhat ambiguous study-specific timescales. For example, studies conducted in an academic setting often use trimesters, semesters, or academic years as developmental units. I ultimately created five bins: 0–4, 4–8, 8–12, 12–24, and 24+ months. These particular cutoffs were advantageous because (1) they align with the academic timescales listed above, where one semester would correspond to three or four months (i.e., the 0–4 month bin) and one academic year to approximately 9 months (e.g., the 8–12 month bin); and (2) they represent different developmental intervals that can be coordinated to investigate the window of maximal opportunity for L2 pronunciation learning, which generally coincides with the first year of intensive L2 exposure (Derwing & Munro, 2015). In sorting studies into bins, I always favored the shorter bin (e.g., I sorted a study examining learners over one year, or 12 months, into the 8–12 month bin rather than the 12–24 month bin).

I also coded studies for a range of methodological features, including the following:

1. Speaker age, which I converted into three bins corresponding to children (0–9 years), adolescents (10–17 years), and adults (18+ years).
2. Context of learning, three levels as follows:
    a. FL refers to participants who were learning the L2 in a classroom setting
    b. Second language study abroad (SL SA) refers to participants who were studying in an L2 environment but intended to return to their native country (i.e., FL learners who had decided to study abroad as part of a university-sponsored program)
    c. Second language (SL) refers to participants who had relocated to an L2 environment on a more permanent basis
3. Speaker L1 and L2
4. Speaker sample size
5. Measurement framework, following Saito and Plonsky's (2019) categories with the addition of a perception category to account for the few studies examining L2 perceptual learning:
    a. Type of outcome: Perception, production of specific features, production of global features
    b. Task type: Perception, controlled production, spontaneous production

I generated multiple entries for publications including more than one level of any of the task features. For instance, Saito et al. (2019) analyzed the production of specific features (e.g., segmental errors, syllable errors) and comprehensibility, a global feature. I therefore created two entries to account for those two feature types. I also generated multiple entries for publications involving more than one L1 group (e.g., Derwing & Munro, 2013) or more than one age group or context of learning (e.g., Muñoz & Llanes, 2014). Finally, Sturm (2019) analyzed the longitudinal development of two distinct learner groups on two timescales in a single publication, so I counted those analyses separately. For these reasons, the total number of studies for any given tabulation may exceed 39 (e.g., studies reporting on more than one age group and/or context of learning would be counted multiple times when considering sample characteristics, studies reporting on more than one type of pronunciation feature would be counted multiple times when considering measurement framework characteristics). Once I had developed the initial coding scheme, a research assistant double-coded all entries. We then discussed and resolved any discrepancies before data aggregation and analysis. For a description of coded features, see Appendix B. The data set with coded entries for all 39 studies is available in the supplementary online materials.

4. General Characteristics

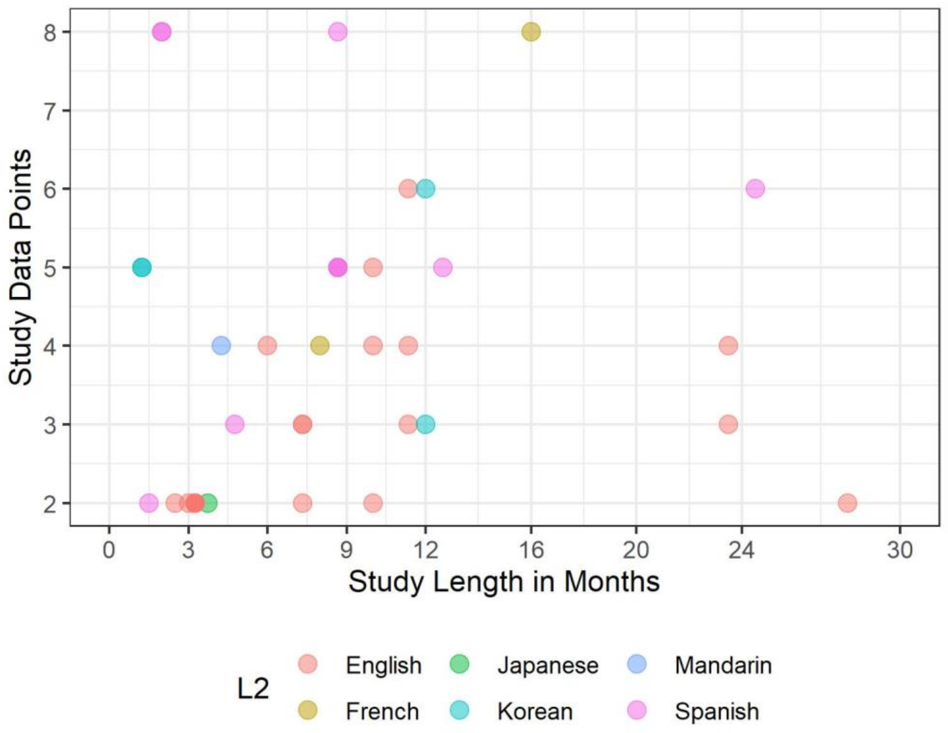
4.1 Longitudinal Characteristics

Studies that self-identify as longitudinal either in the title or keywords can provide insight into researchers' working definition of what counts as longitudinal in L2 pronunciation research. This body of work shows a trend toward a shorter window of analysis and relatively few data points. Of 30 "longitudinal" studies, five tracked learners for more than one year, and 14 included more than three data points. This suggests that the standard of "longitudinal" research is still relatively narrow and does not reflect the longer-term and/or dynamic sampling approach that scholars have increasingly called for.

An analysis of the entire set of studies confirms this definition. Excluding Derwing and Munro's (2013) 7-year (84-month) study, the mean study length was approximately 11 months (*M* = 10.71). Thirty-four studies examined learners over a year or less (0–4 months, *n* = 15; 4–8 months, *n* = 9; 8–12 months, *n* = 10), compared to three studies that analyzed learners over 12–24 months and three that did so over an even longer period (> 24 months). Longitudinal research is as much about sampling frequency and sampling intervals as it is about total length. Here too there is a trend toward fewer data points. Twenty studies analyzed data over two or three data points versus 20 studies that included four or more. As expected, as the number of data points increases, the number of studies at that data density tends to decrease. It is important to bear in mind that studies that examined learners over the same time frame and number of data points may in fact be a series of studies on the same group of learners (e.g., Casillas, 2019, 2020). Regarding spacing, most longitudinal L2 pronunciation research has sampled learners at equal intervals rather than staggering sessions. Figure 2 plots study length against study data points for various L2s. As the figure illustrates, we have not yet achieved a layered view of L2 development on multiple timescales, which would arguably require both a larger set of studies of varying lengths (especially longer-range studies) and studies of similar length but varying sampling frequencies or data densities.

Figure 2

Study Length by Study Data Points



Note. Dots have been staggered along the x-axis to prevent overlap. Darker dots correspond to multiple studies that fall into the same approximate length/data point bin (e.g., L2 English studies that included two data points over a three-month interval). Derwing and Munro's (2013) 7-year (84-month) study has been excluded to facilitate plotting.

4.2 Sample Characteristics

The 39 studies reviewed here include 1,144 speakers. Re-analysis of the same data is common in longitudinal research. However, in some cases, it is not always clear if different research reports analyze the exact same group of participants, a subset of that group, or a similar group drawn from the same learner population, which makes it difficult to determine the total number of unique participants across this set of studies.

Thirty-nine studies reported findings for adult learners, most of whom were university students, whereas studies examining L2 pronunciation development in younger learners have been rare (*n* = 4). With respect to context of learning, studies targeting FL learners are the most common (*n* = 20), followed by SL (*n* = 14) and SL SA (*n* = 9). Notably, there have been three studies that reported on multiple learner context and/or age groups (Baker-Smemoe & Haslam, 2013; Kim, Clayards, & Goad, 2018; Muñoz & Llanes, 2014). Of these, Muñoz and Llanes (2014) merits special mention for its four-way comparison. Most studies have focused on L2 English (*n* = 22), but there is a sizable body of work on L2 Spanish (*n* = 10). Other L2s are not well represented, with four longitudinal studies on L2 Korean, two on French (considering the two distinct longitudinal analyses reported in Sturm, 2019), and only one each on L2 Japanese and L2 Mandarin.

4.3 Measurement and Analytical Characteristics

Measurement and analytical characteristics refer to the tasks and outcomes that researchers choose to examine, as summarized in Saito and Plonsky's (2019) measurement framework, and the conceptual and statistical choices that researchers make when analyzing and presenting results. Regarding the measurement framework, most longitudinal L2 pronunciation research has focused on production; only five studies examined longitudinal changes in L2 perception. Of the production-focused studies, 31 reported on specific features, 12 on global features, and five on both (Akiyama & Saito, 2016; Derwing et al., 2009; Saito & Akiyama, 2017; Saito et al., 2019; Trofimovich et al., 2009). Studies examining both types of features are noteworthy because they have the potential to shed light on how changes in specific features relate to (and potentially underpin) changes in global, listener-based dimensions of pronunciation (e.g., Akiyama & Saito, 2016; Derwing et al., 2009), thereby providing a nuanced, interconnected view of development across multiple levels of analysis.

Target construct must also be considered in terms of task type (e.g., reading, picture description, story narration), given that specific and global constructs can be examined in both controlled and spontaneous speech. Table 1 provides a summary of this comparison. As shown, controlled tasks have typically been used to assess the production of specific features (e.g., individual segments such as L2 vowels), though a small number of studies have also evaluated global features (e.g., comprehensibility) in controlled speech. On the other hand, studies using spontaneous tasks have examined both global and specific features. Surprisingly, few longitudinal reports have compared production in controlled and spontaneous speech (but see, e.g., Saito, 2019), though some researchers have reported such analyses in separate publications (e.g., Derwing et al., 2009; Munro & Derwing, 2008).

Table 1

Summary of Studies According to Target Construct and Task Type.

	Controlled Tasks	Spontaneous Tasks
Global Features	4	8
Specific Features	21	14

The conceptual and statistical choices that researchers make also matter. One of the advantages of longitudinal research is that it allows for a comprehensive treatment of both group and individual data. At a basic level, then, it can speak to individual differences in performance over time. Group-level analyses provide information on mean performance (i.e., how an average individual develops over time), whereas an examination of individual cases can provide insight into the range of observed outcomes. This in turn can stimulate interesting follow-up studies to discover what factors predict differences in rate and shape of development. About half of the longitudinal studies reviewed here presented data on individual learners. Most presented this data in tables and figures to complement group-level statistics, but a few publications specifically focused on individual cases (Holliday, 2015; Sturm, 2019; Zielinski & Pryor, 2020). Worth noting is that some recent studies have included individual-level information in appendices (Casillas, 2020; Kim et al., 2018), which reinforces the fact that the manuscript itself need not be the sole means of content-delivery for data-rich longitudinal research.

Related to this topic is the statistical tests that researchers use. Longitudinal studies tend to generate complex (i.e., dense, multivariate, hierarchical) data sets that can be challenging to analyze. Typically, researchers have relied on ANOVA to quantify change over time and on correlation and regression to examine relationships between individual differences and pronunciation development. These techniques have their limitations. For one, they cannot account for nested data structures or deal with multiple sources of variance (e.g., speakers, items), and while they are effective at modeling linear relationships, they are less reliable at estimating curvilinear patterns. They also assume static predictors, even though many learner differences are inherently time-varying. Mixed-effects modeling is ideally suited to longitudinal data analysis (e.g., Cunnings & Finlayson, 2015). Unlike more traditional statistical tests, mixed-effects models can account for nested data structures while simultaneously estimating variance in multiple facets of the data. Mixed-effects models also offer researchers a means of examining curvilinearity by specifying higher-order polynomials (e.g., a quadratic trend). Perhaps most importantly, mixed-effects models can be fit to unbalanced data sets, which are common in longitudinal work (e.g., due to attrition). Although mixed-effects modeling is common in phonetics research (e.g., Chang, 2019; Kim et al., 2018), SLA pronunciation researchers have yet to fully embrace it.

4.4 Interim Summary

Longitudinal L2 pronunciation research published between 2006 and January 2021 has been characterized by the following trends: shorter designs and fewer data points, a focus on adult (i.e., university) learners of L2 English, an emphasis on controlled tasks, variable reporting of individual data, and a tendency to use analyses that do not respect the hierarchical structure of L2 pronunciation data. In observing these trends, I do not mean to criticize any particular research agenda. Research agendas depend on a variety of factors: interest, funding, geography, professional context, and so on. These trends do, however, create systematic methodological and conceptual gaps in the literature whose collective impact produces a can't-see-the-forest-for-the-trees scenario. Thus, in the following sections, I analyze longitudinal design features in light of two facets of research methodology: participant sample characteristics and measurement framework characteristics. In each area, I analyze both study length and data points to provide a nuanced portrait of the work that has been completed on various learner groups and outcome measures. This analysis provides insight into the types of longitudinal studies that can enhance the scope and generalizability of L2 pronunciation research.

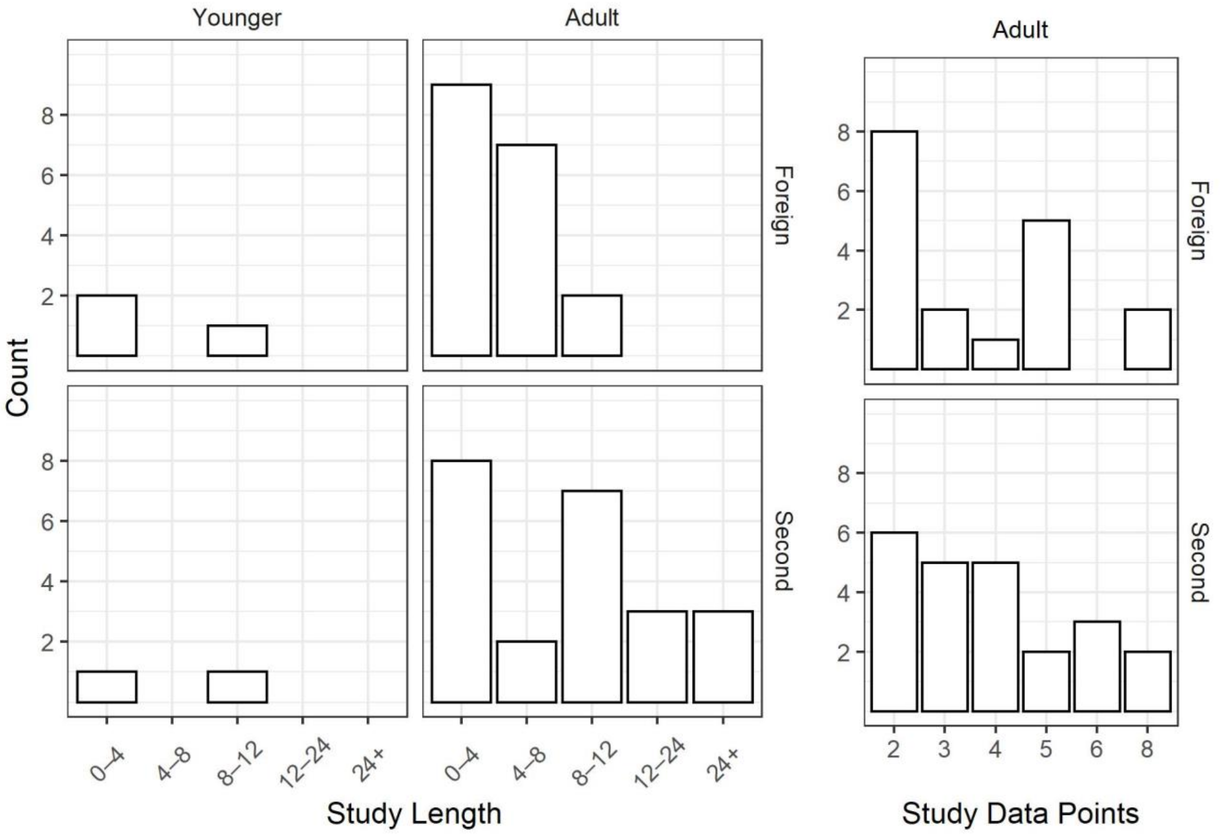
4.5 Longitudinal Design by Participant Sample

Participant sample characteristics include variables such as age, context of learning, and L1-L2 pairings. If one of the basic goals of longitudinal L2 pronunciation research is to understand pronunciation learning as a lifelong endeavor, then as a field we must be concerned with examining L2 speakers of varying ages and proficiency, including individuals in both SL and FL contexts. As shown in Figure 3, since 2006 there have been virtually no longitudinal pronunciation publications dealing with younger L2 learners. An emphasis on adult SLA could be due to a combination of convenience sampling, given that the university students sampled in many studies are comparatively easy to recruit, and practical issues that researchers must surmount (e.g., obtaining consent from minors can be difficult). At the same time, studying learners of varying ages longitudinally is necessary because, as Moyer (2014b) has pointed out, age covaries with a range of social, psychological, and experiential variables known to have an impact on L2 development. Thus, by studying learners at different ages, we can gain more precise insight into how such variables shape L2 pronunciation learning (see, e.g., Muñoz & Llanes, 2014). Here, collaboration is key. Many researchers study L2 development in children, but they may not be interested in pronunciation. Pronunciation scholars could seek to collaborate with these individuals to study younger learners.

Figure 3 also demonstrates that while SL research has examined pronunciation learning over a range of developmental windows, FL research is more circumscribed; to the best of my knowledge, no FL study has investigated pronunciation learning for more than one year. This means that we do not yet have a solid understanding of FL pronunciation learning outcomes across the four-year college curriculum, nor do we understand the variables that predict long-term learning success. Numerous longitudinal FL studies suggest that individual differences in motivation and aptitude predict learning gains over one or two semesters of language instruction (Nagle, 2018; Saito et al., 2018; Saito et al., 2020), but longer-term longitudinal studies hold the key to understanding how the influence of these variables waxes and wanes over time. The fact that a variable is associated with learning over a certain period does not mean that that relationship will persist over time. In other words, a variable that predicts learning at one point in time may not predict learning at another, and even if it does, its predictive value may change. Longitudinal studies can shed light on these time-varying relationships, especially when observation windows encompass transition periods. Transition periods are important because such periods are generally associated with changes in a range of learning-relevant variables, including quantity and quality of language use, instruction, motivation, and so on. Thus, it is during these periods when we might expect to observe changes in relationships between individual differences and pronunciation learning. Simply put, transition periods, irrespective of their source or nature, represent potential windows of maximal opportunity, making them particularly interesting targets for longitudinal work (Ortega & Iberri-Shea, 2005). At the same time, longitudinal researchers must be aware of a self-selection bias, given that more motivated individuals are more likely to continue to study the L2, especially at more advanced levels.

Figure 3

Longitudinal Characteristics by Participant Sample Characteristics



Note. The second language category includes studies addressing both second language (SL) and second language study abroad learners (SL SA) for ease of presentation.

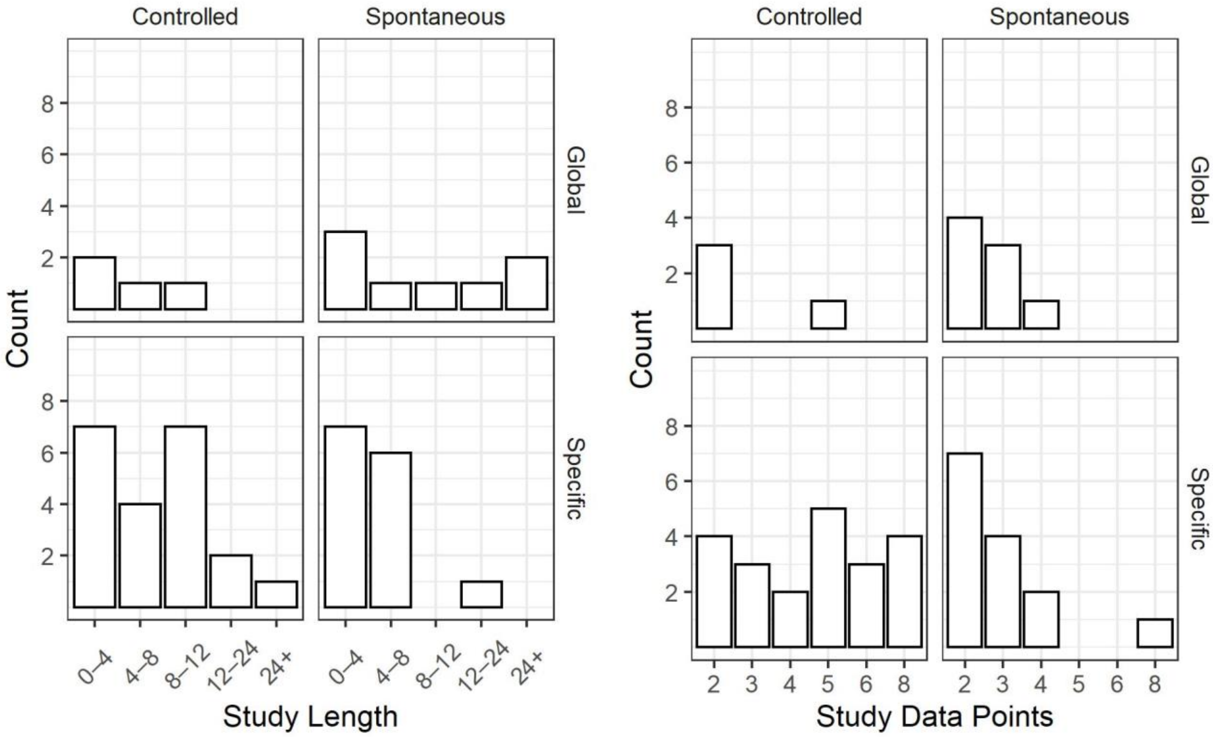
4.6 Longitudinal Design by Measurement Framework

Examining the intersection of longitudinal properties and measurement framework characteristics (i.e., target constructs and task types) can provide insight into current knowledge regarding the development of controlled and spontaneous pronunciation knowledge. Figure 4 plots the current state of longitudinal L2 pronunciation research taking these features into account. Studies that fall into each quadrant address a slightly different topic. On the one hand, the lower left quadrant (Controlled × Specific) encompasses research investigating the development of controlled pronunciation knowledge, or L2 speakers' ability to produce specific pronunciation features on tasks that allow them to focus primarily on their pronunciation. On the other hand, the lower right quadrant (Spontaneous × Specific) represents studies that examine the longitudinal production of specific features in spontaneous speech. These studies can therefore speak to how L2 speakers develop the ability to use specific features (e.g., L2 vowels, consonants, intonation) efficiently on tasks that simulate more realistic speaking conditions. A similar distinction can be made for the upper quadrants: research on global pronunciation development (e.g., their intelligibility, comprehensibility) on tasks that are less (Controlled × Global) vs. more (Spontaneous × Global) controlled/cognitively demanding.

The lower left quadrants (Controlled × Specific) of both the length and data point panels show that the field has accumulated a reasonable spread of studies addressing speakers' controlled pronunciation knowledge on various timescales. The upper right quadrants (Spontaneous × Global) indicate that while there have been studies addressing speakers' global pronunciation proficiency on spontaneous tasks (left panel), these studies have consistently sampled learners over 2–3 data points (right panel), preventing a more nuanced portrait of development at various levels of granularity. The bottom right quadrants reveal a notable absence of studies examining the production of specific features under spontaneous conditions over an extended period (e.g., more than one year). It bears repeating that the target constructs and tasks analyzed in this section necessarily reflect a bias toward adult learners. Thus, virtually any longitudinal pronunciation study addressing younger learners, irrespective of the target structures examined and tasks used, would be a welcome contribution to the field, especially if such studies are designed to capture meaningful transition points.

Figure 4

Longitudinal Characteristics by Measurement Framework Characteristics



Note. Study length is shown in the left panel and study data points in the right panel. Columns refer to task type and rows to construct type following Saito and Plonsky's (2019) framework.

5. Future Directions in Longitudinal L2 Pronunciation Research

Although longitudinal research often evokes a relatively simple pre-post(-delayed) design, longitudinal methods involve complex decision-making related to observation intervals, data points, and data spacing. It comes as no surprise, then, that the term “longitudinal” does not adequately represent this complexity. Given the observed trends toward shorter-term, lower-resolution (i.e., few data points) studies, it seems that the field would benefit from moving toward a new multi-wave standard in longitudinal research. At the same time, researchers should take care to coordinate study length and granularity with the developmental processes they intend to measure, which would include considering potential turning points and transitions brought on by changes in the learning environment, learner psychology, and so on. To that point, it is worth mentioning that dynamic and/or multiwave analyses can also be used to examine behavioral changes on a much shorter timescale (see, e.g., Nagle et al. 2019; Trofimovich et al., 2020).

Longitudinal research is about time, but time is not a simple construct. In a literal sense, learners need time to encode, consolidate, and automatize pronunciation-relevant skills. For instance, sleep appears to play an especially important role in phonetic learning (Earle & Myers, 2014). But time also has another level of interpretation, insofar as the passage of time encodes information about a range of time- and context-sensitive variables that shape development. That is, learners’ patterns of language use, including how often, with whom, and in what contexts they use the L2, inevitably change over time, as does their motivation. These patterns have implications for development. The view that cognitive, affective, and experiential differences form part of a complex, interactive, and time-varying system that catalyzes or constrains pronunciation learning is not new (Moyer, 2014a), but there has not been much research into these interactive and time-varying relationships. Longitudinal studies offer a variety of tools to study such relationships at different levels of resolution. For example, quantitative approaches can shed light on the extent to which changes in motivation, effort, and other individual differences are associated with changes in pronunciation (Nagle, 2018b), and qualitative approaches can provide nuanced insight into dynamic relationships that are not easily captured using traditional quantitative analyses (Zielinski & Pryor, 2020). Mixed-methods longitudinal research perhaps represents a gold standard because it has the potential to generate statistically generalizable findings while simultaneously highlighting the complex dynamics that underlie development (Derwing et al., 2008). Indeed, these considerations are broadly applicable to SLA research at large, where researchers have noted a tension between quantitative methods, which tend to be variable-oriented, and qualitative methods, which tend to be associated with a more ethnographic, or person-centric, approach (Ushioda, 2009). Overall, then, many of the questions Ortega and Iberri-Shea (2005) raised regarding timescales, turning points, and level of granularity still apply today.

6. Recommendations for Conducting and Reporting Longitudinal Work

Longitudinal research is a complex undertaking. Understandably, then, researchers may be reticent to engage in longitudinal work. At the 2017 Pronunciation in Second Language Learning and Teaching conference, Derwing and Munro offered a workshop on longitudinal research. As part of the workshop, participants were asked to complete a survey on longitudinal methods. They shared the following concerns: difficulty planning longitudinal studies, participant attrition, uninteresting results after expending much time and energy, challenging data management, the possibility of needing to make changes to the design while the study is ongoing, and time to publication. Having conducted a number of longitudinal studies and consulted with other researchers engaged in longitudinal work, I would like to address these points.

Concerns about planning a longitudinal study can be boiled down to three interrelated issues: determining the time period over which data will be collected and the number of data points, speakers, and items/listeners to include in the study. At a conceptual level, the length of the observation window and the number of data points are partly fixed by the anticipated developmental timeline of the target structure and the research questions. For instance, if researchers are interested in examining curvilinear trends, then an appropriate number of data points (e.g., 4–6, depending on the complexity of the expected trajectory) would be necessary, and if those same researchers are also interested in capturing transition points, such as before, during, and after study abroad (e.g., Huensch & Tracy-Ventura, 2017), then the observation window would need to encompass those transitions. At a practical level, quantitative researchers are often concerned with the issue of statistical power, or the probability of detecting an effect when one is present. A complete treatment of longitudinal design choices and their relationship to statistical power is far beyond the scope of this paper, but a few basic recommendations can be made. The most general recommendation is to consider where variance in the data is likely to occur and increase sample size in that facet. Thus, for instance, if substantial variance is expected between participants, then it would be advantageous to increase the participant sample size, and if substantial variance is expected in items (e.g., if participants are asked to produce tense and lax vowels in a range of phonetic environments in words of varying frequency and familiarity), then it would be advantageous to increase the number of items included in the study. Researchers are probably accustomed to planning participant sample sizes, but they may not be as accustomed to considering the relationship between item/listener sample sizes and statistical power (for an overview, see Westfall et al., 2014). A similar recommendation can be made for longitudinal design choices. If researchers expect more within-subjects variance (e.g., as might be the case when examining curvilinear change), then it would be beneficial to increase the number of data points included in the study (although rather technical, for commentary see Raudenbush & Xiao-Feng, 2001). Of course, the anticipated effect size also affects power, which means that researchers should familiarize themselves with common effect sizes for L2 research (Plonsky & Oswald, 2014), including L2 pronunciation research (e.g., Lee et al., 2015). Generally, there is an inverse relationship between the effect size and the required sample size to detect it; large effects can be detected in smaller samples, whereas smaller effects can only be reliably detected in large samples. Statistical power aside, there is the practical need to consider the time and cost associated with increasing the number of participants, items, or data sessions. Researchers must also deal with the inevitable reality of participant attrition, which means recruiting a (e.g., 20–25%) larger initial sample to maintain sample size over the course of the study.

Overall, it seems fair to say that longitudinal pronunciation research is still in its infancy (or at least, early adolescence). Researchers can avail themselves of user-friendly tutorials to help estimate sample size (e.g., Norouzian, 2020), while recognizing that meaningful work can be carried out with small sample sizes. As Loewen and Hui (2021) and have observed, small sample sizes are common in many areas of SLA, but small sample sizes are not problematic in and of themselves as long as researchers take a measured approach. Loewen and Hui also remind us that collaborative methods such as multisite designs and replication research, which are a necessary endeavor in any discipline, can help shore up gaps in knowledge, leading to more reliable and generalizable findings over time. Such methods may prove especially useful for researchers interested in carrying out longitudinal studies. Indeed, as the open science movement has gained traction, SLA researchers have increasingly recognized the value of developing, analyzing, and sharing complex (longitudinal) data sets (see, e.g., Saito et al., 2020).[2] What's more, technological advances in data collection, including the use of online data collection platforms such as Amazon Mechanical Turk (Nagle, 2019b; Nagle & Rehman, under review) and Gorilla Experiment Builder (Anwyl-Irvine et al., 2020), are expanding the type and number of participants that researchers can reach. When leveraged appropriately, online platforms provide an efficient means of collecting data from a large pool of participants, though whether such tools can be used for longitudinal data collection (and if so, what best practice would be) remains an open question. In short then, longitudinal research, like any type of research, requires planning. Researchers should take care to engage in reasonable quality control/assurance measures, some of which are unique to longitudinal designs (e.g., participant attrition), but they should not let concern over study planning discourage them from engaging in longitudinal research.

Related to planning are concerns regarding uninteresting results, the need to make changes to the design while the study is underway, and time to publication. Many researchers seem to take a modular view of longitudinal studies, conceptualizing them in terms of a fixed and finite set of a priori research questions to be addressed. Certainly, longitudinal researchers should specify questions during the planning stage, but in conducting longitudinal work it is important to allow room for exploratory and follow-up analyses. As data becomes available for processing and coding, unforeseen patterns may come to light that are worthy of exploration. Such analyses need not raise concerns related to data fishing or hypothesizing after the results are known (HARKing) as long as researchers clearly distinguish between planned and exploratory analyses in scholarly products. In fact, many journals now ask researchers to distinguish between these two types of analyses when reporting findings. Follow-up analyses are also possible. For example, researchers who have carried out a 1-2 year study may decide to follow up with the same participants after a much longer period (e.g., Derwing and Munro's work on L1 Mandarin and Slavic language speakers who immigrated to Canada).

Researchers may realize that they need to make a change to the design after data collection has begun, which can be particularly anxiety-inducing in longitudinal research. Yet, there are ways to strategically collect longitudinal data that allow room for necessary corrections to take place and, indeed, room for innovation. For example, rather than collecting data from a large cohort of (e.g., 100) participants longitudinally, data from multiple cohorts can be collected gradually (20 one year, 20 the next, and so on) until the desired sample size is achieved. In this way, changes can be implemented without compromising the integrity of the design (data from 40 participants follows design A, whereas data from 60 follows design B). It has been my experience that my research interests often change slightly over the course of the study, and if my interests do not change, the way I think about the research does. In other words, over time, as I have published various articles based on a longitudinal project, my understanding of the data has grown, leading me to revise and refine my approach. It is my view that this reflective process, which is a necessary component of processing, analyzing, and disseminating longitudinal research, results in superior research products. As long as researchers clearly link publications dealing with the same data set and/or learner sample, this growth process highlights the ways in which the research project itself has evolved over time. Simply put, in longitudinal work, researchers must recognize and be open to the fact that their perspective will inevitably change over time.

One last concern that deserves attention is time to publication. By processing and analyzing data as they become available, researchers should not have any issue publishing parts of the project in a timely manner. That is, researchers need not (and should not) wait until the research is finished to begin publishing. What’s more, replication has become an important topic in the social sciences, including in SLA. Longitudinal research can play a special role in allowing for robust replications. For instance, if researchers collect data from two longitudinal cohorts, they can build and evaluate a model for one cohort and then fit the same model to the other. Such an analysis can provide a particularly robust check on longitudinal findings and is not out of reach when data collection is staggered (i.e., 30 participants in year one, 30 in year two, and so on). Overall, although longitudinal research can be daunting, especially long-term and/or multi-wave research, we must commit ourselves to such an approach if we are interested in studying development, and we must strive to convey the value of such an approach to reviewers, administrators, and funding agencies.

Finally, I would like to provide a few specific recommendations for reporting longitudinal work. First, longitudinal research clearly requires careful consideration of the time frame over which data will be collected, how often it will be collected, and whether it will be collected over equally spaced intervals. Although all longitudinal publications report this information, it is often on study-specific scales. For example, studies examining FL learners might conceptualize and measure development by semesters or academic years. Because these time frames are dictated by the internal logic of the study itself, it is not always easy to define the length of any given study in unambiguous terms. Moreover, even studies that seem similar may operate on slightly different time scales. A semester at one institution may not be as long as a semester at another. As a field, we should continue to report questions of time and timing on study-specific scales, but we should also include a standard metric, such as months of L2 instruction or L2 exposure. Months appears to be the best unit since it can bridge the shorter-term studies that have been typical of FL work with SL studies, which may examine learners over a longer period. The number of data points should also be reported in a straightforward manner.

With respect to research transparency, as previously mentioned, reanalysis of the same participant sample should be indicated in each publication, citing any previous publications dealing with that sample. One of the advantages of longitudinal L2 pronunciation research is that in many cases researchers generate large data sets amenable to many different analyses. These analyses can, in turn, provide a panoramic view of distinct facets of L2 pronunciation learning, but this vision cannot be fully realized unless researchers and readers can easily link related analyses. What’s more, as previously noted, linking publications associated with large-scale longitudinal research provides insight into how the project itself has evolved over time.

7. Conclusion

Longitudinal research encompasses a broad set of design decisions. Researchers must determine an appropriate window of observation and sampling frequency and coordinate these longitudinal design choices with anticipated developmental timelines. Until now, relatively short-term studies have dominated longitudinal L2 pronunciation research, and these studies routinely include a small number of data points. In this paper, I have argued for a higher-resolution view of the dynamics of pronunciation learning, which hinges upon a new multi-wave standard in longitudinal research design. To be clear, I do not mean to suggest that longer-term studies, or multi-wave studies, are always better. There are no absolutes in research methodology. Rather, I hope to have encouraged greater awareness of longitudinal design options and how those options can be meaningfully manipulated in light of other study elements to enhance current pronunciation research. I have also pointed out strategic gaps in the literature that future work should address, such as studies of pronunciation development in younger learners. I have focused on quantitative longitudinal research because of my own background and because nearly all of the studies surveyed here have used quantitative methods. However, qualitative studies, especially detailed case studies, may be particularly well-positioned to illuminate issues related to learner agency in pronunciation development.

8. Notes

1. I chose to focus on articles published in peer-reviewed journals and conference proceedings because they undergo a rigorous review process and therefore represent research that has been vetted by experts in the field.
2. Of course, researchers must decide at the planning stage if they are interested in making data publicly available so that they can submit the appropriate documents for institutional review and subsequently obtain participants' consent.

9. References

Akiyama, Y., & Saito, K. (2016). Development of comprehensibility and its linguistic correlates: A longitudinal study of video- mediated telecollaboration. *The Modern Language Journal, 100*(3), 585–609. https://doi.org/10.1111/modl.12338

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics, 40*, 134–142. https://doi.org/10.1017/S0267190520000033

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Baker Smemoe, W., & Haslam, N. (2013). The effect of language learning aptitude, strategy use and learning context on L2 pronunciation learning. *Applied Linguistics, 34*(4), 435–456. https://doi.org/10.1093/applin/ams066

Casillas, J. V. (2019). Phonetic category formation is perceptually driven during the early stages of adult L2 development. *Language and Speech*. Advance online publication. https://doi.org/10.1177/0023830919866225

Casillas, J. V. (2020). The longitudinal development of fine-phonetic detail: Stop production in a domestic immersion program. *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12392

Chang, C. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics, 40*(2), 249–268. https://doi.org/10.1016/j.wocn.2011.10.007

Chang, C. B. (2019). Language change and linguistic inquiry in a world of multicompetence: Sustained phonetic drift and its implications for behavioral linguistic research. *Journal of Phonetics, 74,* 96–113. https://doi.org/10.1016/j.wocn.2019.03.001

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 159–181). New York: Routledge.

de Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition, 10*(1), 7–21. https://doi.org/10.1017/s1366728906002732

Derwing, T. M. (2010). Utopian goals for pronunciation teaching. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference* (pp. 24–37). Ames, IA: Iowa State University.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163–185. https://doi.org/10.1111/lang.12000

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.

Derwing, T. M., & Munro, M. J. (2017, August 31). *Overcoming fears about longitudinal pronunciation research: Anxiety reduction through planning, flexibility, and teamwork* [Workshop presentation]. Pronunciation in Second Language Learning and Teaching Conference, Salt Lake City, UT, United States.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(4), 533–557. https://psycnet.apa.org/doi/10.1017/S0272263109990015

Earle, F. S., & Myers, E. B. (2014). Building phonetic categories: an argument for the role of sleep. *Frontiers in Psychology, 5,* Article 1192. https://doi.org/10.3389/fpsyg.2014.01192

Ellis, N. C. (2014). Cognitive AND social language usage. *Studies in Second Language Acquisition, 36*(3), 397–402. https://doi.org/0.1017/S0272263114000035

Gass, S., Loewen, S., & Plonsky, L. (2020). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*. Advance online publication. https://doi.org/10.1017/S0261444819000430

Hanzawa, K. (2018). The development of voice onset time (VOT) in a content-based instruction university program by Japanese learners of English: A longitudinal study. *Canadian Modern Language Review, 74*(4), 502–522. https://doi.org/10.3138/cmlr.2018-0196

Heinrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hiver, P., & Al-Hoorie, A. (2019). *Research methods for complexity theory in applied linguistics*. Multilingual Matters.

Holliday, J. J. (2015). A longitudinal study of the second language acquisition of a three-way stop contrast. *Journal of Phonetics, 50,* 1–14. https://doi.org/10.1016/j.wocn.2015.01.004

Huensch, A., & Tracy-Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *The Modern Language Journal, 101*(2), 275–293. https://doi.org/10.1111/modl.12395

Kim, D., Clayards, M., & Goad, H. (2018). A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics, 67*, 1–20. https://doi.org/10.1016/j.wocn.2017.11.003

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics, 36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Loewen, S., & Hui, B. (2021). Small samples in instructed second language acquisition research. *The Modern Language Journal, 105*(1), 187–193. https://doi.org/10.1111/modl.12700

McAndrews, M. (2019). Short periods of instruction improve learners' phonological categories for L2 suprasegmental features. *System*, *82*, 151–160. https://doi.org/10.1016/j.system.2019.04.007

Moyer, A. (2014a). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics, 35*(4), 418–440. https://doi.org/10.1093/applin/amu012

Moyer, A. (2014b). What's age got to do with it? Accounting for individual factors in second language accent. *Studies in Second Language Learning and Teaching, 3,* 443–464. https://doi.org/10.14746/ssllt.2014.4.3.4

Muñoz, C., & Llanes, À. (2014). Study abroad and changes in degree of foreign accent in children and adults. *The Modern Language Journal*, *98*(1), 432–449. doi:10.1111/j.1540-4781.2014.12059.x

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning, 58*(3), 479–502. https://doi.org/10.1111/j.1467-9922.2008.00448.x

Munro & Derwing (2015). Intelligibility in research and practice: Teaching priorities. In J. Levis & M. Reed (Eds.), *The handbook of English pronunciation* (pp. 377–396). Routledge.

Munro, M. J., Derwing, T. M., & Thomson, R. I. (2015). Setting segmental priorities for English learners: Evidence from a longitudinal study. *International Review of Applied Linguistics in Language Teaching, 53*(1), 39–60. https://doi.org/10.1515/iral-2015-0002

Nagle, C. (2018a). Examining the temporal structure of the perception-production link in second language acquisition: A longitudinal study. *Language Learning*, *68*(1), 234–270. https://doi.org/10.1111/lang.12275

Nagle, C. (2018b). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal*, *102*(1), 1–19. https://doi.org/10.1111/modl.12461

Nagle, C. (2019a). A longitudinal study of voice onset time development in L2 Spanish stops. *Applied Linguistics, 40*(1), 86–107. https://doi.org/10.1093/applin/amx011

Nagle, C. (2019b). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation, 3*(2), 294–323. https://doi.org/10.1075/jslp.18016.nag

Nagle, C., Levis, J., & Todey, E., (2019). The changing face of L2 pronunciation research and teaching. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (pp. 1–9). Ames, IA: Iowa State University.

Nagle, C., & Rehman, I. (under review). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*.

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition, 41*(4), 647–672. https://doi.org/10.1017/s0272263119000044

Norouzian, R. (2020). Sample size planning in quantitative L2 research. *Studies in Second Language Acquisition*, *42*(4), 849–870. https://doi.org/10.1017/s0272263120000017

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25*, 26–45. https://doi.org/10.1017/S0267190505000024

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Raudenbush, S. W., & Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387–401. https://doi.org/10.1037//1082-989X.6.4.387

Saito, K. (2019). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners’ English /ɹ/ pronunciation. *Second Language Research, 35*(2), 149–172. https://doi.org/10.1177/0267658318768342

Saito, K., & Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning, 67*(1), 43–74. https://doi.org/10.1111/lang.12184

Saito, K., Dewaele, J.-M., Abe, M., & In’nami, Y. (2018). Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study. *Language Learning, 68*(3), 709–743. https://doi.org/10.1111/lang.12297

Saito, K., Macmillan, K., Mai, T., Suzukida, Y., Sun, H., Magne, V., Ilkan, M., & Murakami, A. (2020). Developing, analyzing and sharing multivariate datasets: Individual differences in L2 learning revisited. *Annual Review of Applied Linguistics*, *40*, 9–25. https://doi.org/10.1017/S0267190520000045

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning, 69*(3), 652–708. https://doi.org/10.1111/lang.12345

Saito, K., Suzuki, S., Oyama, T., & Akiyama, Y. (2019). How does longitudinal interaction promote second language speech learning? Roles of learner experience and proficiency levels. *Second Language Research*. Advance online publication. https://doi.org/10.1177/0267658319884981

Saito, K., Suzukida, Y., & Sun, H. (2019). Aptitude, experience, and second language pronunciation proficiency development in classroom settings. *Studies in Second Language Acquisition, 41*, 201–225. https://doi.org/10.1017/s0272263117000432

Sturm, J. L. (2019). Current approaches to pronunciation instruction: A longitudinal case study in French. *Foreign Language Annals, 52*(1), 32–44. https://doi.org/10.1111/flan.12376

Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344. https://doi.org/10.1093/applin/amu076

Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice. *Studies in Second Language Acquisition, 31*(4), 609–639. https://doi.org/10.1017/s0272263109990040

Trofimovich, P., Nagle, C. L., Grantham O'Brien, M., Kennedy, S., Taylor Reid, K., & Strachan, L. (2020). Second language comprehensibility as a dynamic construct. *Journal of Second Language Pronunciation*, *6*(3), 430–457. https://doi.org/10.1075/jslp.20003.tro

Ushioda, E. (2009). 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.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of*

Experimental Psychology: General, 143(5), 2020–2045. https://doi.org/10.1037/xge0000014

Zielinski, B., & Pryor, E. (2020). Comprehensibility and everyday English use. *Journal of Second Language Pronunciation, 6*(3), 352–379. https://doi.org/10.1075/jslp.20011.zie

Appendix A

List of Journals Searched (in alphabetical order)

1. *Annual Review of Applied Linguistics*
2. *Applied Linguistics*
3. *Applied Psycholinguistics*
4. *Bilingualism: Language and Cognition*
5. *Computer Assisted Language Learning*
6. *Foreign Language Annals*
7. *Journal of Phonetics*
8. *Journal of Second Language Pronunciation*
9. *Language and Speech*
10. *Language Learning*
11. *Language Learning and Technology*
12. *Language Teaching Research*
13. *The Modern Language Journal*
14. *Second Language Research*
15. *Studies in Second Language Acquisition*
16. *System*
17. *TESOL Quarterly*

Appendix B

Summary and Description of Coded Variables

General Variables

1. Study: This variable enumerates studies in the data set. Multiple entries (see Reason for Multiple) are signaled by decimals (e.g., 1.1, 1.2 both refer to the same study)
2. Authors: A complete list of authors of the study.
3. First Author: The first author of the study.
4. Year: The publication year of the study.
5. Venue: An abbreviation indicating where the study was published.
6. Reason for Multiple: Description of the reason why multiple entries were included for the study, if relevant. For instance, studies have multiple entries because they examined multiple learner age groups (see Participant Age Bin), learners in different contexts of learning (see Context, Expanded Context), or included multiple types of outcome measures or tasks (see SP Construct, SP Task)

Longitudinal Variables

7. Total Length: Total length of the study, as described in the publication.
8. Length Months: Total length of the study in months, calculated based on Total Length.
9. Length Bin: A categorical variable sorting studies into five length bins, expressed in months:
    a. 0–4
    b. 4–8
    c. 8–12
    d. 12–24
    e. 24+
10. Data Points: The number of data points included in the study.
11. Data Spacing: A binary categorical variable that represents whether data was collected at equal or unequal intervals.
12. Unequal Specification: A description of how the data was spaced out if Data Spacing was unequal.

Participant Sample Variables

13. N Speakers: The number of speakers included in the study. For studies reporting both cross-sectional and longitudinal analyses (e.g., Saito et al., 2018), this number reflects only the longitudinal portion of the study.
14. Participant Age: Mean age of participants or participant groups, if reported.
15. Participant Age Bin: A categorical variable sorting studies into three age groups based on Participant Age or a narrative description of sample characteristics:
    a. Child: 0–9 years
    b. Adolescent: 10–17 years
    c. Adult: 18+ years

16. Context: A binary categorical variable that represents the primary learning context of the study:
    a. FL: Studies focusing on foreign language, or instructed, learners.
    b. SL: Studies focusing on second language, or naturalistic, learners.
17. Immersion: A variable that indicates if the SL learners in the study were short-term immersion learners.
18. Expanded Context: A categorical variable that includes three levels for context:
    a. FL: Studies focusing on foreign language learners.
    b. SL: Studies focusing on second language learners who had relocated to an L2 environment.
    c. SL SA: Studies focusing on second language learners who were participating in an immersion or exchange program but intended to return to an L1 environment.
19. L1: The native language(s) of the participants or participant groups.
20. L2: The target language that the participants were learning.

Measurement Framework Variables (Saito & Plonsky, 2019)

21. SP Construct: A categorical variable that represents the type of target feature or outcome:
    a. Perception: Studies examining L2 perception (e.g., Kim, Clayards, & Goad, 2018).
    b. Production – Specific: Studies examining the development of specific features such as L2 vowels, consonants, etc.
    c. Production – Global: Studies examining the development of global features such as comprehensibility.
22. SP Task: A categorical variable that represents the type of task used to elicit data:
    a. Perception
    b. Production – Controlled: Studies examining pronunciation using controlled tasks such as word reading.
    c. Production – Spontaneous: Studies examining pronunciation using spontaneous tasks such as picture description or story narration.
23. SP Coding: A categorical variable that represents how data were coded:
    a. Perception
    b. Production – Coded: Studies in which production data were coded for linguistic features or submitted to acoustic analysis
    c. Production – Rated: Studies in which production data were evaluated by raters (e.g., comprehensibility, accentedness)

Additional Variables

24. Actual Tasks: A short description of the actual tasks used to elicit data.
25. Actual Outcome Measures: A short description of the actual outcome measures included in the study.
26. Statistical Tests: A short description of the statistical tests included in the study.
27. N Raters: The number of raters included in the study, if the study involved impressionistic evaluations (see SP Coding: Production – Rated).
28. Rater Profile: A binary categorical variable that represents rater expertise:

a. Expert/Trained: Studies including raters with training in linguistics or language teaching
b. Novice/Untrained: Studies including naïve or novice raters with a background in linguistics or language teaching

Address for Correspondence

Charles Nagle

Department of World Languages and Cultures

Iowa State University

3102 Pearson Hall

505 Morrill Road, Ames, IA 50011

cnagle@iastate.edu

https://orcid.org/0000-0003-2712-2705