Empathy, Equity, Empowerment

Editors

Cassandra Glynn
Concordia College

Allison Spenader
College of Saint Benedict & Saint John's University

Articles by

- **Gabriela Olivares-Cuhat**
- **Giovanni Zimotti**
- **Jessica Sertling Miller**
- **Amy Kim**
- **Margaret Borowczyk**
- **Rima Elabdali**
- **Margaret E. Malone**
- **Sean Hill**
- **Lisa Brittingham**
- **Clara Burgo**

Central States Conference on the Teaching of Foreign Languages

Empathy, Equity, Empowerment

Central States Conference on the Teaching of Foreign Languages Report 2023

Cassandra Glynn, Editor
Concordia College

Allison Spenader, Editor
College of Saint Benedict &
Saint John's University

**2023 Report of the
Central States Conference on the Teaching of Foreign Languages**

Publisher: Robert M. Terry
116 North Railroad Avenue, Suite 16
Ashland, VA 23005

Central States Conference on the Teaching of Foreign Languages
Daryl Biallas, Executive Director
executive.director@csctfl.org
www.csctfl.org

C P 0 9 8 7 6 5 4 3 2 1

Review and Acceptance Procedures

Central States Conference on the Teaching of Foreign Languages Report

The Central States Conference on the Teaching of Foreign Languages Report is a refereed volume of selected papers based on the theme of the Central States Conference on the Teaching of Foreign Languages. Copies of the publication guidelines are available to authors on the CSCTFL website.

All submissions are read and evaluated by the Editor and at least two other members of the Editorial Board. When all of the reviewers' ratings are received, the Editor makes all final publishing decisions.

2023 Central States Conference on the Teaching of Foreign Languages Report Editorial Board

2022 – 2023 CSCTFL Officers and Directors

Chair	Jill Woerner
Vice Chair	Linda Havas
Executive Director	Daryl Biallas
Recording Secretary	Lesley Chapman
2023 Program Chair and CSCTFL Delegate to ACTFL	Amanda Ramirez
Director	Josh LeGreve
Director	Marcia Davis
Director	Kristopher Morehead
Director	Elke Lorenz
Director	Kris Morehead
Director	Ryan Rockaitis
Director	Janet Rowe
2023 Assistant Program Chair	Marci Harris

Contents

Review and Acceptance Procedures i

Central States Conference Board of Directors 2022-2023 ii

Preface v

Introduction vii

Empathy, Equity, Empowerment

1. **Empowering L2 Learners: The Role of Relational Empathy in L2 Learning** 1
   Gabriela Olivares-Cuhat & Giovanni Zimotti
2. **Academic Integrity in the Virtual Classroom** 21
   Jessica Sertling Miller
3. **The Impact of Hybrid Professional Development on World Language Instructors' Assessment Literacy: Challenges and Opportunities** 39
   Amy Kim, Margaret Borowczyk, Rima Elabdali, Margaret E. Malone
4. **Student Mathematics Achievement in a Mandarin Immersion Program** 65
   Sean Hill
5. **Musical Background and French Pronunciation: A Mixed Methods Study** 103
   Lisa Brittingham
6. **Empathy, Equity, and Empowerment in the Spanish classroom** 125
   Clara Burgo

Preface

Empathy, Equity, Empowerment

The 2023 Central States Conference on the Teaching of Foreign Languages was held in Columbus, Ohio on March 9-11. The CSCTFL Board was ecstatic to be welcoming and supporting educators from across our region this year in a return to in-person professional development and networking.

This year's theme, Empathy, Equity, and Empowerment is reflective of the vital work we as educators do in classrooms every day across the country. The Wikipedia definition of empathy is "the capacity to understand or feel what another person is experiencing from within their frame of reference, that is, the capacity to place oneself in another's position". Our most basic job as language educators is to develop that empathy in and through our students using language and intercultural understanding. These are not just vital to understanding other countries, but vital to understanding our neighbors, our colleagues, our families, and the students that walk our hallways and fill our classrooms.

Building awareness and empathy is the first step toward equity, as we begin to view practices and policies from new perspectives and experiences. Building empathy in ourselves is a continuous process, and when we participate in that process it empowers us to be better advocates for equity. The actions we take for better equity with that knowledge are what empower both students and educators alike.

The 2023 keynote speech was titled "Inquire, Empower & Inspire to Greater Heights: Equity in Action!" and was given by A.C. Quintero of Chicago, Illinois. She has worked in education for 17 years, authored several comprehensible novels, and was the recipient of the 2022 AATSP Diversity, Equity, and Inclusion Award.

Over 10 workshops were offered this year, in addition to the CSCTFL Leadership Academy and the CSCTFL Extension Workshop. The Extension Workshop, given by María Datel of Boston University on behalf of ACTFL, was on integrating Indigenous perspectives on global issues into the language curriculum.

More than 150 sessions were offered this year focusing on inclusivity, equity, social justice, classroom activities, teaching strategies, curriculum development, assessment, intercultural competence, advocacy, and technology. Presenters came from over 30 states across the country to share their expertise and knowledge with attendees, and there were more presentations on inclusive curriculum and classrooms than ever before.

The Central States Conference Report 2023, Empathy, Equity, Empowerment, is a call to language educators to use their voice and their content to build a better world, starting in our classrooms. Thank you to the authors for their work in supporting our students, their peers, and this endeavor.

Amanda J. Ramirez

Empathy, Equity, Empowerment

Cassandra Glynn
Concordia College

Allison Spenader
College of Saint Benedict &
Saint John's University

After attending CSCTFL online for two years, we looked forward to returning to an in person conference in Columbus, Ohio with the theme of *Empathy, Equity, and Empowerment*. However, the state of our field has changed since we last met in person in Minneapolis in 2020, and teachers and students are certainly not the same as they were prior to the pandemic. We are still overcoming a myriad of challenges that the pandemic exacerbated including burnout, loss, racial injustice, and economic insecurity, just to name a few. More than ever, empathy and compassion deserve a central place in education and are vital components of supporting equity and empowerment of both students and teachers in K-16 education.

Issues of *Empathy, Equity, and Empowerment* are more complex than they initially appear. Bialystok and Kukar (2018) posit that empathy can become a "stand in for tolerance;" that is, without interrogating issues of power and privilege, it can be hard to work toward true empathy in schools (p. 30). Attempting to cultivate empathy among privileged students without providing a wider context, historical or social, to understand the stories or the pain of the marginalized group toward whom the empathy is being directed, is problematic (Bialystok & Kukar, 2018; Ennser-Kananen, 2016). In world language education, it can be difficult to convince teachers to move away from only teaching pleasant topics, and Ennser-Kananen urges that "we have to look for the pain and the unsettling within our WL curriculum and instruction in order to adequately address what happens around us" (p. 560). Kubota implores us to examine "...the diverse knowledge reflected in teaching materials and students' voices" (2016, p. 207) so that we might realize contextual relativism and engage in critical pedagogies in our language classrooms. Acknowledging the complexities of stories and experiences, and situating student voices at the center can support the goals of *Empathy, Equity and Empowerment.* All of us, including our colleagues, our students, and their families and communities, have experienced a great deal of change and trauma in the last few years. Therefore, as editors of the CSCTFL Report, we are most grateful to have received submissions this year that focus on supporting all teachers in becoming practitioners of empathetic and equitable practices. This attention to increasing the quality of instruction will lead to the empowerment of learners and teachers across diverse educational contexts.

We would like to thank Robert Terry, Jill Woerner, and Daryl Biallas for their guidance and assistance as we have compiled this year's report. We offer special gratitude to the editorial board for the careful feedback they provided to the authors of the six pieces included in this year's report. The 2023 Central States Conference on the Teaching of Foreign Languages conference theme of *Empathy, Equity, Empowerment* guides this year's submissions in a myriad of ways. The report begins and ends with articles focused on cultivating empathy in the classroom. In between, we have a variety of articles that address issues of equity through academic integrity in college courses and math instruction in a language immersion program in a low SES school. An additional article focuses on providing effective professional development related to language assessment, while another article focuses on using students' background knowledge and experiences in a French class, both of which connect to issues of equity and empowerment.

We begin the report with an article by Gabriele Olivares-Cuhat and Giovanni Zimotti who examine the potential for considering the role of relational empathy in second language acquisition, a concept that is often linked to intercultural communication. They offer a myriad of examples of classroom activities that foster relational empathy while supporting the important goal of social justice learning.

In the next article, Jessica Sertling-Miller argues that academic integrity violations can be linked to issues of access and equity. When students feel a lack of agency and increased stress, that can lead to perceived academic misconduct. Sterling-Miller explores the role of Machine Translation in online teaching and draws on principles of Universal Design for Learning for improving course design. She explores how designing a course with equity and access in mind can help students to uphold academic integrity in their work.

Amy Kim, Margaret Borowyczyk, Rima Elabdabi, and Margaret Malone share their mixed methods research about the long term effects of hybrid professional development among instructors and administrators of less commonly taught languages (LCTLs). They found that although the participants did shift their practices toward a more proficiency-based approach, they also experienced challenges in retaining and using what they had gained in the professional development.

Sean Hill used a hierarchical linear or multilevel modeling (HLM) approach to quantifying students' mathematics achievement in a one-way Chinese (Mandarin) immersion program. The context of the research is a rural school district marked by generational poverty, and the study examined the impact of variables such as free and reduced lunch, special education status, and gender on students' achievement in math while in the program. Hill found that students' preparation and development prior to kindergarten had a significant influence on achievement.

In the fifth article, Lisa Brittingham interrogates whether or not a student's musical background can be capitalized on in the classroom to help them improve their French pronunciation. Using a mixed methods design, Brittingham examined novice level French learners at the post-secondary level and found a positive

relationship between musical background and French pronunciation. The study also led Brittingham to differentiate between pronunciation that is intelligible and pronunciation that is native-like.

We end the report with Clara Burgo's article about centering voices and stories of Latinx *testimonios* to build empathy, equity, and inclusion in the classroom. Burgo focuses on the importance of including Latinx voices through *testimonios* in post-secondary classrooms with both heritage and non-heritage language learners in order to build safe spaces in which students can find validation and share their own opinions and experiences.

Through the diverse scholarly lenses included in the report this year, we hope that you gain new ideas, perspectives, and insights into the various ways in which *Empathy, Equity, and Empowerment* can be enacted in your own practice.

Sincerely,

Cassandra Glynn and Allison Spenader
2023 Report Editors

Bialystok, L. & Kukar, P. (2018). Empathy and authenticity in education. *Theory and Research in Education*, *16*(1), 23-39.

Ennser-Kananen, J. (2016). Pedagogy of pain: New directions for world language instruction. *Modern Language Journal*, 100, 556-564.

Kubota, R. (2016). Critical Content-Based Instruction in the foreign Language Classroom: Critical Issues for Implementation. In L. Cammarata (Ed.), *Content-based foreign language teaching: Curriculum and pedagogy for developing advanced thinking and literacy skills.* (pp. 192-211) Routledge.

1

Empowering L2 Learners: The Role of Relational Empathy in L2 Learning

Gabriela Olivares-Cuhat
University of Northern Iowa
Giovanni Zimotti
University of Iowa

The Challenge

Empathy has emerged as a necessary component for the betterment of social justice in the community, and as such should be viewed as part of the second language (L2) learning process and classroom practices. Concurrently, how does the exercise of empathy in the classroom provide a pedagogical advantage to L2 learners?

Abstract

In this article, we explore the relevance of the concept of relational empathy to L2 learning. To this end, we first provide a short overview of the notion of empathy in second language acquisition (SLA). Then, we present ideas that gave rise to the introduction of relational empathy in the field of intercultural communication, along with potential benefits of taking this perspective in SLA. Next, we describe the post-method framework for L2 teaching as a strategy that is both well-suited for the fostering of relational empathy and is facilitated by the exercise of this approach. Following this lead, we explain in some detail Broome's (2017) guidelines for the development of relational empathy and highlight the congruence of this approach with important principles of L2 teaching and learning. We then offer samples of classroom activities to further demonstrate the value of integrating the exercise of relational empathy in L2 instruction. In conclusion, we observe that these views are consistent with best practices in language instruction, such as the ones given in the *World Readiness Standards for Learning Languages* (National Standards Collaborative Board, 2015) and the Can-Do statements for

Intercultural Communication (National Council of State Supervisors, NCSSFL), and submit that they further social justice through the promotion of a respectful and empowering teaching and learning environment.

Empowering L2 Learners: The Role of Relational Empathy

A capacity for empathy is often posited as a condition for the development of moral values. As stated by Hoffman (2000): "Empathy appears to be congruent with all or most […] justice principles, as well as with caring" (p. 273). What is more, it has been suggested that a disposition for using empathy in communications constitutes a critical element for building better intercultural relationships (Bennett, 1979; DeTurk, 2001). Given that language instruction is at the forefront of promoting intercultural communication, these views are thus in agreement with Mihaly (2021), who, in reaction to the murder of George Floyd and the social unrest that followed, stressed the necessity of making empathy a core component of this field. In addition, there is a sense that the exercise of empathy contributes to making postsecondary students better learners of a second language (L2).

While these considerations seem to support the idea of making empathy a focus of second language learning (L2), the reasons that link empathy to L2 learning call for more scrutiny. It remains difficult to articulate pedagogical recommendations with regards to embedding empathy into the L2 classroom. Thus, the aims of this article are twofold: first to explain how the practice of empathy may also be beneficial to the L2 learner, and second to present examples of classroom activities that may assist teachers in making empathy a central component of the L2 learning experience. To attain these goals, the authors will first review some theoretical concepts.

Conceptual Framework

Even before a name was coined for it, the nature and role of empathy were described and explained from a wide range of perspectives, such as philosophy, poetry, the arts, sociology, psychology, phenomenology, cognitive science, and developmental psychology (Howe, 2013). As generally defined by Demetriou (2018): "Empathy … is a feeling with the person and has the ability to truly understand what another being is going through but also, to harness a deep-seated emotional appreciation of another's situation" (p. 29). In the field of second language acquisition (SLA), Brown (1973) laid a foundation by observing that "Language is one of the primary means of empathizing" (p. 235). Since then, a range of angles have been used by SLA researchers to describe the role that empathy plays in their field. Accordingly, it was viewed as an emotional factor that links a teacher to their students (Arnold & Foncubierta, 2022); as a construct consisting of cognitive, affective, and empathetic elements (Mercer, 2016); within the framework of individual difference research (Dörnyei, 2009); within the context of the field of positive psychology (Oxford, 2016); as a category in personality inventories used by SLA researchers (Dewaele, 2022); as a facet of emotional intelligence (Mercer, 2016; Salovey & Mayer, 1990); and as a positive component of group dynamics (Falout, 2014).

While generally accepted, a link between empathy and L2 learning has only been demonstrated in a few studies. Guiora et al. (1972) successfully correlated authenticity of pronunciation with the ability to correctly assign emotions to facial expressions, using the Micro-Momentary Expression (MME) test. In an empirical study conducted with the participation of 169 college students enrolled in Spanish courses at a midsize Midwestern university, significant correlations were revealed between L2 performance and empathy subscales of an emotional intelligence test developed by Caruso and Mayer (1998) including positive sharing, emotional attention, feel for others, and total empathy scale (Olivares-Cuhat, 2012). As another example, a large-scale empirical study involving the participation of 6,434 subjects (Grin & Faniko, 2012) pointed to significant relationships between world language skills and dimensions of a psychological test that evaluates traits related to cultural empathy as defined by Van der Zee and Van Oudenhoven (2000).

Yet another view of empathy relevant to SLA is how it relates to communication as a main goal of language learning. An emphasis on the importance of communication in L2 learning is found in the 21st Century Skills map, where it is considered as one of the main elements of the *World Readiness Standards* and the Intercultural communication proficiency benchmarks established by NCSSFL-ACTFL (Partnership 21st Century Skills Map). Consistent with this view, Mercer (2016) stated that:

> In the process [of communicating effectively], seeking to understand the mind of the person you are communicating with can greatly assist communication … Ultimately, communication is about understanding how another person might interpret or misinterpret an act of communication and requires you to put yourself into the mind of the other to some degree. (p.100)

As it happens, researchers in intercultural communication have long been preoccupied with the role played by empathy in promoting better communication. In this respect, a key contribution was made by Bennett (1979), who proposed that the development of empathy must account for essential differences that exist between groups (with different languages and cultures) and between individuals (with different personal circumstances and experiences). It is our understanding that these disparities could be extended to other categories including gender, race and socioeconomic status (Freire, 1970) as defined within the application of critical theory to education. Indeed, as articulated by Kumaravadivelu (2006): "The experiences participants bring to the pedagogical setting are shaped, not just by what they experience in the classroom, but also by a broader social, economic, and political environment in which they grow up" (p.174). Rather than trying to imagine oneself in the position of another person, empathy becomes an attempt to participate with and understand the experience of the other person: "Empathy describes a shift in perspective away from our own to an acknowledgment of the other person's different experience" (Bennett, 1979, p. 419). In this context, empathy becomes primarily a relational phenomenon. Again, within the framework of intercultural communication, Broome (1991) furthered this view by proposing the concept of relational empathy, whose nature is dynamic, provisional, and

context-based: "In order for empathy to have intercultural validity, it must be seen as part of the communication process itself, and thus it may be influenced by the variables in that process" (p.240). In the same article, Broome further highlights the contrast between the psychological approach, in which the listener recreates the meaning originally intended by the speaker, and the act of creating a "shared meaning" achieved during interpersonal exchanges. This conceptual distinction becomes clear when contrasting the psychological definition of empathy preferred by Brown (1987): "putting yourself into someone else's shoes" (p. 165) with Broome's (1991) relational perspective stating that: "…The dynamic, contextual, and creative characteristics of a relational conceptualization of empathy can instruct our approach to teaching intercultural communication" (p. 245). These ideas were taken even further by DeTurk (2001), according to whom: "Treating empathy simply as a skill or competency is unrealistic. Relational empathy … can enhance both intercultural understanding and commitment to social justice …" (p. 383). In sum, relational empathy may be defined as the continuous building of a mutual understanding and shared meaning through dialogic interactions.

The main argument of this article is that the same empathy guided processes that are shown to be effective towards building intercultural communication may also be implemented in the L2 environment, which then becomes a tool to shape a more positive—and effective—learning experience. In other words, an advantage of relying on relational empathy (as defined by intercultural communication scholars) in the process of interacting with language learning peers is that it fosters the creation of a safe and equitable environment, which, in turn, is likely to create conditions more conducive to the L2 acquisition process. This latter implication is supported by several fields of inquiry in SLA. First, it has been recognized that negative emotions may have a significant impact on L2 students' learning (MacIntyre & Ayers-Glassey, 2022). For example, such links have been demonstrated by empirical studies on foreign language classroom anxiety (Horwitz, 2010), which have led to the recommendation of providing students with a "supportive classroom environment" (Horwitz, 2013). More generally, the foundation of the critical role played by such reciprocal actions may be found in the cognitive-interactionist L2 theories derived from Long's Interaction Hypothesis (1983) and Swain's Comprehensible Output Hypothesis (1985), which posit that interactions are central to the process of L2 acquisition. Accordingly, "the input and output that arises in interaction links with internal processes such as noticing to promote interlanguage development" (Ellis & Shintani, 2014, p. 223). Second, following sociocultural theories, the quality of a classroom environment may also be linked to the phenomenon of language socialization, i.e., how the practice of a language is closely associated with its use in society. Therefore, according to Ochs (1996): "An important premise of language socialization research is that language socializes not only through its symbolic content but also through its *use*, i.e., through *speaking* as a socially and culturally situated activity" (p. 408). These ideas are grounded on a body of SLA research that was initiated by Vygotsky (1978) and furthered by Leontiev (1978), Wells (1999), and Lantolf (2000), among others. They are based on the general ideas that higher functions of learning may only take

place through social interactions and, as a particular case, view the learning of a language as a relational phenomenon mediated by cultural and symbolic artifacts (Thorne & Lantolf, 2007). As a social phenomenon, progress is then facilitated by multiple interactions that may take place within the framework of classroom activities. As summed up by Ellis and Shintani (2014): "Language development is a social process, mediated by both interpersonal and intrapersonal interaction" (p.223).

A question remains as to whether a learning environment may be transformed by its participants. As it happens, it is a central tenet of the social learning theory proposed by Bandura (1976): "Thus, behavior partly determines which of the many potential environmental influences will come into play and what forms they will take; environmental influences, in turn, partly determine which behavioral repertoires are developed and activated" (p. 195). Along these lines, the benefits that group dynamics exerts on L2 learning were demonstrated by Murphey et al. (2021) who examined "… how language learners influence and bond with each other in a group by sharing their past histories, present feelings, and future aspirations" (p. 290). These ideas were further validated within the framework of a large longitudinal study conducted on L2 students enrolled in Japanese universities (Fukada et al., 2017). The positive behaviors that were observed in this research appear to be congruent with the purported role of relational empathy in intercultural communication, thus reinforcing the notion that this type of behavior is a critical component of a rewarding language learning environment.

Bridge to L2 Instructional Practice

How do the concepts presented in the previous section translate to the L2 classroom? To this end, it should first be acknowledged that there is a strong push in the L2 teaching profession to set aside the use of predefined sets of methods to restore a more balanced relationship between theory and classroom practices, a criticism which has led to the inception of a post-method position to L2 instruction (Leaver, 2021). An important premise to this transition, which was articulated by Kumaravadivelu (2006), is that it must revolve around several principles: particularity, which essentially posits that the pedagogical content is specific to the situational circumstances of the classroom; practicality, which implies that teachers "ought to be enabled to theorize from their practice and practice what they theorize" (p. 123); and possibility, which stresses that social context and cultural identity must be integrated into the classroom experience. It can then be derived from these principles that a post-method framework must preserve the autonomy of the students (who control the use of their learning strategies) and teachers (who engage on a career-long path of self-development).

What could be the role of empathy in this frame of reference? As pointed out by Breen et al. (2001), who discuss ways in which teaching and learning principles learned from experience influence decision making in a specific L2 environment, the reflexive process plays a critical role to make it possible for the teacher to continuously adjust to the ongoing classroom conditions. Breet et al. state: "Pedagogic principles are reflexive in both shaping what the teacher does

whilst being responsive to what the teacher observes about the learners' behavior and their achievements in the class" (p. 473). If we generalize this idea to the entire classroom, this description becomes consistent with the very concept of relational empathy, by which positive relationships arise within a group through interpersonal exchanges and mutual feedback. More generally, a case can be made to the effect that the concept of relational empathy is recurrent in the post-method pedagogy—although not referred to by this name. For instance, within his exploratory practice framework, Allwright (2003) proposes seven principles, including "Put quality of life first," "Work primarily to understand language classroom life," "Involve everybody," "Work to bring people together," and "Work also for mutual development," which are congruent with ideas underlying the fostering of relational empathy. As summarized by Kumaravadivelu (2006): "Collegiality becomes crucial to pedagogic enterprise" (p.196). By the same token, a number of macrostrategies proposed by Kumaravadivelu (2006) would seek to reinforce relational empathy through L2 learning, e.g., "Facilitate negotiated interactions," "Ensure social relevance," and "Raise cultural consciousness" (pp. 202-207). In sum, the promotion of relational empathy could be viewed as a facilitation of the L2 post-method paradigm.

Injecting Relational Empathy into the L2 Classroom

At this point, we should consider the critical question faced by the post-method L2 teacher who would intend to rely on the practice of relational empathy, namely: What strategies are at their disposal? To this end, we propose relying on approaches and solutions that have been informed by researchers in the field of intercultural communication. In doing so, our basic assumption is that the understanding of behaviors observed within the framework of intercultural communication are applicable to the relations that are built within an L2 learning environment. With this in mind, we first recognize a critical contribution made by Deardorff (2006), who identified a set of requisite attitudes for the building of intercultural competence. They consist of three components: "Respect (valuing other cultures, cultural diversity)," "Openness (to intercultural learning and to people from other cultures, withholding judgment)," and "Curiosity and discovery (tolerating ambiguity and uncertainty)." While these principles were proposed with the aim of improving communications with members of other cultures, they may also be of significant benefit within the L2 classroom as "empathy allows us to build trust and respect, reduce tensions, encourage information-sharing, and create a safe environment for interaction" (Broome, 2017, p.1284). Accordingly, these dispositions could sum up the conditions that the teacher should instill among their students as part of an empathy-based learning environment. As a next step, following the notion of relational empathy proposed by Broome (1991), it is crucial for the teacher to cultivate an active engagement of the students with the target culture as well as between themselves. To better understand what this means, one must highlight the imperative to "focus on the creation of shared meaning" and on the idea that "understanding is not viewed as a product, but as an ongoing process occurring between communicators" (p. 1285). Furthermore,

Broome (2017) proposed that the promotion of relational empathy depends on the application of five guiding principles (pp. 1285-87). It should be noted that here as well, these concepts may naturally be expanded to relationships between students within a classroom.

- **View understanding as dynamic and provisional.** This guideline is derived from the recognition that interactions with others are always influenced by preconceived ideas. So, a first step is to accept that a mutual understanding is never reached but must be derived instead from a deliberate process, which therefore is always changing and in progress. This view agrees with the central role that Breen et al. (2001) assigns to reflexiveness in L2 learning, as it involves a constant reevaluation of one's opinions about others during interpersonal exchanges. Thus, "intercultural empathy is developed through an ongoing, corrective process that is dynamic and circular" (p. 1285).
- **Focus on the bridging of difference.** This contrasts with the natural tendency to read an action according to our own interpretations, which forgoes the effort of truly attempting to understand another point of view. The same precept is also proposed by Kumaravadivelu (2006) within the context of the post-method L2 classroom in the form of a macrostrategy recommending to "Minimize perceptual mismatches" (p. 204). In his words: "An awareness of these mismatches can help us effectively intervene whenever we notice or whenever learners indicate problems in carrying out a specified classroom activity" (p. 204).
- **Shift our perspective away from our own experiences.** This suggests that one must resist the temptation to assess others through their own system of values and focus instead on the inner thoughts and motivations that are favored by others. This principle may be also summarized by the "Platinum Rule" proposed by Bennett (1979): "Do unto others as they themselves would have done unto them" (p. 422). As it happens, the benefits brought by this change of reference are well known within the field of transformative language learning (TLL), which for instance claims that learners' success in L2 while studying abroad is conditioned by the need to reach an appropriate level of intercultural development "that permits them to shift from the home culture perspective to that of the target culture" (Davidson et al., 2021, p. 111).
- **Base our interpretation on the overall context of the encounter.** This tenet requires one to consider the realities experienced by the others as an essential factor towards reaching a better understanding of their point of view. From the perspective of L2 learning, this view is congruous with Kumaravadivelu (2006) macrostrategy advocating to "Ensure social relevance", which involves developing an awareness for the socio-economic, political, and cultural conditions that underlie the participants' lives.
- **Strive for a synthesis of perspectives.** This idea posits that it is crucial to seek a common understanding between a diversity of views, even when they are too different to lead to a form of compromise, thus leading to a fundamental "goal of intercultural empathy [...] to create a unique understanding" (p. 1287).

L2 Activities for Promoting Relational Empathy

In this section, the authors propose examples of activities aimed for L2 classrooms at the secondary and post-secondary levels, which are aimed at taking advantage of relational empathy to establish better communications. In doing so, the teacher creates a classroom that minimizes the impact of negative emotions and invites the participants to engage in meaningful interactions. In turn, this should promote the acquisition of the L2 by providing a supportive environment and integrating oral interactions situated within the social and cultural frameworks of the target language. In alignment with Deardorff's recommendations (2006), the cultivation of a learning environment that combines respect, openness, curiosity and discovery should be viewed as a precondition for this type of exercise to be successful. The following examples stem from Broome's (2017) abovementioned five guiding principles for the promotion of empathy in intercultural communications. Each activity includes suggested grade levels, proficiency targets, and recommended functions and forms in the case of a Spanish language classroom.

Activity 1. View Understanding as Dynamic and Provisional

Grade level: 9-16
Proficiency level: Novice through advanced
Functions: Expressing likes/dislikes, considering other's point of view, stating preferences **Forms:** *Gustar (to like), preferer (to prefer),* present tense, subjunctive

The purpose of this activity is to steer a group of students towards considering the point of view of another group, and to do so in a dynamic way, i.e., by leading the class to an inclusive outcome, rather than casting views aside. The following example refers to a fondness expressed by students for a specific musical style as a way to initiate a conversation.

- Select two genres reflecting the interest of the students (e.g., either reggae or hip-hop).
- Request students to select one option and form groups.
- Each group is required to write at least five reasons that support their choice.
- The teacher guides a discussion by contrasting aspects of these styles including musical elements and lyrics. In doing so, the teacher may allude to the social justice themes found in reggae versus the more individualistic view of society pervasive in hip-hop.
- Students defend their choice and listen to the arguments of the other side.
- The teacher asks students to vote for a favorite song from the musical style they did not select in the first place.
- Students are asked to write a short reflection on reasons that incited their classmates to support another style of music than they chose.

Activity 2. Focus on the Bridging of Difference

Grade level: 9-16
Proficiency level: Intermediate (includes a complication) through advanced
Functions: Asking for help, resolving conflict
Forms: *Me puede ayudar (can you help me), podría (would you help me),* present tense, conditional.

In this activity, a role-play scenario is set to enact a situation illustrating a misunderstanding stemming from cultural differences. For instance, a first group of students may impersonate the members of a North American family moving into a South American country, while a second group represent their new neighbors. For instance, a US family from the Midwest moves to a residential neighborhood in Santiago, Chile. While doing so, they expect to experience some friendly interactions with their new neighbors and look forward to doing so while using their hard-won Spanish speaking skills. However, they are disappointed to be apparently ignored (or snubbed?) as nobody seems to even try to approach them. What they do not understand is that Chilean neighbors do not acknowledge newcomers because it is not a cultural norm in this part of the world. At the same time, the Chilean neighbors are put off by the showy attitude of the Midwestern family and dislike their fashion choices. The activity culminates in recreating an encounter taking place as the North American family coincidentally encounters the Chilean family at a grocery store.

- The instructor reviews social customs in Chile and asks the US students to compare them with their own behaviors.
- The students get into groups and select the characters they will represent in the role-play situation (e.g., parents, siblings).
- The teacher informs each group of a need that they encounter, e.g., the US family is out of cash (they miscalculated the exchange rate) and the Chilean family is looking for a ride as their car broke down.
- Both families meet incidentally at the grocery store, but the participants are not aware of the specific need experienced by the other group.
- The students must reach out, identify, and negotiate actions that are in their best mutual interests (while staying in character) by bridging differences.
- Finally, students are asked to write a short reflection on the insight they gained through this exercise.

Activity 3. Shift our Perspective Away From our Own Experiences
Grade level: 9-16
Proficiency level: Intermediate through advanced
Functions: Analyzing literary pieces, expressing an argument, shifting perspectives, integrating others' point of view, writing an argumentative essay
Forms: *Opinar (to give an opinion), parecer (to appear),* present tense

Students are faced with the ambiguous actions of a famous historical figure and put in a situation where they must decenter themselves from an initial opinion they formed about them. In an essay they must then explain how they process the dichotomy between myth and reality and how they may still see the value of a legacy despite the personal failings of its author.

- Students read poems from Pablo Neruda, for instance from *Los veinte poemas de amor,* and then reflect on them.
- In the next session, the instructor presents biographical information on Pablo Neruda that reveals objectionable aspects of his life, which has led to a rejection of his work by some intellectual circles.

- Students are asked to take a position in this debate and submit anonymously a few reasons behind their choice (as perhaps motivated by their own personal circumstances).
- The teachers select, read and discuss some of the feedback with a view to emphasizing the critical dynamics that arise within the classroom.
- Subsequently, students are asked to write an essay in two parts: first defending the work of Pablo Neruda to one of his accusers, and second to argue against the work of Pablo Neruda with a professor of Spanish literature.

Activity 4. Base Our Interpretation on the Overall Context of the Encounter
Grade level: 16 Advanced Placement
Proficiency level: Advanced
Functions: Interpreting oral and written text, expressing understanding, synthesizing different type of information
Forms: Major time frames, tenses

The teacher writes a famous quote on the board, first without offering any further background. After the students react, the teacher provides more details about the text, the life, and the historical circumstances of the author. The students are then given an opportunity to review their previous positions and consider this new information.

- For example, the following quote by Federico García Lorca could be used in an L2 Spanish class "*En España, los muertos están más vivos que en cualquier otro país del mundo*." (In Spain, the dead are more alive than anywhere else in the world.)
- The teacher requests the students to react to this statement and write down their impressions.
- The teacher reads reactions expressed by the students (while preserving their anonymity).
- The teacher then reads the text "*Juego y teoría del duende*" (1933) from which this quote originates.
- The students are then requested to describe the new understanding they have gained from classmates' feedback and their own interpretation of the original text.
- Finally, the students are asked to reflect on how they were better able to interpret the thoughts of García Lorca after reading—and possibly getting inspired by—the passionate talk that the author gave to his original audience.

Activity 5. Strive for a Synthesis of Perspectives
Grade level: 9-16
Proficiency level: Novice through advanced
Functions: Preparing a travel itinerary, finding common ground, negotiating, stating preferences
Forms: Present and future tenses, locations

For this role-play activity, the classroom must plan a study abroad trip that combines L2 learning with other interests. It is up to them to identify a venue that may best satisfy all the participants. For instance:

- The participants are divided into two groups, representing students who are members of the ski club and students who are members of the literature club.
- To secure affordable housing, the two groups must pool their resources and agree on lodging accommodations that will neither be in the mountains nor within an urban cultural center.
- Each group will propose a travel itinerary that, in their minds, strike the best compromise.
- Both groups get together and compare their proposals.
- After negotiating a final decision, the students will generate an itinerary and plan for the study abroad trip.

Activity 6. Cumulative Assignment
Grade level: College
Proficiency level: intermediate to advanced
Functions: Researching information, skimming, scanning, expressing feelings and emotion, finding common ground, negotiating, shifting of perspective
Forms: Major time frames, indirect speech.

This activity immerses the student into the 2019 Chilean protests, also known as *estadillo social* (social uprising). It consisted of a series of demonstrations that started in the capital, Santiago, and expanded to the entire country. The four-day civil unrest was a result of the marked inequalities in the society, the high cost of living, university graduate unemployment, neo-liberal policies, and corruption scandals. As a result, the country suffered the destruction of major means of transportation, businesses, and infrastructure. The president called for a state of emergency and the armed forces took action to prevent more destruction. The army was blamed for the use of excessive force leading to the deaths of 29 protesters and injuring of 2,500 others. After the protests, President Piñera announced a series of measures to improve living conditions, namely, better pensions and healthcare benefits, raising the minimum wage, price controls on energy costs, a new tax for higher-income earners, and a reduction in the salary of elected officials (Wikipedia). The movement also led to a process for the writing of a new Chilean constitution, which is still ongoing.

This cumulative activity intertwines all strategies as part of a roleplay simulation activity, creating situations that will induce the exercise of relational empathy, thus leading to better communication. This activity can be shortened and used during a class period or expanded and used during the semester over the course of a few weeks. In addition, these examples are aligned with Broome's (2017) five guiding principles for the promotion of relational empathy.

Each student impersonates a character involved in this protest with the goal of decentering their cultural beliefs and recognizing the socio-cultural background of their character. To help students learn about the characters' stories and what led them to hold certain ideas (without judging them based on their own feelings and experiences), we decided to flip the usual flow of an activity and introduce the characters' stories without alluding to the social uprising. A few notes before diving into the activity: While many of the characters are real, this type of activity leads to a certain degree of fictionality and dramatization of the event. This activity

can serve as a model that can be adapted to other social justice topics relevant to the L2 classroom.

Step 1. Dynamic Worlds

This activity serves to familiarize students with the topic of the Chilean social uprising. To this end, each student is assigned a character they will impersonate. Some sample characters are given in Table 1. The last column of the table is included to show the involvement of the characters.

- Students conduct an online search about the character they represent.
- They draw a timeline of the life of their character, highlighting the most impactful events in their life.
- They write a list of their characters' personalities, cultural beliefs, and social views.

These steps enable the students to better impersonate the characters in their complexity and thus shape their understanding as dynamic and provisional (guideline 1).

Table 1
Sample Characters of the Activity

	Name	**Profession**	**Involvement**
Group 1 (Protesters)	PareMan	Student	Protester who became famous worldwide thanks to a picture of him holding a stop sign going viral.
	Tía Pikachu	Kindergarten teacher	Protester who became famous for dancing and singing dressed up as Pikachu.
	Gustavo Gatica	Student	Protester who was hit by a rubber bullet and blinded during the protests.
Group 2 (Government)	Sebastián Piñera	President of Chile	In charge of the government that raised the subway fare for the Chilean capital that sparked the protest, he announced a state of emergency to authorize the deployment of the Chilean Army in the street.
	Mario Rozas	*Carabinero* (Chilean police) general director	He was in charge of the police force that repressed the protests and was accused of human rights violations by various ONGs.

Group 3 (Not involved in the Uprising)	Julieta Rojas	*Peluquera* (Hairdresser)	The mother of a student attending the protests
	Gaspar Soto	University professor in Chile	A college professor with family members in favor and against the protest

Step 2. Big Worlds

The purpose of this activity is to guide students to consider the perspectives of different social actors in the Chilean uprising and their actions. Students are divided into groups based on their role in the uprising (e.g., protesters, members of the government, and Chileans not directly involved in the protest). Members of two groups interact and share their cultural/personal beliefs and views of the current situation in the country.

- Group 1 (protesters) and Group 2 (government) discuss with each other their motivations for their participation and reaction in the uprising, as well as their views that they think legitimize their position.
- Group 3 (not directly involved in the uprising) works with groups 1 and 2 to present and discuss their own views (concern for their relatives, their own safety, and the long-term prospects of the movement).
- Students write a reflection on the personal reality of the main actors, their social context, and their involvement. Each group (1-3) must indicate at least one point of view that they did not consider before from a different group.

These steps will enable the students to focus on the bridging of differences (guideline 2).

Step 3. Shifting Worlds

In this activity, students must re-enact a situation in which the characters must decenter themselves and bridge differences. A sample scenario for roleplay is presented in Table 2.

Table 2

Scenario

The *estallido* (social uprising) has escalated and the city infrastructure is being burned and vandalized.

Participants	**Prompt**
Julieta Rojas, mother	Julieta is worried that her son participating in the protest would be hurt and that her downtown hair salon could be damaged.
PareMan, protester	The protester has lost hope for durable change and has begun throwing Molotov bombs on the main square.

- Julieta goes to *La Plaza de Armas* to find her son, who is involved in the protests. She approaches PareMan to find information about her son while criticizing what he is doing.
- Students role-play the interaction.

- In a reflexive assignment, students are asked to provide reasons that explain the positions and actions taken by their counterpart.

In this step, students impersonate a shift of perspective from their own experiences (guideline 3).

Step 4. Real Worlds

The teacher talks about the uprising, how it escalated, the subsequent deployment of the armed forces and the declaration of state of emergency, and the excesses that took place during the repression of the movement.

- Students watch a YouTube video depicting the events (Guardian News, 2019). This original footage shows the violence and distress of the protests in a way that stirs the emotions of the viewers.
- Based on this new perspective, students are asked to reflect on the meaning and impact of the uprising on the lives of the groups they represent.

In this step, the students gain an interpretation of the events in relation to its context (guideline 4).

Step 5. A Better World?

The uprising has eventually led to a series of actions for the writing of a new Chilean constitution, a process that is still ongoing today. The teacher brings this information to the classroom and asks the groups to move forward in time and become the participants of a consultative initiative aimed at defining the process for the writing of a new constitution.

- Participants are asked to research larger aspects of Chilean society whose legitimate claims were not featured prominently in the protests, such as the rights of indigenous people.
- In view of the new perspective gained from the activities, which segment of society should be asked to join a constitutional assembly?
- The three original groups provide recommendations for priorities to be considered by the constitutional assembly.

In this last step, the students strive for a synthesis of perspectives (guideline 5).

Throughout these exercises, it is noteworthy that the activities include elements that can make the classroom environment more favorable to L2 learning. Indeed, the very nature of activities requiring students to shift their perspective, rethink their understanding, find agreements, and synthesize point of views, all within the framework of dynamic interactions and with the aim of improving relationships, is bound to be quite engaging and take the students away from negative emotions that may otherwise be detrimental to the learning of an L2. In other words, the outcome is to induce flow, a concept stemming from the field of positive psychology (MacIntyre & Ayers-Glassey, 2022), which is considered an enabler of an especially deep-seated form of motivation. What is more, these examples provide students with opportunities to practice the L2 in a context that recreates authentic social and cultural conditions with respect to the target language. This is consistent

with Handford's view (2016): "Learning takes place in particular contexts of use, which in turn influence the language used and its appropriateness; learning the socially informed constraints and opportunities of particular situations is seen as fundamental to successful acquisition" (p.152).

Conclusions

This article explored the notion of relational empathy, as proposed by researchers in the field of intercultural communication and promoted the idea of transposing this concept to an L2 learning environment consistent with the views of a post-method condition in language teaching. These considerations reinforce the premise that, in addition to playing a crucial role in promoting social justice, the exercise of empathy makes positive contributions to the learning of an L2. The key element to this insight is to consider empathy as a factor which facilitates improved communications through better interactions, thus resulting in a more rewarding learning experience. Accordingly, examples of activities were provided to implement these ideas: 1) by seeking to explore different social and cultural realities relevant to the target language, whose themes are selected based on the teacher's own experiences while capitalizing on the students' unique cultural and linguistic backgrounds, and 2) by focusing on developing empathic interactions between students in accordance with Broome's five guiding principles, whose aims are not to merely transpose oneself into somebody else's thoughts and feelings, but rather to seek a shared understanding of experiences and values. While this approach was presented within the framework of a post-method condition in L2 instruction, it should be stressed that the use of relational empathy in the classroom could also be used alongside other teaching methods that put an emphasis on the fostering of social justice, such as Critical Content-Based Instruction (Sato et al., 2017) and social justice informed teaching (Wassell et al., 2019). In conclusion, teachers may find in relational empathy a pathway towards developing better interpersonal communication skills in accordance with the guidelines of the *World Readiness Standards* (National Standards Collaborative Board, 2015) and the NCSSFL-ACTFL Can-Do statements for Intercultural Communication (NCSSFL; Partnership for 21st Century Skills, 2011). Furthermore, in pursuing this goal, teachers may also value the idea that these empathetic practices serve their students throughout their lives in the service of building a more just society.

References

Allwright, R. L. (2003). Exploratory practice: Rethinking practitioner research in language teaching. *Language Teaching Research, 7*, 113-141. https://doi.org/10.1191/1362168803lr118oa

Arnold, J., & Foncubierta, J. M. (2022). The humanistic approach. In T. Gregersen & S. Mercer (Eds.), *The Routledge handbook of the psychology of language learning and teaching* (pp. 30-47). Routledge.

Bandura, A. (1976). *Social learning theory.* Prentice Hall.

Bennett, M. (1979). Overcoming the golden rule: Sympathy and empathy. *Annals of the International Communication Association,* 3(1), 407-422 ,https://doi.org/10.1080/23808985.1979.11923774

Breen, M. P., Hird, B., Milton, M., Oliver, R., & Thwaite, A. (2001). Making sense of language teaching: Teachers' principles and classroom practices. *Applied Linguistics, 22(*4), 470-501. https://doi.org/10.1093/applin/22.4.470

Broome, B. (1991). Building shared meaning: Implications of a relational approach to empathy for teaching intercultural communication. *Communication Education, 40,* pp. 235-248. https://doi.org/10.1080/03634529109378847

Broome, B. (2017). Intercultural empathy. In K. Yun (Ed.), *The international encyclopedia of intercultural communication,* (pp. 2-6). Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118783665

Brown, D. (1973). Affective variables in second language acquisition. Language Learning. https://doi.org/10.1111/j.1467-1770.1973.tb00658.x

Brown, D. (1987). *Principles of language learning and teaching* (2nd ed.). Pearson Education.

Caruso, D., & Mayer, J. (1998). A measure of emotional empathy for adolescents and adults. Unpublished manuscript. https://scholars.unh.edu/cgi/viewcontent.cgi?article=1021&context=personality_lab

Cheng, L. (2017). *Intercultural communication.* De Gruyter, Inc.

Davidson, D. E., Garas, N., & Lekic, M. D. (2021). Transformative language learning in the overseas immersion environment: Exploring affordances and intercultural development. In B. L. Leaver, D. E. Davidson, & C. Campbell (Eds.), *Transformative language learning and teaching* (pp. 109-119). Cambridge University Press.

Deardoff, D. (2006). Process model of intercultural competence. *Journal of International Education*, 10, 241-266.

Demetriou, H. (2018). *Empathy, emotion, and education.* Palgrave Macmillan.

DeTurk, S. (2001). Intercultural empathy: myth, competency, or possibility for alliance building. *Communication Education, 50(4),* pp. 374-384. https://doi.org/ 10.1080/03634520109379262

Dewaele, J. M. (2022). Personality. In T. Gregersen & S. Mercer (Eds.), *The Routledge handbook of the psychology of language learning and teaching* (pp. 112-123). Routledge.

Dörnyei, Z. (2009). *The psychology of second language acquisition.* Oxford University Press.

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. Routledge.

Falout, J. (2014). Circular seating arrangements: Approaching the social crux in language classrooms. *Studies in Second Language Learning and Teaching, 4*(2), 275– 300. https://doi.org/ 10.14746/ssllt.2014.4.2.6

Freire, P. (1970). *Pedagogy of the Oppressed*. Continuum.

Fukada, Y., Murphey, T., Falout, J., & Fukuda, T. (2017). Essential motivational group dynamics: A 3-year panel study. In R. Breeze & R. Sancho Guinda, C. (Eds.), *Essential competencies for English-medium university teaching*. Educational Linguistics, 27. Springer https://doi.org/10.1007/978-3-319-40956-6_17

García Lorca, F. (2017). *Antología de conferencias de Lorca: Charla sobre teatro, teoría y juego del duende, Las nanas infantiles*. Independent Publisher.

Grin, F., & Faniko, K. (2012). Foreign language skills and intercultural abilities: operationalization with a large population. *Revue Management & Avenir,* 55, pp.168-184. https:// doi.org/10.3917/mav.055.0168

Guardian News. (2019, November 5). Tens of thousands protest in Chile: We've reached a crisis. [Video]. YouTube. https://www.youtube.com/watch?v=VMYM00i6_L8

Guiora, A.Z., Brannon, R. C. L., & Dull, C.Y. (1972). Empathy and second language learning. *Language Learning, 22*(1), pp.111-130. https://doi.org/10.1111/j.1467-1770.1972.tb00077.x

Handford, M. (2016). The dynamic interplay between language and social context in the language learning classroom: Interpersonal turn-taking ELF learners. In J. King (Ed.), *The dynamic interplay between context and language learner*, (pp. 152-171). Palgrave MacMillan.

Hoffman, M. L. (2000). *Empathy and moral development. Implications for caring and justice.* Cambridge University Press.

Howe, D. (2013). *Empathy: What it is and why it matters.* Palgrave McMillan.

Horwitz, E. (2010). Foreign and second language anxiety. *Language Teaching*, 43(2), 154-167. https://doi.org/10.1017/S026144480999036X

Horwitz, E. (2013). *Becoming a language teacher*. Pearson.

Kumaravadivelu, B. (2006). *Understanding language teaching*. Routledge.

Lantolf, J. P. (2000). Second language learning as a mediated process. *Language Teaching, 33* (1), 79–96.

Leaver, B. L. (2021). Transformative language learning and teaching: The next paradigm shift and its historical context. In B. L. Leaver, D.E. Davidson, & C. Campbell (Eds.), *Transformative language learning and teaching* (pp. 13-22). Cambridge University Press.

Leontiev, A. N. (1978). *Activity, consciousness, and personality.* Prentice Hall.

Long, M. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics,* 4(2), 126-41. https://doi.org/10.1093/applin/4.2.126

MacIntyre, P., & Ayers-Glassey, S. (2022). Positive psychology. In T. Gregersen & S. Mercer (Eds.), *The Routledge handbook of the psychology of language learning and teaching* (pp.61-73). Routledge.

Mihaly, D. (2021). Empathy in Spanish instruction. *Hispania*, *104*(3), pp. 352-355. https://doi.org/10.1353/hpn.2021.0088

Mercer, S. (2016). Seeing the world through your eyes: Empathy in language learning and teaching. In P. MacIntyre, T. Gregersen, & S. Mercer, (Eds.), *Positive psychology in SLA* (pp. 91-111). Multilingual Matters.

Murphey, T., Fukada, Y., Fukuda, T., & Falout, J. (2022). Group dynamics. In T. Gregersen & S. Mercer (Eds.), *The Routledge handbook of the psychology of language learning and teaching* (pp. 285-299). Routledge.

National Standards Collaborative Board. (2015). *World-readiness standards for learning languages*. (4th ed.). ACTFL.

Neruda, P. (2020). *Los veinte poemas de amor y una canción desesperada*. Editorial L'Aleph.

NCSSFL-ACTFL(2017). *Can-Do statements for intercultural communication.* https://www.actfl.org/educator-resources/ncssfl-actfl-can-do-statements

Neuliep, J. W. (2017). *Intercultural communication: A contextual approach*. 7th Edition. Sage.

Ochs, E. (1996). Linguistic resources for socializing humanity. In J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 407-437). Cambridge University Press.

Olivares-Cuhat, G. (2012). Does empathy make a difference in the foreign language classroom. *Alicanto. Journal of Literature, Linguistics and Education*, *5*, 62-72. http://www.revistahumanidades.uda.cl

Oxford, R. (2016). Toward a psychology of well-being for language teachers: The EMPATHICS vision. In P. MacIntyre, T. Gregersen, & S. Mercer, (Eds.), *Positive psychology in SLA* (pp. 10-87). Multilingual Matters.

Partnership 21st Century Skills Map. (2011) https://www.actfl.org/sites/default/files/resources/21st%20Century%20Skills%20Map-World%20Languages.pdf

Salovey, P., & Mayer, J. D. (1990). Emotional Intelligence. *Imagination, Cognition and Personality*, *9*(3), 185–211. https://doi.org/10.2190/DUGG-P24E-52WK-6CDG

Sato, S., Hasegawa, A., & Kumagai, Y., & Kamiyoshiu, U. (2017). Content-based instruction (CBI) for the social future: A recommendation for critical content-based language instruction (CCBI). *L2 Journal*, 9(3). https://doi.org/10.5070/L29334164

Swain, M. (1985). Communicative competence: some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in Second Language Acquisition* (pp. 235-56). Newbury House.

Thorne, S. L., & Lantolf, J. P. (2007). Sociocultural theory and second language learning. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 201-224). Lawrence Erlbaum.

Van der Zee, K. I., & Van Oudenhoven, J. P. (2000). The multicultural personality questionnaire: A multidimensional instrument of multicultural effectiveness. *European Journal of Personality, 14*, pp. 291-309. https://doi.org/10.1002/1099-0984(200007/08)14:4<291::AID-PER377>3.0.CO;2-6

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Wassell, B., Wesely, P., & Glynn, C. (2019). Agents of change: Reimagining curriculum and instruction in world language classrooms through social justice education. *Journal of Curriculum and Pedagogy,* 16, 263-283.

Wells, G. (1999). *Dialogic inquiry: Towards a sociocultural practice and theory of education.* Cambridge University Press.
Wikipedia. 2019-2022 Chilean protests.

Appendix

Resources for *Estallido Social en Chile*

A tres años del estallido social, Gabriel Boric pide unión a los chilenos. https://elpais.com/internacional/2021-10-18/chile-en-un-mar-de-incertidumbre-a-dos-anos-del-estallido-social.html

Convención constitucional. https://es.wikipedia.org/wiki/Convenci%C3%B3n_Constitucional_(Chile)

Chile, en un mar de incertidumbre a dos años del estallido social. https://elpais.com/internacional/2021-10-18/chile-en-un-mar-de-incertidumbre-a-dos-anos-del-estallido-social.html

Estallido social. https://es.wikipedia.org/wiki/Estallido_social

La cronología del estallido social de Chile. https://www.dw.com/es/la-cronolog%C3%ADa-del-estallido-social-de-chile/a-51407726

La histórica resistencia mapuche en Chile y el estallido social actual. https://ladiaria.com.uy/chile/articulo/2020/8/la-historica-resistencia-mapuche-en-chile-y-el-estallido-social-actual/#:~:text=Cualquier%20observador%20del%20estallido%20social%20que%20se%20inici%C3%B3,la%20covid-19%20pusieron%20en%20pausa%20a%20las%20movilizaciones.

2

Academic Integrity in the Virtual Language Class

Jessica Sertling Miller
University of Wisconsin-Eau Claire

The Challenge

How can language educators promote academic integrity when machine translation is ubiquitous and easy to use? How can they empower all students to make informed choices regarding the resources available to them? Instructional design models that are equity-based can reduce factors contributing to misconduct while positively impacting success.

Abstract

Many factors motivate academic dishonesty in language classes, including stress and easy access to online translators. While instructors may look for ways to deter cheating, policing learners may create anxiety and negatively affect their sense of agency. Creating conditions that limit factors contributing to misconduct instead of implementing punitive measures can lead learners on an aspirational path to uphold standards of academic integrity. Such approaches are described and discussed through the lens of instructional design. They are applicable to all types of language classes, and especially suitable to asynchronous online courses. First, machine translation technologies are considered. Then, academic integrity is connected to accessibility and issues of equity. Equity-based design ideas that align with standards of accessibility and support proficiency development are shared. Specifically, principles of Universal Design for Learning and the Transparency in Learning and Teaching can help all students maintain integrity by removing obstacles and reducing stress factors. Those models are unpacked and practical advice for language educators is provided.

Keywords: academic integrity, inclusive practice, online machine translation, online teaching, pedagogy, transparency, universal design

Academic Integrity in the Virtual Language Class

Language educators strive to offer growth opportunities for students by creating and curating course content while carefully choosing how material is presented. Ensuring that learners are exposed to language that is slightly beyond their level yet comprehensible is part of Krashen's well-established *i*+1 Input Hypothesis (1985) that has informed language education for decades. Still, students sometimes produce work at a proficiency level beyond the +1 language input modeled in class. This is a typical sign that machine translation (MT) technology was used and that students submitted MT output as their own. Jolley and Maimone (2022) reviewed research on MT published in the last 30 years and concluded that, although many educators consider it to be unethical, its use by students is unavoidable. O'Neill (2019) found that 87.7% of students report using MT even when the technology is prohibited in academic integrity policies.

Despite the frequency of academic misconduct, instructors in higher education rarely pursue sanctions. Lynch et al. (2021) researched academic integrity in nursing programs and found that teachers feel not only ill-equipped to detect cheating with certainty, but also do not believe punishment is worth the added workload. In most cases, they prefer to handle cases on their own by considering the students' contexts and using restorative approaches to improve future behavior. Robinson and Openo (2021) corroborated that faculty feel emotionally exhausted and deeply affected by the process of dealing with academic dishonesty, compounded by the stress they see their students experiencing. Bretag (2013) supports moving beyond punishment to instead focus efforts on supporting a culture of integrity. As for MT specifically, Jolley and Maimone (2022) identified two mindsets: (1) educators who equate MT use with cheating tend to recommend detecting it, addressing it when it occurs, and adapting to prevent it from happening; (2) those who regard MT as a tool capable of improving language quality usually recommend integrating it routinely in the curriculum.

This paper supports adapting course design to reduce unwarranted MT use (mindset 1), but also argues for normalizing MT use (mindset 2) as a resource, with the acknowledgement that it can have both positive and negative impacts on learning. Beyond issues of workload and pedagogy, this seemingly ambivalent position is informed by considerations of Equity, Diversity, and Inclusion (EDI) which is a concern at the mid-size liberal arts public university in the American Midwest where I have been teaching French for over 15 years. O'Neill (2013) showed that language teachers successfully identify the use of MT in their students' work 70.7% of the time. Innes (2019) found the average rate to be 74.04%. Those numbers imply that false positives in MT identification occur about two to three times out of ten. Were academic misconduct policies followed strictly, at least two students out of ten would be wrongly accused, with substantial ramifications.

This paper proposes strategies to communicate more transparently and effectively with students on matters of academic integrity and MT use in particular.

MT use can be normalized, but also reduced through course design. This approach is motivated by the need for action to uphold individual and institutional EDI commitments and driven by the recent call to apply UDL principles for a more inclusive approach to academic integrity (Davis, 2022). These approaches are especially relevant to language teachers in higher education who have some control over their classroom policies, assessment types, and Learning Management Systems.

Literature Review

Davis (2022) reported high levels of anxiety in students under investigation for misconduct, affecting their sense of belonging in their academic setting. Furthermore, certain groups such as international students, first-generation students, students whose native language is not that in which school policies are written, students of color, Indigenous students, are overrepresented in investigations related to academic conduct (Davis, 2022; Eaton, 2020; Eaton 2022; Robinson & Openo, 2021; Sanni-Anibire et al., 2021). Eaton (2020, 2022) warned that this bias places racialized minorities under disproportionate scrutiny, and therefore they are more at risk of being reported for misconduct. That is a concern that needs to be addressed urgently.

An Inclusive Approach to Academic Integrity

Reedy et al. (2021) suggested increasing comprehension of academic integrity policies by making them more accessible to diverse learners with, for instance, creating flow charts for visual interpretation. In the same spirit, Davis (2022) recommended following Universal Design for Learning (UDL) principles for official documents to be more inclusive and easier to interpret. She suggested that cases of suspected misconduct could be reduced by removing legalese from documents relating to academic conduct policies. That could positively impact a large number of students: those whose native language is different from the dominant language at their institution; those with reading impairments; anyone too overwhelmed by that type of language to engage with it. Khoo and Kang (2022) helped English language learners avoid academic integrity violations and saw improvement in their written proficiency with a four-week online writing course during which students were empowered to practice ethical behaviors in risk-free journaling tasks with personalized feedback. A salient piece of advice they offer is to give learners agency through low-stake practice that will increase their confidence, thus reducing their perceived need for outside help.

In the virtual language class, one strategy is to improve instructional design. Adopting accessibility guidelines when designing the syllabus with policies on integrity, fine-tuning pages in the Learning Management System (LMS -e.g., Blackboard Learn, Canvas, D2L Brightspace, Moodle), allowing for multiple access points to course content, and providing transparent assessment prompts are a few strategies discussed. Frameworks grounded in research are described – UDL and Transparency in Learning and Teaching (TiLT) specifically. Concrete advice and examples based on those models are provided. Instructors who have ownership

of the material shared on the LMS will find these recommendations immediately applicable. For educators that are part of a larger team, these approaches will foster much-needed conversations on how to provide consistency across classes to improve students' experiences and strengthen their language programs.

Machine Translation Technology

Many free tools are available online to assist language students: dictionaries, conjugators, translators, and communities of multilingual people. There is nothing inherently wrong with using available resources. Among that wide array of resources, this paper focuses on MT because it is ubiquitous and easy to use to translate entire sentences and paragraphs. The quality of the language produced when students use MT technologies has been shown to exceed that produced without them, but these immediate gains do not transform into durable proficiency development (Hellmich, 2021; Lee, 2020; Niño, 2020; O'Neill, 2019). It appears that students do not retain the information they translate. Unlike online dictionaries (e.g. www.Larousse.fr) in which students first need to input a word in the target language to obtain a definition in the same language, or online translation dictionaries (e.g., WordReference) that translate single words and provide contextualized sentences with grammatical information, MT technology (e.g. Google Translate) allows them to input text of any length in one language and receive a translation that they effortlessly copy and paste into their work.

Instructor Attitudes about MT

Determining whether MT is a friend or a foe is never straightforward and always context-dependent (Correa, 2014; Groves & Mundt, 2015; Klekovkina & Denié-Higney, 2022). Language teachers now generally advocate for training students in MT best practices (Clifford et al., 2013; Ducar & Schocket, 2018; Jolley & Maimone, 2015; O'Neill, 2019; Pardo-Ballester, 2022; White & Heidrich, 2013). Similarly, academic integrity scholars call for increased guidance to help students learn and practice ethical behaviors in their disciplines (Bretag, 2013; Bretag, 2019; Çelik & Lancaster, 2021; Davis, 2022; Eaton, 2022; Khoo & Kang, 2022; Luck et al., 2022; Sefcik et al., 2020). The type of training, however, will likely differ from class to class as instructors have personal views on MT and academic integrity. According to Clifford et al. (2013), 42% of instructors surveyed equated any use of MT as cheating, 21% as not cheating, and 37% opted for the "other" category, citing that context and frequency of use muddied boundaries. When investigating the type of usage that is ethical vs. unethical, Jolley and Maimone (2015) found that 87.18% of surveyed instructors deemed MT use for individual words to be ethical, with exactly as many stating that translating paragraphs or entire texts is unethical. If instructors cannot themselves consistently establish the border between ethical and unethical use of MT, how can learners be expected to abide?

Neural Machine Translation

In 2016, Google Translate upgraded to Neural Machine Translation (NMT), a deep learning system based on an artificial neural network that continually improves the accuracy of its translations by modeling its output on natural

language, taken from millions of examples available online and from user queries. Ducar and Schocket (2018) noted that with the advance of technology, use of online translators will be easiest to detect not through translation oddities, but because the production will be far above the expected proficiency level. Klimova et al. (2022) call for more experiment research involving NMT since its use is unavoidable; they also think it can potentially help learners if used systematically.

In November 2022, the artificial intelligence lab OpenAI launched *ChatGPT*, a chatbot service that can answer questions of any kind convincingly, summarize, find errors, and, among other things, generate coherent academic papers (Bowman, 2022). Google is currently developing similar technology (Metz, 2023). Such powerful programs have the potential to grow into useful tools to facilitate access to knowledge but also to spread misinformation because falsehoods are created when the bot does not know an answer (Bowman, 2022; Metz, 2023). ChatGPT immediately turned into a pathway to academic misconduct. Students have already been caught submitting ChatGPT output as their own (Cassidy, 2023). In reaction to that, a Princeton student released an app capable of identifying whether a paper was written by ChatGPT (Bowman, 2023). It is unclear if the bot can be asked to write in a language other than English, but the situation is evolving at the speed of digital light so language educators need to stay informed.

Academic Integrity by Design

What educators can focus on instead is designing online modules that put learners on an ethical path. In a Teaching in Higher Ed podcast interview (Stachowiak, 2017), academic integrity scholar Newton emphasized the importance of assessment design in reducing academic misconduct. Among other things, he mentioned enabling students to show what they learned, testing the targeted skills (e.g., narrowing the scope of an exam, not mixing different competences in the same task) and doing so in multiple low-stake assessments to relieve stress and lighten the workload at the end of the term, which was already suggested by McAllister and Watkins (2012). Applied to a virtual language course component, this might look like short weekly assignments in varied formats (e.g., recordings, written submissions, concept maps) with clear directions relative to MT. A scoring rubric would be provided ahead of time, and room for language errors should be evident in the fact that the maximum score should be reachable despite mistakes, according to the targeted proficiency level. Bretag et al. (2019) explained that assessments should be designed with academic integrity in mind. An in-depth discussion with examples of such designs follows.

Focus on Equity

Eaton (2022) explained that there are three established lenses through which academic misconduct is typically framed: it is seen as a moral issue, a policy issue, and a pedagogical issue. She proposed to add a fourth: equity. She remarked that there can be no integrity without equity and declared: "Promoting academic integrity is about more than upholding rules and policies, especially when they perpetuate systems of privilege for some and oppression for others" (pp. 6-7). She

referred to barriers experienced by equity-deserving groups that are overrepresented in misconduct reporting (e.g., international students, first-generation students, students who do not master the dominant language, students of color, Indigenous students).

Eaton's call to action is: "Make equity, diversity, inclusion, decolonization and Indigenization an academic integrity imperative" (p. 9). She supports applying principles of UDL as described by Davis (2022) to make learning material more accessible and comprehensible. Gierdowski and Galanek (2020) found that 19% of undergraduate students in the US disclosed a disability in 2015 and suspect that the actual number is higher as many who could declare a disability choose not to do so. Prioritizing accessibility would have a wide-ranging positive impact on learning and integrity for those needing accommodations, whether the need is disclosed or not, as well as for the equity-deserving groups that are overrepresented in misconduct reports. Ultimately, everyone benefits from clear instructions that do not make assumptions about learners' prior experiences, abilities, or access to resources.

One concrete recommendation for online language classes is to be more intentional in the design of the material educators share via their Learning Management System (LMS). In a virtual course the LMS is the classroom and the point of interaction between students and teachers. Paying close attention to its design is an effective way to facilitate communication, remove barriers, and reduce stress. McAllister and Watkins (2012) emphasized a decade ago that the design of online courses can influence student success and minimize dishonesty. Since then, research-based guidelines and templates have been developed to facilitate such redesigns.

Universal Design for Learning

The UDL guidelines (CAST, 2018) are freely available online at www.udlguidelines.cast.org. They aim to provide educators in any discipline with "a framework to improve and optimize teaching and learning for all people based on scientific insights into how humans learn". They offer concrete suggestions to apply design changes so that all students can engage with the learning material, anticipating needs that may be overlooked by educators who do not experience the same barriers. Instead of addressing needs as they arise (e.g., when a student requires accommodations), UDL strives to eliminate them. Consequently, learners who prefer not to formally request accommodations, or who are not aware of the accommodation process, or who may not know that they qualify for it, encounter as few obstacles in their learning as possible. In the end, everybody benefits from these design changes as barriers to learning are torn down. Academic integrity becomes easier to maintain because learners can focus on completing the task instead of spending time and energy trying to follow directions that are confusing to them, that minimize their agency, or that are not scaffolded in manageable steps. UDL differs from the Americans with Disabilities Act (ADA) because it is not legislation. The ADA protects the civil rights of people with disabilities, guarantees equal opportunity in accommodations, and defines accessibility standards used in regulation. UDL guides best practices and supports educators who want to create an accessible and equitable learning environment, even in situations where the ADA would not mandate it.

As an aid to teachers willing to apply UDL principles, CAST (2018) developed a detailed graphic organizer that encompasses a variety of concepts meant to represent the needs of all learners. Each section of the organizer contains several key ideas, labeled "checkpoints." Every checkpoint is clickable and leads to further information backed by

research. Because the organizer is comprehensive, users are encouraged to focus on one specific challenge or goal at a time, as opposed to try to tackle everything at once. Table 1 shows a simplified version of the UDL graphic organizer with concrete examples found within each checkpoint. These examples were selected here because they are relevant to equity-focused academic integrity in the virtual language classroom.

Table 1

Adapted UDL Graphic Organizer

	Provide multiple means of…		
	Engagement (why)	Representation (what)	Action and Expression (how)
Access	Provide choices Invite personal responses Vary the level of novelty or risk	Display information in a flexible format Provide written transcripts for videos Provide descriptions for all images	Provide alternatives to marking with pen/mouse Select software that works seamlessly with keyboard alternatives and alt keys
Build	Encourage division of long-term goals into short-term objectives Emphasize process, effort, improvement Support opportunities for peer interactions Provide feedback that encourages perseverance	Embed support for vocab within the text Make connections to previously learned structures Allow the use of Text-to-Speech Embed visual supports for vocab (pictures, videos) Provide electronic translation tools or links to multilingual glossaries	Compose in multiple media such as text, speech, drawing, comics, film Provide spellcheckers, grammar checkers, word prediction software Provide differentiated models to emulate
Internalize	Provide prompts, rubrics that focus on self-regulatory goals Provide scaffolds and feedback for managing frustration Give feedback on progress in a manner that is understandable and timely	Activate relevant prior knowledge Highlight or emphasize key elements Give explicit prompts for each step Provide checklists, organizers, electronic reminders	Provide scaffolds to estimate effort, resources, and difficulty Embed prompts to "stop and think" before acting Provide templates for organizing information Show representations of progress

Applying UDL guidelines does not mean revamping an entire LMS course. Altering the presentation of policies slightly or the structure of assessment prompts are small, durable, and replicable investments with the potential to be

deeply impactful. In a *Teaching in Higher Ed* podcast interview (Stachowiak, 2018), Tobin, an expert in technology-mediated education, emphasized that the UDL framework does not decrease the rigor of a course content: it lowers the barrier to access that rigorous content. Davis (2022) finds UDL principles advantageous because they reframe inclusion as universal. When these guidelines are heeded, materials are accessible to all from the beginning and do not need to be adapted for a few students later. Consequently, even those who prefer not to disclose a disability will see their needs better met.

Applications in the Virtual Language Class

Stress is a factor contributing to breaches in academic integrity (Eaton, 2022). Clicking through the LMS and being unable to locate a document crucial to successfully complete an assignment is time-consuming and frustrating. Trying to understand lengthy policies on academic integrity written in jargon that may sound intimidating and punitive can induce anxiety (Khoo & Kang, 2022; Sanni-Anibire et al., 2021). McNeill (2022) recommends moving away from language associated with disciplinary procedures (e.g., cheating, misconduct) in favor of aspirational terms (e.g., academic purpose) linking academic integrity to educative approaches in order to build and meet community standards. In this section, the checkpoints provided in the aforementioned UDL graphic organizer are connected to steps educators can take toward a more positive and inclusive strategy to promote academic integrity.

The LMS-Based Syllabus

UDL guidelines recommend creating a safe space for learners (UDL graphic organizer checkpoint 7.3) and fostering a sense of community (checkpoint 8.3). The LMS, even more so in an online class, is a good place to share practical information with students on academic integrity best practices (Luck et al., 2022). Gierdowski and Galanek (2020) identified among their key findings that students with disabilities consider the LMS to be an essential element to access course content and would like their teachers to use it more. With the support of UDL guidelines, an LMS-based syllabus becomes a tool to promote academic integrity by granting open access to resources that help students make choices that are in their interest.

Since the LMS plays a central role in today's educational settings, instructors need to be intentional about the form and content of the material they post if they want students to engage with it. Stating policies clearly in a location that is easy to find is important. A syllabus created in the LMS rather than as one file emailed to students or buried among other documents in the LMS can fulfill that role adequately. The LMS-based syllabus can take the form of a collapsible module/unit with separate pages for each topic (e.g., course description, topical outline, attendance, rubrics, academic integrity) at the top of the LMS course. Students can see and access it anytime they land on the course site and browse its subsections quickly. The syllabus transforms from a document consulted once and put away on the first day of classes into a reference guide visible throughout

the course. Following accessible standards for digital text, providing descriptions for images (checkpoint 1.3), and minimizing verbiage so that students readily find the information for occasional reference can increase and maximize interactions with the syllabus. A clearly written policy on academic integrity that is practical and convenient to locate is more likely to be accessed and processed. The practices of organizing information sequentially into smaller elements and avoiding 'information dumping' that will deter absorbing the content align with UDL principles on guiding information processing (checkpoint 3.3).

An LMS-based syllabus also makes it easy for teachers to create and schedule reminders. Repeating important policies ensures that the information stays fresh, especially as rules often differ for each class and each instructor, including those regarding MT use. UDL principles support highlighting goals consistently by using reminders that focus learners' attention (checkpoint 8.1). Academic integrity policies and practices deserve to be revisited regularly, even for just a minute. In the LMS, instructors can quickly convert key sections of the syllabus into announcements, thus reiterating select policies throughout the semester, as long as those reminders are limited to a few essential topics lest students get into the habit of dismissing them should they be too frequent. These announcements can also take the form of brief video messages (i.e., a minute or less) in an effort to use multiple media for communication (UDL checkpoint 5.1).

Transcripts and Subtitles

In the virtual language class, one UDL principle that supports language learning is to present information in different modalities (checkpoint 5.1). Written material often outweighs audio/video material because it is easier to manage and assess. UDL guidelines are reminders that educators' preferred styles are not adapted to all learners. Giving multiple entry points to the same content increases accessibility. For instance, many news sources publish short videos accompanied by transcripts or subtitles in the target language. Students can be invited to start with the video before moving on to the transcript/subtitles or vice versa, compare both versions, or simply choose one. If the goal is general comprehension of a news report, then the modality is of less importance; if given the choice, selecting a resource that provides options benefits all learners.

UDL guidelines recommend using transcripts and subtitles with videos (checkpoint 1.2). In some cases, textual support of audio/video material can be required to comply with ADA regulations. But language teachers who want to test audio comprehension may hesitate. Recent research on the effect of subtitles has shown no effect on proficiency (van der Zee et al., 2017), meaning that using them had no benefits and no drawbacks on language development. The authors recommend using subtitles to help students with hearing disabilities. Korucu-Kış (2021) analyzed existing research on captioning effects on L2 listening: findings so far have been inconclusive. She explained these results by citing factors that are difficult to control (e.g., background knowledge of listeners, type of media). Her final reflection agreed with that of van der Zee et al. (2017) in that captions do not necessarily help or impede comprehension. Another team of researchers

found a positive effect on vocabulary acquisition, especially with subtitles in the target language (Reynolds et al., 2022). Therefore, subtitles are tools that not only make videos more accessible but also do not obstruct language growth, and even have the potential to accelerate vocabulary acquisition. They should not be automatically discounted and should not be viewed as cheating.

When learners do not understand most of a video, perhaps because its level is beyond the *i*+1 comprehensible input (Krashen, 1985), they may experience stress, panic, or feel inadequate as they assume that they should understand more since the teacher is presenting this material. As mentioned already, stress and anxiety factor heavily on reasons to cheat (Eaton, 2022; Jolley & Maimone, 2022; Stachowiak, 2017). Subtitles and transcripts (even partial) can relieve some of the students' stress, with the understanding that they will sometimes be assessed on audio material with less support of that kind, and the assurance that it will be within their reach. In his *Teaching in Higher Ed* interview (Stachowiak, 2018), Tobin said that format choice should be given when format does not make a difference with the goal of the learning activity. However, he underscored that when format is part of the task requirement, it is appropriate to not offer a choice. Applied to language teaching and learning, this means that instructors remain free to use videos without subtitles if they deem it in alignment with the objectives. However, they should pick material that is level-appropriate and scaffolded to avoid student behavior that would go against their decision and violate academic integrity policies. Scaffolding could include providing vocabulary to help students express their thoughts on a particular topic, which would limit their perceived need to visit online translation sites.

Machine Translation

In the section on machine translation technology, the literature review established that MT use by students is inevitable, with or without nefarious intent. As technology advances, online translators' output becomes more accurate and perhaps more tempting for students. In light of these trends combined with factors that contribute to students rationalizing MT use even when prohibited (e.g., time constraints, unclear directions, cultural acceptance of MT), and keeping in mind recent calls to consider academic integrity as an equity issue (Davis, 2022; Eaton, 2022), a proposed solution to MT-based academic misconduct is to allow the use of online translators and to stop trying to detect it. As Klekovkina and Denié-Higney (2022) discussed the results of their study on MT, they concluded that mindset adjustments would be necessary for educators to make online translators allies.

That approach aligns with the ideas on academic integrity considered thus far and with UDL guidelines for a more equitable learning environment. While UDL checkpoint 2.4 supports access to online translation tools, that recommendation likely targets students outside of language courses who may not fully comprehend course content without linguistic assistance. The perspective of language educators differs but is not unrelated. Not understanding material with certainty is naturally unsettling. It can be a new experience met with negative emotions by monolingual students. When teachers respond to that emotional response by limiting the resources

that would provide comfort (e.g., access to subtitles, translators) and by increasing the threat of disciplinary action, they are likely creating further distress. Sefcik et al. (2020) examined academic integrity programs and noted that participants expressed the need to focus on values, skills, and outcomes instead of rules and sanctions. Authorizing MT use in language courses would shift the focus on values, skills, and outcomes, provided learners are made aware of the reasoning and allotted time to practice using MT. Furthermore, having MT as an option empowers learners as they have more control over what they want to communicate. Giving learners agency helps them uphold academic integrity (Khoo & Kang, 2022; Klekovkina & Denié-Higney, 2022). UDL principles also encourage optimizing individual choice and learners' autonomy (checkpoint 7.1). Providing support to use technologies effectively is also part of UDL guidelines (checkpoint 4.2). Applied to MT, this means that students should be shown how to access a teacher-recommended MT site, the best ways of using it for language development, and should be led to reflect on the advantages and disadvantages of MT as they practice working with and without the technology. The goal is for students to have the ability and agency to make an informed decision as to what they consider appropriate MT use given their specific context.

In conjunction with that training, mistakes should be normalized and even incentivized. Directions for all assignments should make it clear that communication at the target proficiency level will be successful despite language mistakes. Functional accuracy should bear more weight than grammatical accuracy to prevent learners from vilifying mistakes. A straightforward rubric showing that the maximum grade can be attained despite errors can relieve stress experienced by students who believe only perfection warrants a top grade. Part of the assignment could further include submitting a sentence that proved problematic, with points gained just for demonstrating struggle and resilience, similar to what Klekovkina and Denié-Higney (2022) described doing to limit MT usage, highlight the value of recycling familiar structures in creative ways, and reinforce comprehensibility, not accuracy, as a measure of success. Then, instructors can provide focused feedback on segments that learners identified as wanting to improve, thus increasing mastery-oriented feedback to sustain engagement and motivation (UDL checkpoint 8.4).

When communication, and not perfection, is the goal of a language assessment, test anxiety can be reduced because the focus is on what students can do, not on what they cannot do. Errors should be treated as valuable learning tools. Risk-taking should be celebrated for its enormous value in language learning and communication. Targeted corrective feedback combined with comments highlighting students' accomplishment and overall progress can turn assessments into meaningful resources and powerful motors for growth. Klekovkina and Denié-Higney (2022) linked the shift from MT-generated correctness to functional comprehensibility with students' agency: they can derive satisfaction and motivation when they realize that they do not depend on a machine to communicate their ideas.

Assignment directions should also include the instructor's recommendation on MT use for each task (e.g., encouraged, neutral, discouraged, prohibited) along with a rationale to remind students of the advantages and drawbacks. Finally, since students are routinely asked to cite their sources when writing academic papers,

indicating which machine translation technology service they used, if any, with a list of terms or phrases generated by MT, should become customary. Citing sources is an ethical academic practice; normalizing such MT citation and underlining or rewriting the output obtained via MT would additionally raise awareness on how much that tool was utilized to avoid dependency, and reinforce the new structures generated by MT that, if pertinent to students, should be memorized.

Transparency in Learning and Teaching

Another equity-minded approach is that of Transparency in Learning and Teaching (TiLT). That framework provides a helpful template to design assessments along some UDL principles, although it is not affiliated or explicitly tied to UDL. A description and resources are available on the TiLT Higher Ed website (www.tilthighered.com). The TiLT structure ensures that all essential elements of a prompt are included. This protocol has been shown to improve student success. As they design their assessment tasks instructors can simply fill out each category of the template with relevant information. There are three main areas to be addressed: Purpose, Task, and Criteria. The first describes why students are asked to complete it, the second lists all the steps required to complete it successfully, and the third describes the evaluation process. All those elements have been identified by research as significantly impactful on student success, especially equity-deserving groups (Winkelmes et al., 2019). The TiLT framework aligns with UDL because its Purpose/Task/Criteria template breaks down an entire assignment into smaller organized subsections, which supports planning development by dividing long-term goals into reachable short-term objectives (checkpoint 6.2), helps learners manage information and resources by providing checklists (checkpoint 6.3), and guides information processing by progressively and sequentially releasing the relevant information (checkpoint 3.3).

The TiLT template can be adapted in many ways, allowing instructors to customize it to their style and discipline. However, one pitfall to avoid is to give students too much information. At first, it may seem like the more details are provided, the better. Paring down content to what is absolutely necessary helps students focus and ensures that they read directions with their full attention instead of scanning the document. If the assignment becomes too long, educators should take it as a sign to break it down into multiple assignments. Under Purpose, instructors list the set of skills that learners will practice, what knowledge they will gain, and how those elements relate to students' life. The Task section is particularly helpful because it contains what students are expected to do and how to do it. That includes preparatory work (e.g., reading a specific chapter, reviewing notes), something which is often overlooked because considered tangent to the task itself. Yet preparatory work is neither intuitive nor insignificant. UDL guidelines emphasize the importance of priming background knowledge so that everyone starts the assignment on an equal footing (checkpoint 3.1). A note on the instructor's recommended use of MT as mentioned earlier would fit in that section. The Criteria piece is a checklist of expectations, with a grading rubric detailing the evaluation process. The checklist prevents oversights and helps students monitor

their progress (UDL checkpoint 6.4). Examples showing what type of work meets expectations and what type does not help clarify expectations. A subsection on academic integrity can be appended if desired by copying and pasting the statement in the syllabus, offering students a refresher on this important policy (UDL checkpoint 8.1, "heighten salience of goals and objectives").

Estimated Completion Times

The TiLT framework does not include estimated completion times for each task. Personally, I have found that adding that bit of information can be eye-opening for instructors and students alike. Preparatory work, reflection, completion of multi-step tasks, and proofreading should be taken into consideration when designing an assessment. From my teaching experience at a public university, when given an approximate time requirement to complete a task on their own, students often report having needed much longer than predicted. This is not surprising since they are not experts. Teachers sometimes forget what being a novice learner in their field is like. Deciding how much time students should reasonably devote to an assignment, ensuring that all tasks can be completed within that range, and communicating this plan with students can help them uphold academic integrity. UDL guidelines point out that experts identify the most important features in information quickly and allocate their time efficiently, whereas non-experts need the help of explicit cues (checkpoint 3.2).

From a practical perspective, an explicit completion time estimate is essential for many students, especially at the university level. Knowing how much time outside of class will be necessary for success helps with caregiving responsibilities, job schedules, and self-care. I have found that when students are unable to infer this on their own, or when there is a misalignment between their expectation and the actual time required, they are likely to experience stress and frustration. This problem is likely heightened in asynchronous online classes because there is no start and end time to class, unlike for in-person meetings.

Furthermore, some students may fall into the trap of perfectionism and take hours to complete an assignment that would have lasted 45 minutes in class. Stress and perfectionism are elements conducive to academic dishonesty as discussed earlier, which could lead to overusing MT. I observed this trend during the pandemic when work was exclusively submitted online. An otherwise successful student started using MT extensively in my intermediate-level class. When I asked her to explain, she cited stress and perfectionism. At the time, assessments in my classes did not indicate estimated completion times. Realistic expectations of time investment are crucial in helping students succeed with integrity (Ducar, 2018). In their Good Samaritan study, social psychologists Darley and Batson (1973) found that people in a hurry are less likely to help someone in distress. In some cases, participants did not even notice the individual needing assistance and stepped over them on their way to deliver a sermon on the parable of the Good Samaritan. The researchers concluded that "ethics becomes a luxury as the speed of our daily lives increases" (Darley & Batson, 1973, p. 107). As time demands increase for a myriad of reasons, learners' attention is stretched thin, which can lead to poor

ethical choices in order to meet time constraints. Creating a consistent routine with more but smaller assessments helps students organize their schedule with confidence and avoid the stress of having one week (i.e., midterm or final weeks) during which they will have more work than time allows to be completed well. Here is an example of a presentational task prompt following the principles described, and a model of the policy relating academic integrity shared with students via the LMS. In my classes, I administer 11 assessments in a semester, and rotate the modes of communication being evaluated (e.g., interpretive, presentational, interpersonal). The lowest grade is dropped. As a supporter of Open Educational Resources (OER), I make all my lesson plans available (details are posted on my teaching page: https://people.uwec.edu/millerjs/).

Conclusion

The virtual language class may seem rife with opportunities to cheat with online translators at students' fingertips. This should not push educators to view their class through a lens of suspicion. Ultimately, most students want to learn; little satisfaction is derived in the long term from completing every assignment through a translator. Eaton (2022) stated: "If we are going to truly commit to equity in academic integrity, we must not only accept other ways of knowing, being, and doing, we must embrace them" (p. 4). The present review and reflection on academic integrity in the virtual language class supports embracing multiple approaches to teaching and learning, multiple access points to material shared in the LMS, and multiple ways of using MT in practice and in assessments. This holistic approach has the potential to empower every learner and free teachers of the burden to detect academic misconduct and enforce disciplinary measures. Students should focus on learning, educators on teaching.

As discussed, instructional design can direct students towards integrity by reducing obstacles and factors that may lead to violations. Frameworks like UDL and TiLT provide research-grounded support (e.g., the UDL graphic organizer, the TiLT template) to enhance any teaching material, which will benefit all learners, especially among equity-deserving groups. Several strategies that fit those models were described. For one, pivoting away from high-stakes assessments toward smaller and regular evaluations can decrease stress and motivate learners to communicate in the target language. Second, a flexible policy regarding the use of online translators can give learners agency by letting them decide if they want MT assistance and to what extent, empowering them to better control their own learning based on their individual circumstances. Third, assessment directions that are written clearly, succinctly, and that repeat the instructor's stance on MT, can help calibrate academic integrity values. Those values are not immune to change over time for students experiencing stress and other challenges. These prompts should include accurate completion time estimates for students to allot the appropriate amount of attention. Finally, targeting functional accuracy, i.e., comprehensibility, in learners' language production rather than correctness can reframe the objective of assessments as an opportunity to grow and show what one can do. Focused corrective feedback balanced with positive reinforcement and acceptance of mistakes as natural steps

toward progress further drives the point that evaluations are invaluable resources to be welcomed. Acknowledging the challenges and rewards of the language acquisition journey, giving all students access to the resources available to them, and helping them use those judiciously, are actions educators can take to support standards of proficiency development, tenets of equity, and values of academic integrity.

References

Bowman E. (2022, December 19). *A new AI chatbot might do your homework for you. But it's still not an A+ student*. NPR. https://www.npr.org/2022/12/19/1143912956

Bowman E. (2023, January 9). *A college student created an app that can tell whether AI wrote an essay*. NPR. https://www.npr.org/2023/01/09/1147549845

Bretag, T. (2013). Challenges in addressing plagiarism in education. *PLOS Medicine*, *10*(12), 1-5. https://doi.org/10.1371/journal.pmed.1001574

Bretag, T., Harper, R., Ellis, C., van Haeringen, K., Newton, P., Rozenberg, P., & Saddiqui, S. (2019). Contract cheating and assessment design: Exploring the connection. Final report. https://ltr.edu.au/resources/SP16-5383_BretagandHarper_FinalReport_2019.pdf

Cassidy, C. (2023, January 9). *Australian universities to return to 'pen and paper' exams after students caught using AI to write essays*. The Guardian. https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays

CAST (2018). Universal Design for Learning Guidelines version 2.2. http://udlguidelines.cast.org

Çelik, Ö., & Lancaster, T. (2021). Violations of and threats to academic integrity in online English language teaching. *The Literacy Trek, 7*(1), 34-54. https://doi.org/10.47216/literacytrek.932316

Clifford, J., Merschel, L., & Munné, J. (2013). Surveying the landscape: What is the role of machine translation in language learning? *Research in Education and Learning Innovation Archives, 10*, 108-121. https://doi.org/10.7203/attic.10.2228

Correa, M. (2014). Leaving the "peer" out of peer-editing: Online translators as a pedagogical tool in the Spanish as a second language classroom. *Latin American Journal of Content and Language Integrated Learning, 7*(1), 1-20. http://doi.org/10.5294/laclil.2014.7.1.1

Darley, J. M., & Batson, C.D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, *27*(1), 100-108. http://doi.org/10.1037/h0034449

Davis, M. (2022). Examining and improving inclusive practice in institutional academic integrity policies, procedures, teaching and support. *International Journal for Educational Integrity*, *18*(14), 1-21. https://doi.org/10.1007/s40979-022-00108-x

Ducar, C., & Schocket, D. H. (2018). Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google Translate. *Foreign Language Annals*, 51(*3*), 779-795. https://doi.org/10.1111/flan.12366

Eaton, S. E. (2020). Race-based data in student conduct: A call to action. Werklund School of Education, University of Calgary, Calgary, AB. pp. 1-11. http://hdl.handle.net/1880/112157

Eaton, S. E. (2022). New priorities for academic integrity: equity, diversity, inclusion, decolonization and Indigenization. *International Journal for Educational Integrity*, *18*(10), 1-12. https://doi.org/10.1007/s40979-022-00105-0

Gierdowski, D., & Galanek, J. (2020). ECAR study of the technology needs of students with disabilities. Research report. Louisville, CO: ECAR. https://www.educause.edu/ecar/research-publications/ecar-study-of-the-technology-needs-of-students-with-disabilities/2020/introduction-and-key-findings

Groves, M., & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, *37*, 112-121. https://doi.org/10.1016/j.esp.2014.09.001

Hellmich, E. A. (2021). Machine translation in foreign language writing: student use to guide pedagogical practice. *Alsic*, *24*(1), 1-16. https://doi.org/10.4000/alsic.5705

Innes, A. (2019). Differentiating between translation and student translation: Red flags and salient lexicogrammatical features. *Lublin Studies in Modern Languages and Literature*, *43*(4), 1-13. http://doi.org/10.17951/lsmll.2019.43.4.1-13

Jiménez-Crespo, M. A. (2017). The role of translation technologies in Spanish language learning. *Journal of Spanish Language Teaching*, *4*(2), 181-193. https://doi.org/10.1080/23247797.2017.1408949

Jolley, J. R., & Maimone, L. (2015). Free online machine translation: Use and perceptions by Spanish students and instructors. In A. J. Moeller (Ed.), *Learn languages, explore cultures, transform lives: 2015 Report of the Central States Conference on the Teaching of Foreign Languages* (pp. 181-200). R. M. Terry. https://files.eric.ed.gov/fulltext/ED598272.pdf

Jolley, J. R., & Maimone, L. (2022). Thirty years of machine translation in language teaching and learning: A review of the literature. *L2 Journal*, *14*(1), 26-44. http://doi.org/10.5070/L214151760

Khoo, E., & Kang, S. (2022). Proactive learner empowerment: towards a transformative academic integrity approach for English language learners. *International Journal for Educational Integrity*, *18*(24), 1-24. https://doi.org/10.1007/s40979-022-00111-2

Klekovkina, V., & Denié-Higney, L. (2022). Machine translation: Friend or foe in the language classroom? *L2 Journal, 14*(1), 105-135. https://doi.org/10.5070/L214151723

Klimova, B., Pikhart, M., Delorme Benites, A., Lehr, C., & Sanchez-Stockhammer, C. (2022). Neural machine translation in foreign language teaching and learning: A systematic review. *Education and Information Technologies*, 1-20. https://doi.org/10.1007/s10639-022-11194-2

Korucu-Kış, S. (2021). On the effectiveness and limitations of captioning in L2 listening. *International Journal of Modern Education Studies, 5*(2), 516- 536. http://doi.org/10.51383/ijonmes.2021.153

Krashen, S. (1985). *The input hypothesis: Issue and implications*. Longman.

Lee, S. M. (2020). The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning*, *33*(3), 157-175. https://doi.org/10.1080/09588221.2018.1553186

Luck, J-A., Chugh R., Turnbull, D., & Rytas Pember, E. (2022). Glitches and hitches: Sessional academic staff viewpoints on academic integrity and academic misconduct. *Higher Education Research & Development*, *41*(*4*), 1152-1167. https://doi.org/10.1080/07294360.2021.1890697

Lynch, J., Salamonson, Y., Glew, P., & Ramjan, L. M. (2021). "I'm not an investigator and I'm not a police officer" - a faculty's view on academic integrity in an undergraduate nursing degree. *International Journal for Educational Integrity*, *17*(19), 1-14. https://doi.org/10.1007/s40979-021-00086-6

McAllister, C. & Watkins, P. (2012) Increasing academic integrity in online classes by fostering the development of self-regulated learning skills. *Clearing House*, *85*(3), 96-101. https://doi.org/10.1080/00098655.2011.642420

McNeill, L. (2022). Changing 'hearts' and minds: Pedagogical and institutional practices to foster academic integrity. In S. Eaton & S. Christensen Hughes (Eds.), *Academic integrity in Canada: An enduring and essential challenge* (pp 487-503). Springer. https://doi.org/10.1007/978-3-030-83255-1

Metz, C. (2023, January 10). *A.I. is becoming more conversational. but will it get more honest?* The New York Times. https://www.nytimes.com/2023/01/10/science/character-ai-chatbot-intelligence.html

Niño, A. (2020). Exploring the use of online machine translation for independent language learning. *Research in Learning Technology*, 28, 1-32. https://doi.org/10.1111/flan.12640

O'Neill, E. M. (2013). Online translator usage in foreign language writing. *Dimension*, *11*(1), 74-88. https://files.eric.ed.gov/fulltext/EJ1211306.pdf

O'Neill, E. M. (2019). Online translator, dictionary, and search engine use among L2 students. *CALL-EJ*, *20*(1), 154-177. http://callej.org/journal/20-1/O%27Neill2019.pdf

Pardo-Ballester, C. (2022). A case study: A technological insight on teaching translation. *Foreign Language Annals*, *55*(3), 894-913. https://doi.org/10.1111/flan.12640

Reedy, A. K., Wurm, P. A. S., Janssen, A., & Lockley, A. (2021). A community of practice approach to enhancing academic integrity policy translation: A case study. *International Journal for Educational Integrity*. *17*(12), 1-20. https://doi.org/10.1007/s40979-021-00080-y

Reynolds, B. L., Cui, Y., Kao, C.-W., & Thomas, N. (2022). Vocabulary acquisition through viewing captioned and subtitled video: A scoping review and meta-analysis. *Systems*, *10*(133), 1-20. https://doi.org/10.3390/systems10050133

Robinson, R., & Openo, J. A. (2021). The emotional labour of academic integrity: How does it feel? *Canadian Perspectives on Academic Integrity*, *4*(1), 20-39. https://doi.org/10.11575/CPAI.V4I1.71350

Sanni-Anibire, H., Stoesz, B. M, Gervais, L., & Vogt L. (2021). International students' knowledge and emotions related to academic integrity at Canadian postsecondary institutions. *International Journal for Educational Integrity*, *17*(21), 1-16. https://doi.org/10.1007/s40979-021-00088-4

Sefcik, L., Striepe, M., & Yorke, J. (2020). Mapping the landscape of academic integrity education programs: What approaches are effective? *Assessment & Evaluation In Higher Education*, *45*(1), 30-43. https://doi.org/10.1080/02602938.2019.1604942

Stachowiak, B. (Host & Producer). (2017, June 15). Reach everyone, teach everyone (No 227) [Audio podcast episode]. In *Teaching in Higher Ed*. Innovate Learning LLC. https://teachinginhighered.com/podcast/reach-everyone-teach-everyone

Stachowiak, B. (Host & Producer). (2018, October 18). Promoting academic integrity (No 157) [Audio podcast episode]. In *Teaching in Higher Ed*. Innovate Learning LLC. https://teachinginhighered.com/podcast/promoting-academic-integrity

van der Zee, T., Admiraal, W., Paas F., Saab, N., & Giesbers, B. (2017). Effects of subtitles, complexity, and language proficiency on learning from online education videos. *Journal of Media Psychology*, *29*(1), 18-30. https://doi.org/10.1027/1864-1105/a000208

White, K. D., & Heidrich, E. (2013). Our policies, their text: German language students' strategies with and beliefs about web-based machine translation. *Die Unterrichtspraxis/Teaching German*, *46*(2), 230-250. https://doi.org/10.1111/tger.10143

Winkelmes, M.-A., Boye, A., & Tapp, S. (2019). *Transparent design in higher education teaching and leadership: A guide to implementing the transparency framework institution-wide to improve learning and retention*. Stylus Publishing.

3

The Impact of Hybrid Professional Development on World Language Instructors' Assessment Literacy: Challenges and Opportunities

Amy Kim
Georgetown University
Margaret Borowczyk
Georgetown University
Rima Elabdali
Georgetown University
Margaret Malone
ACTFL

Challenge Statement

Professional development for language instructors is intended to support improvements in teaching and learning, but are such efforts effective? This study examines the influences, both short and long term, of a proficiency-based, hybrid professional development workshop focused on language assessment for postsecondary instructors of LCTLs

Abstract

Despite the widespread use of the ACTFL Proficiency Guidelines in many world language (WL) programs, relevant professional development (PD) opportunities for language instructors are scarce in U.S. post-secondary education contexts (Fulcher, 2012). This mixed-methods study focuses on a hybrid PD program (online and face-to-face) on oral proficiency assessment tailored toward 33 WL instructors and administrators and the follow-up questionnaire and interviews examining the short- and long-term influence on instructors' teaching,

assessment literacy and practices. The results showed significant short-term gains related to the understanding and use of the basic assessment concepts and the ACTFL OPI. The follow-up activities revealed important changes in participants' classroom teaching and assessment practices, towards more proficiency-based approaches, although some changes were difficult to sustain over time. This project offers perspectives on the influences of training in assessment literacy on WL teaching and assessment practices and the challenges instructors face.

The Impact of Hybrid Professional Development on World Language Instructors' Assessment Literacy: Challenges and Opportunities

With the ongoing interest in proficiency-based teaching and assessment, it has become a responsibility of world language (WL) educators (also commonly referred to as foreign language educators) to develop the assessment literacy and capacities necessary to communicate assessment results (Stiggins, 1999), determine appropriate tools for intended assessment uses (Inbar-Lourie, 2008), and meaningfully assess student outcomes (Scarino, 2013). Many WL higher education programs have immersed themselves in the proficiency-based approach that is committed to the systematic, rigorous reevaluation of their curricular goals and designs. Since the 1980s, the ACTFL Proficiency Guidelines and the accompanying Oral Proficiency Interviews (OPI) have been one of the most prominent scales and assessments employed for an array of purposes, such as for establishing entry points/exit goals and supporting high-quality teaching (Cox et al., 2018; Malone, 2010; Norris & Pfeiffer, 2003). WL educators are often expected to utilize ratings of student outcomes aligned with the ACTFL Proficiency Guidelines to meet the local needs of their institutions, and therefore continually refresh their assessment literacy skills through professional development (Malone, 2017; Taras & Davies, 2017; Taylor, 2009).

Assessment literacy has been a focus of scholarly discussions in language education since the onset of the 21st century (Boyles, 2005; Brindley, 2001; Davis, 2008; Inbar-Lourie, 2008). What is termed as language assessment literacy (henceforth LAL) has been theoretically conceptualized and captured into a plethora of componential models and frameworks. These frameworks consisted, by and large, of knowledge about assessment and language, familiarity with fundamental assessment concepts and principles, and skills to utilize assessment and interpret the results.

The combined focus on language assessment literacy and the ACTFL Guidelines has resulted in research that indicates that the principles and practices gained from ACTFL OPI training can result in washback, or the positive influences assessment practices can have on teaching and learning, for WL programs at the postsecondary level (Liskin-Gasparro, 2003; Malone & Montee, 2010; Norris & Pfeiffer, 2003). Compared to the demand for language assessment, however, professional development opportunities (PD) on the ACTFL Guidelines and OPI are limited, often constrained by funding, time and geography for many WL educators (Malone & Montee, 2010). In addition, many working with less

commonly taught (LCTL) languages have limited resources and language assessment PD opportunities. Empirical findings on the extent to which PD opportunities influence teaching at the classroom level remain limited, and there are challenges to making PD opportunities widely available.

To address these challenges, the current research examines the short- and long-term impact of a hybrid online/in-person PD workshop for LCTL teachers' assessment literacy. Based on longitudinal data from participant questionnaires and interviews, we investigate the influence of the workshop on educators' perception of student abilities, teaching and assessment practices, and curricular design. Thus, this study sheds light on the extent to which an improved understanding of the ACTFL Guidelines and proficiency-based-based assessment can influence positive instructional change in post-secondary settings.

Review of the Literature

Concerns about assessment literacy in general abound in the literature; prior studies have indicated that teachers may lack the skills needed to select or develop measures appropriate for their contexts (Taras & Davis, 2017; Wright et al., 2016; Xu & Brown, 2016). Similar concerns emerge in studies of language teachers, which suggest inadequate levels of assessment training (Fulcher, 2012; López Mendoza & Bernal Arandia, 2009) and the need for training in foreign-language assessment (Magnan, 1991; Terry, 1992). While there is ample evidence of inadequate assessment practices, pre-service assessment courses and seminars remain the dominant model for preparing WL teachers. This model, as Allen and Maxim (2013) argued, does not account for long-term responsibilities and contextual realities of WL teaching. Further, the ever-changing nature of language testing as a field highlights the need for on-going PD in language assessment (Malone, 2008).

In the educational sphere, a considerable body of research highlights a positive relationship between teachers' participation in PD and teacher efficacy (Shaha et al., 2015). Using Tschannen et al.'s (2001) framework, Swanson (2012) defines teacher efficacy as "a judgment about a person's belief in his or her ability to bring about desired outcomes of student engagement and learning" (p. 81) and emphasizes the importance of it for all teachers. Developing instructor LAL is arguably a type of teacher efficacy; if instructors believe that they can use assessment to increase student engagement and learning, then developing LAL, in the case of the present study, through an OPI workshop, may increase instructors' beliefs about their ability. L2 research shows that PD has a positive impact on teaching and learning in general. For example, PD offers language teachers opportunities to establish communal support. Such social networks focused on exploring innovative practices can be a catalyst for adopting and sustaining practices (Haviland et al., 2010; Hurlbut, 2015), because teachers are able to explore the feasibility of new practices under local constraints and discuss ways to adapt them to the regulations and culture of their institutions. PD can also encourage language teachers to move beyond teacher-centered teaching and grammar-oriented assessment towards learner-centered approaches and performance-based assessment practices (Hildén

& Fröjdendahl, 2018; Hurlbut, 2015). Additionally, PD has the potential to foster language teachers' assessment literacy. Teachers who engage in assessment-focused PD have demonstrated higher confidence in their assessment skills and are better able to reflect on their assessment practices (Hildén & Fröjdendahl, 2018; Hurlbut, 2015). Thus, PD focused on oral proficiency in general and the ACTFL OPI in particular may support development of language teacher efficacy.

Despite the broad generalizations on the success of PD, many argue that coherent, rigorous, long-term investigation of PD is rare yet essential in understanding the ways in which the gains remain over time (Boyles, 2005; Haviland et al., 2010; Shaha et al., 2015). Some argued that the effects of PD have been often investigated episodically, typically by conducting evaluation surveys immediately after the PD for quality assurance (Desimone, 2009). Boyles (2005) recommends that teacher development should be ongoing, take place through both online and face-to-face formats and occur in a variety of contexts and be instituted as part of teaching preparation programs. As Haviland et al. (2010) point out, "ongoing, focused faculty professional development on assessment, together with visible administrative support, can play a positive role in nurturing faculty understanding, confidence, and attitudes regarding assessment" (p. 274-275). The purpose of this project was to offer such opportunities and explore its influence on teachers' language assessment literacy.

Language Assessment Literacy

It is important that language educators have a grasp of assessment literacy so that they can both understand and apply these principles in their classes and explain assessment purposes and outcomes to their students and other stakeholders (Fulcher, 2012). Scholars have also emphasized the importance of embedding LAL in the contingencies of the classroom context. Early on, Brindley (2001) recognized the material constraints that impact classroom assessment practices and called for more flexible curriculum-related assessment. This recognition mirrors calls for bridging the gap between testing and assessment cultures (Inbar-Lourie, 2008) and emphasis on teachers' ability to apply their LAL within the context of their classroom practices (Boyles, 2005; Malone, 2013).

While conceptualizations of LAL have continued to evolve, there is not a single LAL model generally acknowledged to date (Vogt et al., 2020). We draw on Fulcher's (2012) characterization of language assessment literacy as:

> The knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom-based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order to understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals. (p. 125)

Contextualized within teachers' assessment needs, Fulcher's (2012) model of LAL aligns with goals of the current study which explores the influences of professional

development on teachers' assessment literacy. Drawing on survey data eliciting teachers' assessment needs, Fulcher (2012) delineated three main components of LAL: practices, principles, and contexts. Practices involve the knowledge, skills, and abilities employed during language assessment practices. Principles are the foundational knowledge of assessment concepts and processes that serve to guide practice. Contexts refer to the wider social, political, and historical context in which assessment practices and principles are situated. Fulcher argues for a procedural approach to demonstrate principles in practices and embeds both principles and practices in the wider context of language assessment. The PD program described in the current study focused on the classroom as a context in which the principles and practices of the ACTFL Proficiency Guidelines and OPI are embedded.

The ACTFL Proficiency Guidelines and ACTFL OPI Workshops

The ACTFL Proficiency Guidelines are a set of language proficiency descriptors that have shaped much contemporary thinking about proficiency-based instruction and assessment in the United States and beyond. The ACTFL Proficiency Guidelines and accompanying tests (including the ACTFL Oral Proficiency Interview (OPI), a standardized test developed by ACTFL to determine learners' oral proficiency level) are used widely for a variety of purposes, such as for establishing entry points/exit goals and supporting high-quality teaching (Malone, 2018). While there are controversies about the overreliance on the ACTFL Proficiency Guidelines and oral proficiency scales on setting educational outcome standards, research has indicated that the ACTFL Proficiency Guidelines are beneficial in establishing appropriate checkpoints for further adaptation of program curricula and materials (Liskin-Gasparro, 2003; Norris & Pfeiffer, 2003). Furthermore, Cox et al. (2018) highlight that despite criticisms of the Guidelines for their lack of grounding in empirical research, the definitions of proficiency levels the guidelines establish allow educational stakeholders "to hold each other accountable for learning outcomes and reasonable expectations on how language will be acquired" (p. 112).

ACTFL OPI workshops provide an opportunity for participants to not only understand the ACTFL Proficiency Guidelines but also the principles and practices of oral proficiency assessment. In order to become a certified OPI tester for the full, four-level scale, candidates must complete a four-day workshop and submit two rounds of sample interviews to a certified ACTFL facilitator. Thus, understanding of oral proficiency continues to develop beyond the workshop and into the certification progress.

This rigorous approach to assessment is challenging for many language instructors, therefore, those seeking limited certification can instead participate in an abbreviated workshop. Participants in this study took part first in a modular-based, online introduction followed by a three-day, face-to- face workshop that served as an introduction to the ACTFL Proficiency Guidelines and performance-based assessment, with an emphasis on how to apply these guidelines and assessments to the classroom. Liskin-Gasparro (1984) argued that although the direct use of the OPIs in the classroom by WL teachers is not advisable, "the

interactive and global focus of the oral interview can provide a model for the development of communicative oral classroom tests" (p. 68) and thus promote language assessment literacy which has been shown to be underdeveloped in many WL instructors (Fulcher, 2012; López Mendoza & Bernal Arandia, 2009). As such, improving participants' assessment literacy was a central goal of the workshop described in this study. Therefore, participants in this study practiced conducting ACTFL OPIs and used these techniques to design in-class oral assessments and interactive tasks that would challenge and develop students' oral proficiency.

Methods

The current project is a STARTALK-funded program, begun in 2017, to provide blended PD to different groups of critical languages educators during the summer. In this case, critical languages include Arabic, Chinese, Dari, Portuguese, Hindi, Korean, Persian, Russian, Swahili, Turkish, and Urdu. In 2019, the program also became open to a limited number of Spanish language educators who represented a small percentage of participants (N=3). The research component of the project employed a mixed-methods approach to examine the short and long-term influence of PD on teacher efficacy, in terms of their level of knowledge/skills and its transferability, revealing some external factors that affect teacher efficacy and implementation of program-level changes over the long term. The current paper presents a combination of data collected from two cohorts. Research activities took place between May 2018 and February 2019 for Cohort 1, and was repeated one year later (May 2019 through February 2020) for Cohort 2.

Program Overview

Drawing on Fulcher's (2012) framework of LAL, our blended PD workshop was designed to build a foundational understanding about basic concepts of language assessment (e.g., backward design) as well as the ACTFL OPI (i.e., Principles), and to focus on transferring these language assessment concepts and processes into classroom contexts (i.e., Practices & Contexts). The online modules were self-directed, addressing important concepts related to the PD (e.g., origins of the OPI) and facilitating participants' understanding through mini quizzes, forum discussions, and virtual office hours. The online modules were followed by a three-day, in person workshop in which participants were able to review their understanding of conceptual knowledge (day 1) and discuss how to adapt the ACTFL Proficiency Guidelines to the needs of their own programs (days 2 and 3). The PD was fairly consistent across the two years in terms of content in the online modules and in-person workshop. Participants in both cohorts provided self-assessments and feedback before and after the online modules to help tailor the in-person workshop to emphasize areas of concern for the participants. Participant feedback revealed their limited sense of familiarity in conducting OPIs (Cf. Table 2 below); thus, the workshop provided opportunities to practice the OPI with volunteer WL learners.

Participants

A total of 33 in-service instructors and administrators working with (at least one) critical language in postsecondary education took part in the study.

Both cohorts were recruited from publicly available lists of language faculty at institutions of higher education from across the United States. Across both cohorts, priority was given to participants with no background in the ACTFL OPI and to participants from the Mid-Atlantic region. Although the two cohorts were fairly similar across years, there were two basic differences between the first and the second cohort. As mentioned earlier, STARTALK allowed a few participants from Spanish language faculty in the second cohort, and the research team selected three faculty members who teach Spanish for this cohort. Second, based on feedback from the review team for the first year of the program. an emphasis was put on recruiting faculty from community colleges for the second cohort. Among them, 94% were instructors of critical languages or Spanish while 6% were administrators; 73% were affiliated with four-year institutions while 27% of the participants came from community colleges. While about 78% of the participants were affiliated with universities in the East Coast (e.g., DC, MD, VA, PA), the remaining 12% were affiliated with universities in the Midwest (e.g., MI, MO) and West Coast (CA, AZ). All participants were provided with resources to travel to the face-to-face workshop. The participants were heterogeneous with respect to the target language(s). The Spanish language instructors represented both four-year and community colleges and had similar backgrounds to the LCTL instructors. Table 1 summarizes the background characteristics of the participants, organized by cohort.

Table 1
Background characteristics of the participants (n = 33)

		Cohort 1 (n=15)	Cohort 2 (n=18)
		Number (percentage)	Number (percentage)
Position			
	Instructor	14 (93%)	17 (94%)
	Administrator	1 (7%)	1 (6%)
Types of institutions			
	Four-year universities	12 (80%)	12 (67%)
	Community colleges	3 (20%)	6 (33%)
Geographical area			
	East Coast	13 (87%)	13 (72%)
	Midwest	1 (7%)	3 (17%)
	West Coast	1 (7%)	2 (11%)
Target language			
	Mandarin Chinese	4 (27%)	5 (28%)
	Arabic	2 (13%)	4 (22%)

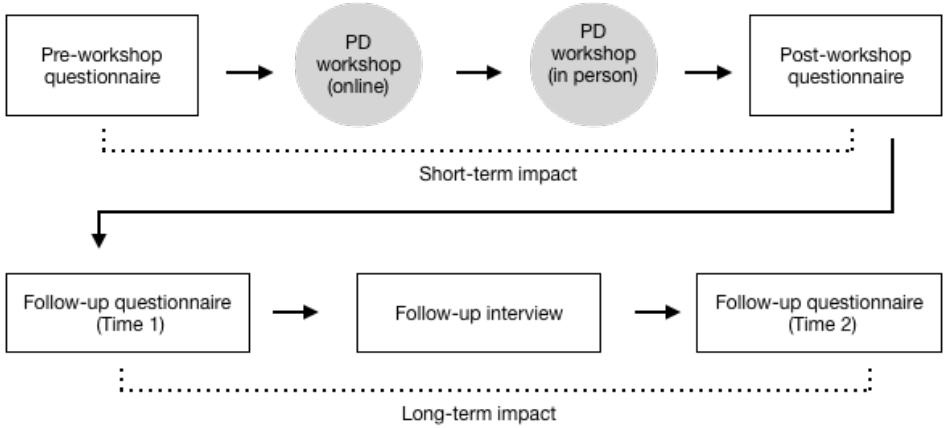
Korean	3 (20%)	3 (17%)
Portuguese	1 (7%)	2 (11%)
Russian	2 (13%)	0 (0%)
Hindi	0 (0%)	0 (0%)
Turkish	1 (7%)	0 (0%)
Persian	1 (7%)	0 (0%)
Swahili	0 (0%)	1 (6%)
Spanish	0 (0%)	3 (17%)
Mixed (administrator)	1 (7%)	0 (0%)

Instruments and Procedure

The study used a series of questionnaires and interviews, conducted to collect quantitative and qualitative data from participants. Figure 1 below summarizes the details of the data collection procedure.

Figure 1

Summary of data collection procedure



There were three major instruments used in the present study:

- Workshop questionnaires: The workshop questionnaires were administered online before (pre-) and immediately after (post-) the workshop in order to gauge the workshop's short-term impact. These questionnaires captured participants' biodata, participants' level of satisfaction from the overall PD program, and level of familiarity and comfort with core assessment concepts (e.g., backward design, the ACTFL Proficiency Guidelines) on a 4-point Likert-scale.
- Follow-up questionnaires: The follow-up questionnaires were individually crafted to the specific goals set by each participant and were administered online at two different times Time 1 (3-4 months after the PD) and Time

2 (7-8 months after the PD). The items in the follow-up questionnaires included questions, such as "Has the workshop influenced the ways you provide feedback to your students in any way in your Arabic 101 course?" Participants in Cohort 1 were presented with binary choices between 'yes' or 'not really.' Those in Cohort 2 were asked to respond on a 5-point Likert scale.[1] In both years, participants were given additional space for further comments. The follow-up questionnaires were designed to examine the workshop's long-term impact on participants' perceived ability to transfer their knowledge into practice.

- Semi-structured interview protocol: Between the two follow-up questionnaires, semi-structured interviews were conducted with several participants. The interviews were guided by the individual participants' responses to the first round of the follow-up questionnaire. The goal of the interview was to gain a deeper understanding of ways in which changes, if any, were being implemented. Interviews were conducted in person, via telephone or videoconferencing and included discussion of challenges associated with the planning and implementation. All interview data were transcribed and double-coded by the research team members.

Data Analyses

The study collected a mixture of quantitative and qualitative data. The workshop questionnaire elicited quantitative data from Likert-scale. This quantitative data were entered into SPSS, then analyzed using paired-sample t-test to examine whether significant changes were found after the PD workshop. Numerical data and open-ended responses collected through the follow-up questionnaires were compiled using Excel. The interviews were coded based on a coding scheme, which was developed after reading and annotating the interview transcripts from a pilot study in summer 2017 based on an inductive coding process (Strauss & Corbin, 1994). Details of the coding schemes are summarized in Appendix A. All interview data were double-coded by the research team using the software NVivo (QSR International, n.d.). Any discrepancies were discussed and noted on a separate file.

Results

Short-Term Gains in Knowledge: Pre- and Post-Workshop Questionnaire Results

Data from the workshop questionnaires suggested that there was an increasing level of familiarity and comfort regarding the key assessment concepts relevant to the ACTFL Proficiency Guidelines and the OPI. Table 2 shows the two-tailed paired-sample t-test results for the difference between the pre- and post-workshop questionnaires.

1. It was realized after the 2018-2019 follow-up that the impact of the PD programs were often conceived on a continuum, rather than on dichotomous categories (e.g., 'yes' and 'not really'). Thus, the follow-up questionnaires in 2019-2020 employed a Likert-scale.

Table 2
T-test results from pre- and post-workshop questionnaires (n=33)

LAL		Pre	Post	Effect size	t-test
Principles	Backward design	2.82 (0.98)	3.30 (0.68)	0.94	2.97*
	The ACTFL proficiency guidelines	2.88 (0.74)	3.38 (0.56)	0.80	3.72**
↓↑	ACTFL OPI	2.48 (0.97)	3.36 (0.65)	1.08	4.67**
	Elicitation of speech for an ACTFL OPI	2.06 (1.06)	3.21 (0.74)	1.06	6.22**
Practice	Rating an ACTFL OPI	2.15 (1.06)	3.18 (0.64)	1.10	5.36**
	Designing oral interpersonal communication tasks	2.94 (0.90)	3.33 (0.65)	1.06	2.14*
↓↑	Planning units based on proficiency goals	2.97 (0.88)	3.15 (0.62)	0.95	1.10
	Providing feedback to students	2.88 (0.89)	3.18 (0.64)	1.05	1.67
Context	Using assessment results to make changes	2.88 (0.93)	3.12 (0.60)	0.83	1.58

Note: Answer choices range from 1 to 4, where 1 = not familiar/comfortable at all, 2 = slightly familiar/comfortable, 3= fairly familiar/ comfortable, 4 = extremely familiar/ comfortable.

* indicates $p < 0.05$; ** indicates $p < 0.001$

Results indicate that short-term gains were achieved in all areas; however, the differences were found to be statistically significant in only six areas ($p < 0.05$), all of which corresponded to the Principles and Practices components of LAL, including 'Backward design,' 'The ACTFL proficiency guidelines,' 'ACTFL oral proficiency interview (OPI),' 'Elicitation of speech for an ACTFL OPI,' 'Rating of an ACTFL OPI,' and 'Designing oral interpersonal communication tasks.' The pre-post differences were not statistically significant in areas corresponding to the Context component of LAL ($p > 0.05$), including 'Planning units based on proficiency goals,' 'Providing feedback to students,' and 'Using assessment results to make changes.' This demonstrates the need for long-term evaluation of PD interventions, since instructors gain the most experience applying their knowledge contextually when they are back in the classroom. Therefore, the impact of the workshop on the Context component of LAL is explored more deeply in the long-term follow-up.

Also revealed in the short-term analyses was the practical significance expressed in terms of effect size (i.e., magnitude of effect). The effect sizes indicate that the magnitude of the effects, in terms of Plonsky and Oswald's (2014) field-

specific benchmark for within-group contrasts, were medium (between 1.06 to 1.10) for three areas closely related to the ACTFL OPI. Effect sizes of other areas were small (between 0.8 to 0.94). Another point worth noting is the medium effect sizes (between .83 to 1.05) for nonsignificant items, which may signify practical significance (as opposed to statistical significance). Taken together, the findings suggest that the short-term gains were achieved, particularly in the areas related to the understanding and use of the ACTFL OPI.

Long-Term Gains: Follow-Up Questionnaires and Interview Results

At the beginning of the following fall semesters, participants were invited to partake in the follow-up activities. Overall, the response rate ranged widely, ranging from 38% to 83%. The initial response rates were quite high in both Cohorts (73% and 83% respectively); however, response rates in subsequent rounds were lower (53% and 38% respectively). Details are summarized in Table 3.

Table 3
Response rates of the follow-up activities

		Invited	Responded	Response rate
Cohort 1	Follow-up questionnaire (Time 1)	15	11	73%
	Follow-up semi-structured interview	15	8	53%
	Follow-up questionnaire (Time 2)	15	8	53%
Cohort 2	Follow-up questionnaire (Time 1)	18	15	83%
	Follow-up semi-structured interview	18	7	38%
	Follow-up questionnaire (Time 2)	18	7	38%

Note: response rate = responded/invited

Evidence from The Follow-Up Questionnaires

The follow-up questionnaires asked respondents to first indicate whether the PD activities influenced the ways they are currently teaching, which includes their perception of students' ability levels, any changes to the curriculum/syllabus, and ways to provide feedback, therefore targeting the Contexts component of LAL. These questionnaires were administered twice after the PD. The first follow-up questionnaires were administered at the beginning of the new semester (September/October), followed by the second one at the beginning of the next semester (January/February). They included close-ended questions, followed by additional space for further comments. Participants in Cohort 1 were asked to provide 'yes' or' not really' responses to the questions. Table 5 shows the percentages of Cohort 1 participants who indicated yes to each question.

Table 4
Results of the follow-up questionnaires from Cohort 1

LAL	Has the workshop influenced the…?	Time 1 (n = 11)	Time 2 (n = 9)
	Backward design	64%	67%
	Providing feedback	64%	53% [a]
Context	Perceptions of student abilities	100%	89%
	Overall teaching practices	91%	78%
	Curriculum/syllabus	73%	56%

Note: Questionnaire for Time 1 took place in September, 2018 and Time 2 in January/February, 2019. The percentages were calculated by dividing the number of 'yes' responses by the number of total responses.
[a] Due to one missing data point, the total number of responses was 8.

As Table 5 shows, the participants in Cohort 2 responded to a 5-point Likert scale. With the Likert data, t-tests with unequal variances were performed to investigate the sustainability of impact.

Table 5
T-test results obtained from the follow-up questionnaires from Cohort 2

LAL	Has the workshop influenced the…?	Time 1 (n =15)	Time 2 (n = 7)	Effect size	t-test
	Backward design	3.96 (1.09)	3.36 (1.15)	0.42	1.37
	Perceptions of student abilities	4.45 (1.03)	3.92 (0.92)	0.21	0.63
Context	Overall teaching practices	4.30 (0.95)	3.86 (0.77)	0.45	1.24
	Curriculum/syllabus	3.33 (1.30)	2.79 (1.25)	0.32	1.16
	Providing feedback	3.78 (1.15)	3.36 (1.39)	0.09	0.31

Note: Answer choices range from 1 to 4, where 1 = not at all, 2 = a little, 3= a moderate amount, 4 = a lot, and 5 = a great deal. Questionnaire for Time 1 took place in October, 2019 and Time 2 in February, 2020.

Though different metrics were used to measure the long-term influences of the workshop, the means from Time 1 indicated a pattern: PD had a strong positive impact on participants' perception of student abilities and overall teaching practices at the beginning of the fall semester, while having a modest impact on other areas such as the curriculum/syllabus, backward design and the provision of feedback. The decreases observed in Time 2 suggested that the overall impacts decline over time. This pattern resembled that of Cohort 1: the percentage of participants who believed in the PD's influence decreased by between 11% to 17% in most items, though there was a 3% increase in terms of the use of the backward design. Data from Cohort 2 showed that the mean differences between Time 1 and Time 2 decreased for all items. Based on the descriptive statistics, it appeared that the relatively higher effects observed at the beginning of the semester were not sustained.

However, the t-test results with unequal variances revealed that the differences were not statistically significant ($p > 0.05$) for all items. The lack of significance between Time 1 and Time 2 may be due to the small sample size. Furthermore, the effect sizes of all items were marginal (Plonsky & Oswald, 2014). This means that the decreased impact observed in ratings between Time 1 and Time 2 are negligible. It may be difficult to draw meaningful conclusions based solely on quantitative data (without qualitative backing). The open-ended responses provided more nuanced insights to the PD's long-term influence.

Open-Ended Responses

The open-ended responses lent support and context to the quantitative finding that the PD had the largest impact on instructor's perceptions of student abilities. Namely, the workshop helped instructors identify clear benchmarks for their students' proficiency and therefore fostered more individualized and realistic expectations of what students could be expected to understand and produce after a given lesson. One Mandarin instructor noted, "I adjusted my class's expectation based on the student's current levels and tried not to push students to produce the functions [that] are two or more levels beyond what they can do." One Spanish instructor also discussed adjusting her expectations of student abilities, saying "It has changed how much material I try to cover in a single class and even a single semester. I am going to delete some of my future assignments so I can have more time recycling material." Having more realistic goals for what students will be able to do with the target language given their proficiency level was therefore key in allowing instructors to craft more appropriate lessons and curricula. It also helped instructors better identify individualized needs in classrooms with mixed levels. One Korean instructor explained:

> In a midterm speaking test, as a lower risk assignment, I emulated the OPI model, asked students to pick scenarios I created, appropriate to what we have covered. I discovered some of my students had much more prior knowledge of Korean than others and it helped me to adjust my lessons for them. If I had just asked them to memorize and present, I wouldn't have known this about them.

By applying principles of the ACTFL OPI to classroom assessments of speaking, instructors were able to assess not only the extent to which students grasped lessons taught in that class, but the variation in ACTFL proficiency levels that exist within the class. Then, instructors adjusted their lessons for the specific levels of their students.

Participants also discussed shifting their approach to providing feedback to students, the most pervasive shift being a focus on communicative competence rather than grammatical accuracy. A Mandarin instructor explained, "It changed my [per]ception of speaking abilities versus accuracy. I can now more easily accept inaccurate, from the grammatical point of view, performance as successful if the communication/functional goal is achieved." This sentiment was echoed by a Korean instructor who noted, "I am focusing on their level-appropriate

communicative competency in giving feedback rather than being nitpicky about their grammar correctness." This suggests that the workshop's emphasis on applying ACTFL's Can-Do statements (which describe what learners should be able to do at each ACTFL proficiency level) to classroom instruction allowed participants to adjust their focus from grammatical accuracy to what students were able to accomplish communicatively with their target language, and center their feedback around improving communication.

Quantitative results revealed that the workshop had a small effect on participants' approach to curricular design and their application of backwards design to curricular planning. The open-ended responses provide insight into why the workshop had a smaller impact in those areas, in contrast to perceptions of student abilities and overall teaching practices. In many cases, instructors saw themselves as having limited control over curriculum design because of departmental requirements; this factor also impeded the extent to which they could apply the principles of backwards design to their syllabi. Specific nuances and caveats to this issue are explored in the following section on interview responses.

However, for the participants who discussed implementing curricular changes as a result of the workshop, they mentioned recycling materials to target the greatest amount of linguistic skills, adapting textbook materials to suit more interactive tasks, and implementing backwards design for individual lessons. During the face-to-face workshop, the facilitator had discussed strategies for recycling materials in ways that target all four skills (speaking, reading, writing, listening), to foster deeper engagement with a given topic, and the questionnaire responses revealed that in several cases, this advice had an effect on participants' curricular practices. One Spanish instructor noted that she "mixed all the skills of reading, listening, and speaking in the class. So now the students read a text and then watch a video on that topic which leads up to a discussion on what their opinion is relating to what they learned." Furthermore, participants' increased familiarity with designing interactive tasks allowed them to adapt existing textbook-based materials to better address student needs. One Mandarin instructor explained, "I refer to the textbook for some information, but I adopted and revised- revised it, adapted it, modified it to suit the needs of my students." Therefore, even when instructors must follow a designated textbook, they can use insights from the workshop to adapt textbook lessons to their students' proficiency levels.

Finally, some instructors were able to apply the principles of backwards design to their post-workshop curricula and lessons. One Spanish instructor noted that although her department sets curricular standards for each Spanish level, she makes an effort to translate departmental standards into ACTFL proficiency goals for her students and uses them as a starting point for her syllabus design. She explained, "I think about what I want students to be able to do at the end of the semester and make sure all lessons and assessments are helping them towards that end." Approaching backwards design from a more micro perspective, at the level of the individual lesson, one Mandarin instructor wrote, "I will first think of the learning 'goals' or lesson Can-dos statement before planning my teaching activities. I also pay attention [to] how to check for understanding/assessment

tools now." Therefore, part of her lesson planning now involves identifying what she wants students to be able to do by the end of the lesson and planning formative assessments around those goals. Overall, the questionnaire revealed that the workshop influenced participants' curriculum design by encouraging adapting course materials to target all linguistic skills and thinking about the end goal of the lesson or course in the initial planning stages.

The extent to which participants were able to sustain these changes seven to eight months after the PD workshop decreased in all areas from the initial September/October questionnaire to the January/February questionnaire; the later questionnaire revealed that although participants retained an understanding of how to assess students' abilities and provide feedback in consistent with the workshop, they found themselves defaulting to pre-workshop strategies as the demands of the semester grew. Furthermore, some found it difficult to fully incorporate ACTFL's proficiency standards and performance-based assessments into their curricula as the semester went on. For example, a Korean instructor wrote, "I try to do more performance-based activities but it requires too much time in the beginner's level." In a similar vein, an Arabic instructor noted, "I think that I need more practice to learn to apply the ACTFL standards into lesson objectives." This underscores the need for continued support in the months following the PD workshop to reinforce the tools and concepts that were introduced in the original intervention.

Evidence from The Interviews

The semi-structured interviews conducted between the two questionnaire administrations provided more elaborated and detailed explanations for the quantitative results, along with specific classroom examples. The list of themes emerging from the interviews (i.e., coding scheme), as well as the rate of inter-rater reliability in coding for these themes can be found in Appendices A and B. This section will discuss the two most frequent themes, teaching and assessment, which highlight changes the participants made to their teaching practices, including their course structure, teaching methodology, or classroom activities, and changes to their assessment practices, including placement tests, checks for understanding, and formative oral proficiency assessments.

Teaching

The participants commented on a myriad of factors and concepts covered in the PD that made it possible to bring immediate changes to their teaching practices. These changes included using open-ended questions and scaffolding techniques, designing individualized and student-centered lessons, incorporating small group work, and using authentic instructional materials as well as virtual tools to promote interaction. For example, one Mandarin instructor said, "I feel the metaphor of ceiling and of the floor you know has just started me to think um of preparing different materials, different activities for different students." Learning how to design tasks and activities that gauge students' comfort zones versus their linguistic limits helped this educator foster a more individualized approach to

learning. Another Mandarin instructor discussed integrating authentic materials via tasks that used Chinese currency to facilitate tasks regarding shopping: "We also had fake money, fake Chinese money, I actually also brought in real money. So all of those prompts I think created an environment where the students could perform tasks." A Russian instructor discussed facilitating student presentations on the history and politics of Russian cities by using information from Russian news outlets.

Participants also explained that because of the PD, they learned of additional resources and approaches to foster authentic interaction among classmates and with target language-speaking partners. One Portuguese instructor noted that the PD facilitator encouraged her to start a video-based discussion board for her students, on which students could respond to each other's posts about a topic using audio and video recordings, thus promoting asynchronous spoken interaction. Furthermore, a Spanish instructor explained that during the workshop, she was introduced to a mobile app called *Conversifi* that has allowed her to connect her students with native Spanish speakers for a language exchange:

> Conversifi, that was the other thing. So there's this online like virtual exchange with Spanish language learners and English language learners. So my students have connected with students of English in Spain and Latin America and they had 15-minute conversations.

Taken together, these reflections suggest that participants drew on lessons from the workshop to create classrooms that incorporated a student-centered approach and maximized opportunities for authentic interaction.

Assessment

Regarding changes to assessment practices, participants repeatedly discussed using OPI techniques at the beginning of the semester for placement purposes and using the ACTFL Proficiency Guidelines to guide the development of assessment tasks and grading rubrics throughout the semester. One Arabic instructor discussed adapting the OPI to assess student levels at the beginning of his semester:

> The workshop helped me assess students' proficiency. Before the workshop, I did not pay any attention to student's levels or their abilities at the beginning of the course. I used to just start working on the assigned textbook and I did not emphasize the conversation exercises in the textbook. After the workshop, I conducted OPIs with my students at the beginning of the course to assess their levels and understand their abilities.

Similarly, a Spanish instructor conducted an OPI-based speaking assessment with the students in her Intermediate Spanish class at the beginning of the semester to assess their readiness for the course, and in some cases, encouraged students to enroll in a different level. Additionally, participants continued drawing on the ACTFL Proficiency Guidelines and OPI techniques throughout the semester to develop appropriate assessment tasks and grading rubrics. Many participants indicated that the program enabled them to build flexibility into the ways they perceive student abilities and conduct assessment. For example, one Portuguese instructor said:

> Because if I look back, the way I did oral exam before, it was more structured and when you ask specific questions that are [the] same for each student then you're expecting [the] same answers in the same category. And you'll be kind of penalizing the student if the student does not produce what you're expecting. So that's the conversation I had with her [trainee] and you know and talking about how to be more open to more creative language and how to make that happen. And I think that model from the OPI gives you some guidance on how to do that.

Another Portuguese instructor echoed this sentiment, focusing on how she adjusted her perceptions of student abilities in presentation tasks:

> One thing that it was very good actually in the training was that going over the assessment that we could use in the classroom because especially for fluency, and proficiency, like when the students are making a presentation especially at the basic level, sometimes it's very easy to just like watch it and assess it from the perspective of like, a higher, like a native speaker. And then just like it just makes me more conscientious, to be more aware, and pay more attention to that. That's just like what is expected at their level, and how much they are fulfilling their expectation of that level and not at a higher level.

Therefore, the data show that participants benefited from applying the ACTFL Proficiency Guidelines as a yardstick for measuring functional language ability, which supported development and grading of formative assessments.

While the examples above suggest that positive changes occurred, the instructors also reported challenges that limited their ability to implement these changes. Although the study included primarily LCTL instructors, three Spanish instructors were included in Cohort 2; they accounted for slightly less than 10% of the total participants in the study. Only one participated actively in the follow up activities. Generally, there did not appear to be any notable differences between the Spanish professor's responses and those of the LCTL instructors except that the Spanish instructor occasionally referred to department-level initiatives. As a part of a fairly large language department, she reflected on the workshop's potential across instructors of Spanish at her institution.

The two main challenges reported across multiple interviews were departmental restrictions on curricular change and a lack of knowledge in how to apply concepts from the PD to heritage speakers.

Departmental Restrictions

In most cases, participants were employed in language departments where they either received a pre-determined syllabus, or where full-time faculty in the department collectively determined the curricula for each language level. Therefore, in many cases, participants were powerless to redesign their syllabi according to the principles they learned during the workshop, restricting their ability to explore the Contexts component of LAL. Some instructors noted that, because they were working with a pre-determined curriculum, the workshop had

little to no influence on their curricular design. Instead, they applied what they had learned in a more limited manner, namely to the way they provided feedback to their students or the way they perceived their students' language abilities. Other instructors however, interpreted their agency a bit more expansively, by finding ways to incorporate performance-based assessment, interactive tasks, and backwards design within individual lessons, to stay within the broad framework of the departmental curriculum, but innovate in key areas. Other participants even introduced the ACTFL Proficiency Guidelines and central lessons from the workshop to their colleagues in departmental meetings to influence changes in department-wide curricula. However, to assist all participants, regardless of their stature and level of influence in a language department, future iterations of the PD could provide guidance to participants on how to apply workshop principles to their syllabi in cases where they must innovate within a predetermined structure. Such restrictions and limitations occurred regardless of language taught.

Applying Lessons to Heritage Speakers

Instructors who taught classes specifically for heritage speakers, or who had heritage speakers in their classes, frequently shared the perception that they did not feel confident applying the lessons they had learned in the PD to these speakers. For example, when curricula for heritage speaker classes were oriented towards improving literacy skills and expanding students' formal registers, instructors struggled to apply the ACTFL Proficiency Guidelines-Speaking to written work. The OPI, of course, is focused on speaking proficiency, so this criticism is logical. Furthermore, some noted that they felt ill-equipped to provide useful feedback to heritage speakers, especially in cases where the student spoke a stigmatized or non-standard dialect of the heritage language. It is important to acknowledge these issues in PD to ensure that participants understand how key concepts in proficiency and assessment can be applied to the full range of learners.

Discussion

The central aim of this study was to investigate the short- and long-term influences of a hybrid professional development program on WL teachers' language assessment literacy. While some testing and educational organizations have previously provided PD programs on WL assessment (Malone & Montee, 2010), there was no robust empirical evidence on the extent to which learning is transferred to classroom practices over time. In order to address this gap, we evaluated the impact of our program on two levels: the magnitude of the program's short-term impact on participants' understanding of assessment concepts and material changes implemented as a result of the program. Therefore, this study was the first to explore whether short-term gains following PD in language assessment literacy have a lasting impact, using self-perception data. In this section, we discuss some specific findings relevant to future studies.

The pre- and post-workshop questionnaire results revealed that significant short-term gains were achieved in the Principles and Practices components of LAL, namely participants' familiarity with the principles of backwards design, the ACTFL Proficiency Guidelines, the ACTFL OPI, eliciting speech for an ACTFL OPI,

rating an ACTFL OPI, and designing oral interpersonal communication tasks. The follow-up questionnaires and interviews revealed progress that participants made in applying LAL Principles and Practices to their classroom contexts. Specifically, they perceived that the workshop had the greatest impact on their perception of student abilities and overall teaching practices. Namely, the workshop provided participants with a standards-based framework for interpreting student abilities, which allowed them to set realistic, level-appropriate expectations for student performance. It also promoted the use of feedback oriented towards improving students' communicative competence rather than placing a primary focus on grammatical accuracy. The principles of backwards design inspired participants to design more of their lessons around clearly stated proficiency goals. Regarding the impact on overall teaching practices, participants reported using more open-ended questions and scaffolding techniques, designing more individualized and student-centered lessons, and using more authentic instructional materials as well as virtual tools to promote interaction.

Concerning the impact on classroom assessment practices, participants discussed using OPI techniques for placement purposes and the ACTFL Proficiency Guidelines to develop classroom assessment tasks and grading rubrics. These findings lend support to the hypothesis, articulated by Malone and Montee (2010), that training in OPI techniques could result in positive washback through a more informed approach to developing students' communicative competence.

The long-term follow up revealed that the extent to which the workshop influenced these positive changes in curricular practices decreased over time, although the change was small in range and not statistically significant. The open-ended questionnaire responses revealed that in some cases, participants defaulted to their pre-workshop methods of providing feedback and assessing student abilities, as the demands of the semester increased. At the same time, participants reported retaining a theoretical understanding of how to implement changes to their classroom practices consistent with the PD. Some participants found it difficult to continue designing and implementing performance-based assessments and interactive tasks in a limited amount of instructional time, while still complying with department-set achievement benchmarks. As Boyles (2005) and Haviland et al. (2010) have suggested, this tension highlights the need to provide on-going professional development and follow-up, together with programmatic commitments to continuously support the instructors.

In addition to the challenge of sustaining changes over the long term, instructors also faced two broader challenges: (1) applying proficiency-based instruction to predetermined curricula consistent with traditional grammar-based approaches and (2) applying formative assessments based on OPI principles for heritage learners. In several cases, instructors felt a lack of agency in being able to apply lessons from the PD, especially the principles of backwards design, in cases where they could not make changes to their curriculum or syllabus. In other cases, they decided to forego comprehensive changes to their curriculum or syllabus, and instead found ways to draw on workshop principles to design individual lessons around interactive activities. In all cases, participants could potentially benefit

from explicit discussions during the PD of strategies for identifying spaces within a predetermined curriculum to implement their increased assessment literacy and familiarity with the ACTFL Guidelines. Because collaboration with colleagues can also be highly beneficial to exploring different ways to implement PD principles under local institutional constraints (Haviland et al., 2010), the program could find ways to facilitate collaboration and share insights among PD participants in the months following the workshop. The impact of this collaboration on participants' classroom practices could then be investigated. Likely, though, the most comprehensive changes would result from collaboration at the department level, where participants can work with their own colleagues to influence curriculum design. Such an approach would support alignment in proficiency-based approaches across course sequences, as well as help develop expectations for attainment across levels. With the increased adoption of the ACTFL Proficiency Guidelines across U.S. colleges and universities (Malone & Montee, 2010), such continuity may become increasingly feasible.

Another challenge identified was working with heritage learners: participants who were teaching heritage speakers in their classes noted that they struggled to apply the ACTFL Guidelines to their lesson design and assessments. They were also unsure of how to provide feedback in a way that honored students' non-standard varieties of the heritage language while supporting their progress towards higher proficiency levels. Valdés (1989) and Draper and Hicks (2000) argued that since the ACTFL Proficiency Guidelines were designed to assess foreign language learners rather than heritage language learners, the developmental hierarchy described in the Guidelines does not fit the heritage learners' language developmental sequence, and therefore should not be used to assess heritage learners' progress. After using the OPI to assess the speaking ability of Intermediate-level Hindi foreign and heritage language learners, Ilieva (2012) recommended that the guidelines should differentiate between both learner groups by separating those features that are characteristic of just HLLs. Martin (2010), however, argued that there is no consensus on the need for different sets of assessments or criteria to assess heritage and foreign language learners, especially when what is being assessed is the ability to perform specific communicative tasks, rather than the ability to match native speaker performance. Despite this lack of consensus among researchers, it may be beneficial, during PD, to explain areas in which OPI techniques may be applicable to the design and assessment of oral tasks for heritage speakers, and areas in which they may be beyond the scope of the lesson, such as in heritage courses that emphasize writing and the development of formal registers. The field would also benefit from more studies investigating the utility of PD in ACTFL Guidelines and OPI techniques to teaching heritage speaker courses, especially in programs that have implemented such PD successfully.

Limitations

The study was limited by the small sample size in each cohort (lacking enough power to detect meaningful differences). Additionally, the study elicited limited responses. The follow-up data represented the experiences of 38% to 83% of the

total participants, and the extent to which low response rate has resulted in (non) response bias is unknown. Furthermore, the influences reported in the current study are self-reported ratings and therefore may suffer from social desirability bias. Note, however, that the study captured the challenges that hamper implementation and areas that deserve further improvement, which may indicate that the bias was minimal. Given these limitations, no claims can be made about the generalizability of the results beyond the present sample (though it was not the goal of the study to do so) and the results should be interpreted with caution. Future studies should be carried out with a larger sample, accounting for external factors (e.g., time, institutional context) that may moderate the effects of PD.

Conclusion

This study investigated both the long- and short-term impact of PD in assessment literacy on instructors of several LCTLs and Spanish in higher education. Findings demonstrated that in the short term, the PD workshop influenced participants' assessment literacy and their ability to design lessons, communication tasks, and assessments in ways that target students' oral proficiency. This development supports Swanson's (2012) assertion that knowledge can transform language teacher efficacy and that PD can contribute to such efficacy (Desimone et al., 2002; Shaha et al., 2015). However, participants found some curricular changes, such as an increase in performance-based assessment, difficult to sustain in the long-term and explained that they would benefit from additional practice in applying the workshop principles to their classrooms, months after the original PD. Participants were also hampered in applying the lessons they learned in PD due to their limited agency in changing a department-determined curriculum. Others discussed the limited applicability of the ACTFL Guidelines to heritage speaker performance.

Therefore, though ACTFL/OPI-based PD efforts seem to be useful in promoting a proficiency-based approach to WL teaching and assessment, language teachers need more explicit and sustained guidance around how to implement new approaches given departmental restrictions and around how to apply the ACTFL Guidelines to heritage language learners. Language departments sending faculty members to such workshops with the intention of implementing changes to curricula and proficiency standards across all courses in a given department could promote proficiency-based approaches and avoid conflict with traditional grammar-based approaches within the same sequence. In this way, visible administrative support (Haviland et al., 2010) would be crucial to ensuring the widespread impact of proficiency-based PD within language departments. Furthermore, one-time interventions may not guarantee long term gains, so ongoing PD and continuing collaboration among workshop participants may be needed to support sustainable curricular change. Future studies will need to investigate what types of follow-up interventions are most effective in facilitating the retention of principles acquired during OPI training. Because much teacher PD on language assessment literacy consists of singular workshops, the goal of researchers and practitioners in the upcoming years will be to gain a better understanding of the impact of PD in the months and years following an

intervention, and to discover ways of supporting follow-up activities and colleague collaboration that promote sustained improvement.

Acknowledgements

This work was supported by STARTALK [Grant # H98230-18-1-0047]. We would also like to thank Todd H. McKay for his contributions at initial stages of this research; Jahurul Islam, Rachel Myers, Zhuosi Luo, and Ho Fai (Viggo) Cheng for assisting with transcription and coding.

References

Allen, H. W., & Maxim, H. H. (Eds.). (2013). *Educating the future foreign language professoriate for the 21st century*. Boston: Heinle Cengage. Boyles, P. (2005). Assessment literacy. In M. H. Rosenbusch (Ed.), *New visions in action: National assessment summit papers* (pp. 11–23). Ames, IA: Iowa State University.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, McNamara, T., O'Loughlin, K.. (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 126–136). Cambridge, UK: Cambridge University Press.

Cox, T. L., Malone, M.E., & Winke, P. (2018). Future directions in assessment: Influences of standards and implications for language learning. *Foreign Language Annals*, *51*,104-115.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181-199.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24*(2), 81-112.

Draper, J., & Hicks, J. (2000). Teaching heritage languages: The road traveled. In J. B. Webb & B. L. Miller (Eds.), *Teaching heritage language learners: Voices from the Classroom* (pp. 14-38). New York: American Council of Teachers of Foreign Languages.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Haviland, D., Shin, S., & Turley, S. (2010). Now I'm Ready: The impact of a professional development initiative on faculty concerns with program assessment. *Innovative Higher Education, 35*(4), 261–275.

Hildén, R., & Fröjdendahl, B. (2018). The dawn of assessment literacy – exploring the conceptions of Finnish student teachers in foreign languages. *Apples: Journal of Applied Language Studies, 12*(1), 1–24.

Hurlbut, S. L. (2015). *Exploring foreign language classroom-based assessment through subject-specific professional development: A participatory action research case study*. (Publication No. 3717957) [Doctoral dissertation]. ProQuest.

Ilieva, G. N. (2012). Hindi heritage language learners' performance during OPIs: Characteristics and pedagogical implications. *Heritage Language Journal*, *9*(2), 18-36.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*(3), 385-402.

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals*, *36*(4), 483-490.

Liskin-Gasparro, J. E. (1984). The ACTFL proficiency guidelines: Gateway to testing and curriculum. *Foreign Language Annals*, *17*(5), 475-489.

López Mendoza, A. A., & Bernal Arandia, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *Profile Issues in Teachers Professional Development*, *11*(2), 55-70.

Magnan, S. S. (1991). Just do it: Directing TAs toward task-based and process-oriented testing. In R. V. Teschner (Ed.), *Assessing foreign language proficiency of undergraduates. Issues in language program direction: A series of annual volumes* (pp. 135–162). Boston: Heinle.

Malone, M. E. (2008). Training in language assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education, Vol. 7: Language testing and assessment* (2nd ed., pp. 225-233). New York: Springer Science and Business Media.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329-344.

Malone, M. E. (2017). Developing Instructor Proficiency in Oral Language Assessment. In N. Van Deusen-Scholl & S. May (Eds.), *Second and Foreign Language Education. Encyclopedia of Language and Education* (pp.439-452). Springer International Publishing.

Malone, M. E., & Montee, M. J. (2010). Oral proficiency assessment: Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass, 4*(10), 972-986.

Martin, C. (2010). Assessing the oral proficiency of adult learners, "heritage" and "native" speakers using the ILR descriptors and ACTFL guidelines: Considering the challenges. *Russian Language Journal, 60,* 167-181.

Norris, J. M., & Pfeiffer, P. C. (2003). Exploring the uses and usefulness of ACTFL oral proficiency ratings and standards in college foreign language departments. *Foreign Language Annals*, *36*(4), 572-581.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912.

QSR International (n.d.). NVivo qualitative data analysis software (Version 12). QSR International Pty Ltd. Retrieved from https://www.qsrinternational.com/nvivo/home

Shaha, S. H., Glassett, K. F., & Ellsworth, H. (2015). Long-term impact of on-demand professional development on student performance: A longitudinal multi-state study. *Journal of International Education Research*, 11(1), 29-34.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing*, *30*(3), 309-327.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational measurement: Issues and Practice*, *18*(1), 23-27.

Strauss, A. & Corbin, J. (1994). Grounded theory methodology: an overview. In: Denzin, N.K., Lincoln, Y.S. (Eds.), *Handbook of Qualitative Research (*pp. 273–285). Thousand Oaks: Sage Publications.

Swanson, P. (2012). Second/foreign language teacher efficacy and its relationship to professional attrition. *Canadian Modern Language Review*, *68*(1), 78-101.

Taras, M., & Davies, M. (2017). Assessment beliefs of higher education staff developers. *London Review of Education*, 15(1), 126–140.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21-36.

Terry, R. M. (1992). Improving inter-rater reliability in scoring tests in multi-section courses. In J. C. Walz (Ed.), *Development and supervision of teaching assistants in foreign languages. Issues in language program direction: A series of annual volumes* (pp. 229–262). Boston: Heinle.

Valdés, G. (1989). Teaching Spanish to Hispanic bilinguals: A look at oral proficiency testing and the proficiency movement. *Hispania*, *72*(2), 392-401.

Vogt, K., Tsagari, D., Csépes, I., Green, A., & Sifakis, N. (2020). Linking learners' perspectives on language assessment practices to teachers' assessment litTeracy enhancement (TALE): Insights from four European countries. *Language Assessment Quarterly*, *17*(4), 410-433.

Wright, H.P., Foran, A.R., Holmes, H. & Lou, Y. (2016). Classroom assessments: What type of learning is being measured? Paper presented at American Educational Research Association, Washington, DC.

Xu, Y., & Brown, G. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*(C), 149–162.

Appendix A

Description of the codes used to analyze the interview data

Code	Description
Authenticity	Data coded as "Authenticity" indicate that participants are thinking about the extent to which teaching and learning align with the 'real-world' language use.
Agency	Data coded as "Agency" indicate that the participants feel a lack of personal control in affecting course design, syllabi, or lesson planning because of external factors and may or may not discuss ways of overcoming that challenge
Curriculum, Syllabus	Data coded as "Curriculum and Syllabus" indicate that participants have made general changes to their curriculum or syllabus. Participants do not provide any detailed information about what aspect of their curriculum or syllabus they changed, however.
Goals, Objectives	Data coded as "Goals and Objectives" indicate that participants are thinking in novel ways about or have made changes to the goals and objectives associated with teaching and learning. Goals and objectives may relate to (a) broader, course-level, (b) individual unit or class goals, or (c) students' own learning.
Teaching	Data coded as "Teaching" indicate that participants have, in some way, made changes to their teaching practice on account of having attended the program. Changes may relate to course/unit structure, teaching methodology, or activities.
Materials	Data coded as "Materials" indicate that participants have made changes to the teaching and/or learning materials in any or all of the courses they teach.
Testing, Assessment	Data coded as "Testing" indicate that participants have made changes to their testing programs. Changes may be about (a) placement, (b) classroom formative or summative assessment, or (c) rubrics used to rate or score performances.
Feedback	Data coded as "Feedback" indicate that participants are increasingly aware of or have altered how they provide feedback about language use to students.
Novel ideas	Data coded as "New Ideas" suggest that participants, as a result of their participation in the program, have new ideas for changes they would like to make to any aspect of their program, including goals/objectives, materials, teaching, or testing.
Student levels	Data coded as "Student Levels" indicate that participants have gained confidence in their ability to rate students' performances (a) at a particular level (e.g., Novice-mid, Intermediate-low, etc.) or (b) in relation to can-do statements or benchmarks.

Program	Data coded as "Program" consist of either (a) praise for the program or (b) recommendations and/or requests for ongoing program work.
Student reactions	Data coded as "Student Reactions" consist of data that document students' reactions to changes that their instructors have made to any aspect of their teaching and learning practice.

Appendix B

Inter-rater reliability in coding themes

Themes	% of agreement	
	2018	2019
Authenticity	94.72	99.17
Curriculum/syllabus	98.34	98.66
Feedback	99.79	96.07
Goals, objectives	99.04	98.49
Materials	99.31	97.98
Student levels	99.13	94.84
Student reactions	96.74	98.24
Teaching	94.99	90.60
Testing, assessment	95.44	93.34
PD workshop/program	96.70	97.52
Novel ideas	98.25	97.56
Agency	-	97.77

Note. The inter-coder reliability ranged between 90-99%. The code "Agency" was introduced in 2019.

4

Student Mathematics Achievement in a Mandarin Immersion Program

Sean R. Hill
Central Michigan University

Challenge Statement

Dual language programs have increased in popularity and it is difficult to determine whether academic differences exist among students due to placement within immersion or traditional classrooms. This analysis uses multilevel modeling and a pre-kindergarten assessment to examine factors related to differences in test scores.

Abstract

With the rise in dual language immersion programs across the United States, researchers and educational leaders have examined the academic achievement of students in these programs with increasing scrutiny. Investigations have reported that there may be an initial decline in immersion student performance relative to their peers in traditional classrooms. This is then typically followed by immersion students reaching parity and sometimes surpassing their peers. Unfortunately, many studies report on academic achievement without carefully controlling for student differences prior to beginning their elementary education. This study examines student performance in a rural school district marked by generational poverty that implemented a one-way Mandarin Chinese immersion program through hierarchical linear or multilevel modeling (HLM). Initial analyses indicated superior performance by students in the immersion group in mathematics, with variables such as free-and-reduced lunch status, special education status, and gender playing a significant role in the prediction of academic achievement. However, once a pre-kindergarten assessment was included in the model, the immersion advantage disappeared. This indicated that the primary predictor of student achievement between the groups was aligned to their preparation and development prior to kindergarten, and those differences

persisted over the consecutive grade levels. Although HLM is an appropriate tool for missing data, absent pre-kindergarten screening scores and progress monitoring limited the generalizability and confidence of these findings.

Keywords: academic achievement, Mandarin, Chinese, dual language immersion, multilevel modeling, hierarchical linear modeling

Introduction

The number of dual language immersion programs has experienced dramatic growth across the United States over the past two decades (Center for Applied Linguistics [CAL], 2011; Steele et al., 2017; Watzinger-Tharp et al., 2021). It is difficult to ascertain the number of programs that exist today because there is no national registry of these programs. CAL's last update occurred in 2011, and there were 448 schools listed with immersion programs at that time. Utah alone boasted approximately 250 programs for the 2020-2021 academic year (Utah State Board of Education, 2021). With other states doubling their immersion offerings (Steele et al., 2017), there are estimated to be thousands of programs now across the country.

There are many different program models that fit under the umbrella term of dual language immersion. Across all models, however, the main goal of these programs is to ensure that students receive 50-100% of their core instruction in the target language (TL) in order to have strong language skills in two languages by the end of high school. Some universities have even now begun to teach 300-level college courses to high school students (Bowman et al., 2020) and students can complete a language minor shortly after entering college. Many of the early studies on immersion examined students' proficiency in the TL (Bild & Swain, 1989; Genesee & Jared, 2008; Swain & Lapkin, 1982) and that examination continues today (Padilla et al., 2013; Parks, 2020). Similarly, early studies sought to scrutinize whether the time spent in the TL adversely affected the development and proficiency of students' first languages. Soon after, however, many investigators began to examine student academic achievement in other content areas over a variety of languages and with a variety of methods (Artzer, 1990; Hill, 2018; Padilla et al., 2013; Steele et al., 2017; Thomas et al., 1993; Watzinger-Tharp et al., 2018; Willig, 1985). Findings typically indicated that students in immersion programs have initially depressed academic achievement, but over a few years scored on par with, and sometimes better than, students in traditional classrooms.

The accuracy of these claims can be called into question for multiple reasons. Because the continuation of the immersion programs and related funding are often dependent on student achievement, it is important to note that at times immersion students are offered additional resources relative to students in traditional classrooms (Essama, 2007). This speaks to an underlying concern with making comparisons between students in immersion programs and their peers. Students may come into school having had disparate opportunities and the inclusion of additional resources can exasperate the differences between potentially unequal groups. This concern is apparent when researchers make use of propensity matching (Watzinger-Tharp et al., 2016) or the examination of schools that offer

enrollment in immersion programs through lottery systems (Steele et al., 2017). Propensity matching is a quasi-experimental statistical technique that creates an artificial control group matched to a treatment group on important characteristics to approximate experimental design and causality effects from an intervention. Lotteries make it impossible to know if there are truly equal groups because parents typically must express a desire to enroll their children in an immersion classroom in order to be placed in an enrollment lottery system. This desire to enroll students in a specific academic program may act as a proxy variable for parental investment in education and create unequal groups prior to formal K-12 education. This parental investment, therefore, can also manifest throughout a child's educational career as additional resources, such as content-specific tutoring support, which could moderate small initial differences in kindergarten readiness or depressed performance in the early grades. Parents that do not express a desire to enroll their children in an immersion program may differ from those that do, and those differences have not been tracked. An additional difficulty present in many school situations is the inconsistency with which data is collected. When students are absent from school, and traditional data approaches like regression and the ANOVA family of tests exclude students with missing data from analyses.

The current investigation examines student performance in a school that maintained a one-way partial Mandarin Chinese immersion model over four years. A one-way (or foreign language) dual language immersion program consists of students whose first language is not the TL. In this situation, all of the students spoke English as their first and primary language. As a 50/50 partial immersion model, the instructional day was evenly divided between Mandarin Chinese and English. This study is also unique because it examines one small, rural elementary school in Michigan with marked generational poverty and includes pre-kindergarten student achievement scores. This offers an opportunity to examine student mathematics achievement and growth, controlling for group differences in kindergarten readiness. This investigation hypothesizes that students in the immersion program perform better than their peers in mathematics performance. In addition, it is expected that students that are placed in the immersion program upon parent request are better prepared for kindergarten, and will also score higher on yearly mathematics assessments. These group differences will likely persist throughout their placement in the immersion program but are expected to be moderated factors such as free-and-reduced lunch and special education status.

Literature review

L2 instruction and Student L1 Development

The research boom around language immersion programs in the United States typically pays homage to the first Canadian French language immersion programs in the 1960s. While the recent explosion in US language immersion programs are sought after by parents endeavoring to increase the social capital of their children (Williams, 2017), bilingual education was seen as detrimental to language development in the 1960s and beyond. This is why the success of the students in achieving proficiency in two languages in those first Canadian programs was so

important to combating the idea of the bilingual deficit hypothesis (Oller et al., 1997). The bilingual deficit hypothesis posited that bilingual students do not reach a high level of proficiency in either their first or second language. For example, they may have less developed vocabularies in both languages and rely on code-switching to fully convey a message. Specifically, the full linguistic repertoire of a bilingual student in this hypothesis is not honored as it is via a translanguaging lens (Wei, 2018). Indeed, MacNamara (1966) introduced the balance hypothesis, which described a zero-sum game for language development. Children who learned a language other than English in the household were expected to have poorer English fluency than children that solely heard and spoke English. Effectively, MacNamara (1966) argued there was a cap on language development, and increases in fluency in one language led to decreased fluency in another. Hence, language immersion programs had been expected to decrease the overall linguistic growth of students.

While the studies looking at language fluency suffered from serious methodological flaws and linguistic elitism (Cummins, 1977), the performance of the anglophone children in French immersion began to lead to what has been called the bilingual advantage hypothesis (Fortune & Tedick, 2003; Oller et al, 1997). When tested, these children eventually reached parity with their counterparts in English-only classrooms, and some French immersion students, unexpectedly, performed better in English literacy development than peers from English-only programs despite having spent most of their early grades immersed in the target language (TL). This led to the proposal of the Linguistic Interdependence Hypothesis (Cummins, 1979). Commonly referred to as the two ice-berg model, Cummins envisioned a substrate of linguistic knowledge that is shared between two languages. The part of the iceberg above the water for each language refer to the surface features of each language, which demarks vocabulary, grammar, and syntax. Within this conceptualization, students learning through the TL could also better understand and contrast the surface features of their first language. In the shared underwater level, students could increase their performance in English literacy development while learning content through French because of the shared linguistic knowledge. The surprising parity of the French immersion students to their English-only peers led to the first examinations of immersion student fluency and literacy development in both English and French across many school districts (Swain & Lapkin, 1982). This focus on immersion student performance relative to their peers in traditional classrooms has continued to be a hot topic both in the United States and abroad (Bild & Swain, 1989; Genesee & Jared, 2008; Lo & Lo, 2014; Padilla et al., 2013; Parks, 2020; Steele et al., 2017; Watzinger-Tharp et al., 2016).

The Rise of DLI Relative to Full Immersion Programs

Educators, researchers, and parents quickly wanted to examine the performance of students in English and other academic areas after having received much of their instruction in the TL. Programs labeled partial immersion appeared to allay the fears that students were not receiving enough instruction in English, and were not prepared for high stakes assessments (Fortune & Tedick, 2003; Pipia,

2013). Partial immersion programs differentiated themselves with the total or full immersion programs because the goal was to present content to students in a 50/50 model, where half of the school time was instructed in the TL and the remaining content was taught in English. In fact, this move to less contact time in the TL was reported as detrimental to long-term student performance relative to the gains that students demonstrated in full or total immersion programs (Bournot-Trites & Reeder, 2001). Therefore, higher overall performance was expected for immersion students that spent more time learning content through the TL.

Regardless, the current dual language immersion (DLI) model, especially as designed in Utah (Watzinger-Tharp et al., 2016), is becoming increasingly common and maintains 50% instruction in English and the TL. This is especially relevant for schools that face the pressures of high-stakes testing and state-level assessments that pit schools against each other in top-to-bottom lists, wherein the bottom five percent of schools faced potential closure and the firing of all staff (Higgins, 2016). Further, if students register deficits in English language arts, for example, they can also be retained in third grade (Westall et al., 2022). These consequences are even more poignant when a summary of ten years of the academic effects of the early Canadian programs found that immersion student performance was typically suppressed relative to their peers in traditional classrooms in the early grades before eventually reaching parity in later grades (Swain & Lapkin, 1982). Effectively, there are very real consequences for students, teachers, administrators, and schools where performance lags among early immersion students in the full or total models, so it is understandable why the current 50/50 model is a common program choice, therefore, it is important to understand how these programs affect achievement in different content areas.

Mathematics Achievement in Immersion Programs

While the central focus of the immersion achievement research has examined linguistic outcomes of English and the TL, the examination of student achievement in other content areas has become increasingly important. Mathematics and English language arts achievement, for example, are the primary measures of student achievement in the early elementary grades. Therefore, it is important to examine the prior findings of immersion students' mathematics achievement. While many studies detail the benefits of immersion education on student achievement in mathematics (Arthur, 2004; Artzer, 1990; Essama, 2007; Jacobsen, 2013; Mukai et al., 2005; Marian et al., 2013; Padilla et al., 2013; Strickland & Hickey, 2016; Thomas et al., 1993), others find no effect for immersion education on mathematics achievement or growth (Haj-Broussard, 2005; Thomas et al., 1993; Watzinger et al., 2016). Even a finding of no significance for immersion programming indicates that immersion students perform as well as those from traditional classrooms.

On the other hand, some researchers indicated that students in immersion programs experience decreased performance in courses taught through the TL (Lo & Lo, 2014). However, Lo and Lo (2014) did not find decreased achievement in mathematics taught in the TL for immersion students. This contradicted the

finding that math performance was indeed lower for students in immersion groups relative to peers in traditional classrooms (Thomas et al., 1993). To better assess the mixed findings for math achievement, Hill (2018) performed a meta-analysis that examined student achievement across ten different programs which provided enough information to calculate effect sizes. For math achievement, there was a small but significant ($p < .001$) mean effect size (Hedge's *g*) of 0.20 with a fixed effects model and 0.36 with the random effects model. The I^2 was large, however, which indicated that there was considerable underlying variability associated with those differences. Despite those results, a failsafe *K* analysis indicated that only a few studies would be required to be included that would negate the positive effect size associated with immersion students' math achievement scores. It is plausible that studies that demonstrated this negative finding may not have been published because schools with underperforming programs would have quickly ended those programs due to concerns with high stakes testing and consequences for funding.

Lastly, the analyses surrounding the academic achievement of students in immersion programs suffers from one particular and noteworthy flaw, the potential of unequal groups. Additionally, the absence of student achievement data prior to entrance in immersion education makes testing pre-kindergarten differences difficult, and this is not captured in most datasets. Indeed, multiple authors have noted that immersion programs can quickly become gentrified as parents seek to build social capital in their children, and programs that began as two-way programs often end as one-way programs, often pushing out students and families from lower socioeconomic backgrounds (Dorner et al., 2021; Gándara, 2021, Valdez et al., 2016; Williams, 2017; Woody, 2020). This is especially important because this supports the idea that immersion programs and the comparison of academic achievement of students within those programs rests on groups that were unequal from the beginning. If immersion students tend to be wealthier (Chung, 2020; Parkes, 2008; Valdez et al., 2016) and have more parental support for their educational development, it is plausible that their performance upon entrance to kindergarten is likely to be higher than comparison groups. This performance bump, therefore, could mitigate the effects of the initial suppression on test scores and lead to growing disparity in performance over the years.

For instance, Watzinger-Tharp et al. (2016) were very cognizant of this possibility and they "used rigorous statistical methods to ensure an equitable comparison of students participating in DLI with students not in DLI" (p. 7). They utilized individual students as comparisons against themselves for achievement gains and also made use of propensity matching for comparisons. They also used multi-level regression analysis to help control for student- and school-level effects. Even though they showed no results for immersion programming on the effect of third-grade math achievement, they did lament that "no prior math achievement scores were available to permit us to rule out the possibility that DLI and non-DLI students differed prior to participation" (p. 11). Their analysis of fourth grade results used third-grade performance in math and English language arts as a control, and they did find a small positive effect associated with DLI programs for math achievement.

Similarly, Steele et al. (2017) were cognizant of the problem with initial group differences and examined students' achievement with immersion students chosen through a lottery system. Their findings did not indicate that mathematics performance was different between DLI and non-DLI groups. However, it was still possible that there were initial group differences among the parents intending to enroll their children in language immersion programs relative to those that did not intend to enroll children in immersion programs. As indicated above, it is possible that parents from higher socioeconomic backgrounds were more interested in increasing social capital in their children than the parents who did not choose to enroll in the lottery for access to those programs (Dorner et al., 2021; Gándara, 2021, Valdez, et al., 2016; Williams, 2017).

Purpose and Research Questions

The current study is vital because it examines student mathematics performance and controls for student achievement prior to kindergarten and the beginning of immersion education. Further, yearly progress monitoring data is available for students in kindergarten through third grade, allowing a glimpse into early performance over time. Therefore, math performance can be examined between immersion and traditional students and the pre-kindergarten screening score can be used to examine children from similar backgrounds in the two program models. Anecdotally, some parents were also unaware that they had checked the box for their children to participate in the immersion program, which could slightly mitigate the unequal groups problem. The multi-level regression analysis with HLM also allows for missing data, whereas traditional regression or ANOVA analyses require the removal of students with missing data from analyses. Further, the multi-level modeling allows longitudinal student scores to be nested within each student, so that their scores can be analyzed over multiple time periods. Contrary to a repeated measures ANOVA, the assumption about equal time periods between tests does not have to be met. This is particularly important for data sets in K-12 because student absences often lead to disparate testing dates. Lastly, the nesting of data is an important characteristic of multi-level modeling because the assumption of independence of samples would yield inflated effects and increased standard errors in a one-level regression.

The purpose of this study is to determine the effect that a 50/50 one-way dual language Mandarin immersion program has on student academic achievement in mathematics. In particular, this investigation will evaluate whether program type (immersion vs. traditional) has an effect on academic achievement in mathematics, after controlling for the effects of free-and reduced lunch status (a marker for socioeconomic status), special education status, gender, and grade level. Additionally, the inclusion of group differences prior to enrollment in kindergarten is possible with the inclusion of a kindergarten readiness screen. This variable is absent from most of the available research in the field. The study will then explore these potential effects between the program type and other variables through the use of hierarchical linear (multilevel) modeling. Specifically, the research questions are:

1. Do students in a Mandarin Chinese immersion program perform better over time in mathematics relative to peers in traditional classrooms?
2. Do program differences disappear when controlling for pre-kindergarten readiness?

This investigation expects to find a positive program effect on mathematics achievement for students enrolled in the Mandarin immersion program, but that effect is expected to be moderated and mediated by other variables. For example, it is expected that students that receive free-and-reduced lunch and special education services will demonstrate decreased performance in both the immersion and traditional groups. It is also hypothesized that students that score higher on pre-kindergarten readiness will demonstrate higher mathematics achievement over the years.

Methods

Research Design

This investigation is a quasi-experimental design and utilizes a panel and cohort sequential design. While the population of students in the immersion program remains a constant cohort, there is more variation in the traditional classrooms because the population of the school is highly mobile. The purpose of the design is to compare the mathematics achievement between immersion and traditional students, controlling for the effects of other factors. The data collection occurs longitudinally over multiple years from the 2010-2011 school year to the 2013-2014 school year. While the first cohort of student scores occurred a decade ago, the fact that there is early yearly data and a pre-kindergarten screen provides an opportunity for a candid analysis of mathematics growth with the ability to examine initial group differences prior to formal kindergarten education. Additionally, the purpose is to build a model that predicts student performance between the two groups.

Population

This investigation examines the achievement scores of 427 kindergarten through third grade students in a school that offered a one-way Mandarin dual language program in addition to a traditional program housed within the same building. There were 92 students in the kindergarten cohort, 113 in the first-grade cohort, 118 in the second-grade cohort, and 104 in the third-grade cohort. Students in the third-grade immersion groups had been learning Mandarin since their kindergarten year. By comparing school district students housed within the same building, albeit in different programs, it is expected that the student groups will be more similar than they would be by comparing an immersion school to a traditional school within or across school districts. While parents choose whether to place their children in the immersion program or not, there are no building and district level effects to confound the results. This contrasts with much of the previous research that compares students across multiple buildings and districts without accounting for those effects.

In this immersion model, students spent approximately half the instructional day in English with an American teacher and the remainder of the day in Mandarin with a Chinese teacher hired from the Confucius Institute at Michigan State University. The Confucius Institute provided the same training model and coursework for all its foreign Chinese teachers, in addition to professional development for the elementary

school's administrator and American paired teachers. Students in the traditional program spent all day with the same instructor and received 100% of their instruction in English, unless they had a world language enrichment course. Students' enrichment was a weekly session in Spanish except for the 2010-2011 year when the enrichment language was Mandarin. The immersion students learned mathematics and science during the Mandarin portion of the day in addition to Chinese language arts. Social studies and English language arts were taught with the American teacher. All students attended the same enrichment courses throughout the year: art, music, physical education, and Spanish.

Sample

A census was used in order to obtain a sufficient sample size to make comparisons of performance among demographic subgroups. All student scores were examined within the school building. Because a census was used for data collection, there was no randomization in sampling. The sample school is effectively a convenience sample because it was chosen due to its location close to the researcher and previous affiliation allowed favorable access to student data. However, there are very few schools in Michigan that meet the program description described in the previous section, and this provided the opportunity to investigate the program model with a high percentage of rural students who live in poverty. The obtained results and explanatory models may generalize to additional schools, but these models and conclusions will have to be investigated in multiple sites to determine their validity across multiple population samples.

Data Collection

The school was contacted via a letter addressed to the superintendent, principal, and school board president asking for permission to examine the academic achievement of students enrolled in the immersion program relative to students enrolled in traditional classrooms. The researcher petitioned administration for the achievement data and demographic information listed in the variables section. The school compiled an Excel file for the researcher that contained no individually identifiable information from the data storehouse used.

Variables

There are many factors that can contribute to student achievement on progress monitoring and high stakes testing. Although program type will be analyzed based on student performance in mathematics, additional variables are added to examine a predictive model for achievement. Table 1 delineates the variables that were used in the study and how they were coded.

Independent Variables

The principle independent variable of interest is a dichotomous categorical behavioral for program type: immersion or traditional. Student demographic information includes dichotomous coding for free or reduced lunch status (as a marker for socioeconomic status), special education status, and gender. Additionally, grade level was also included and it was on an ordinal scale from zero to three. The Brigance Pre-kindergarten screen was administered to some students prior to the

beginning of kindergarten and is described below. Other variables that had been considered were race/ethnicity, first language, and previous Mandarin instruction. It was decided to not include those variables because there were no or too few students that differed from the majority.

Table 1
Variables, Data Type, and Operationalized Coding

Variables	Type	Operationalization
Independent		
Program Type	Categorical	Coded: 0 = Traditional, 1 = Immersion
Free or Reduced Lunch Status	Categorical	Coded: 0 = No, 1 = Yes
Special Education Status	Categorical	Coded: 0 = No, 1 = Yes
Gender	Categorical	Coded: 0 = Female, 1 = Male
Grade Level	Ordinal	Coded: 0 = Kindergarten, 1 = first grade, 2 = second grade, 3 = third grade
Brigance	Continuous	Scales from 0 to 100
Dependent		
MEAP Mathematics	Continuous	Scale*
Star Math	Continuous	Scale from 0 to 1400

Note. *Possible scale total change yearly and by grade level. For the 2013-2104 school year, the MEAP mathematics maximum is 412 points and the minimum is 203 points. (Michigan Department of Education, 2013)

The distribution of students is presented in Table 2 where the percentage of students in each category is described.

Table 2
Dataset Aggregate Frequencies

Classification	*N* (Percent)
Immersion Classification	
Traditional	268 (62.8%)
Immersion	159 (37.2%)
Free/Reduced Lunch Classification*	
No	127 (29.8%)
Yes	299 (70.2%)

Special Education Classification*	
General Services	374 (87.8%)
Special Education Services	52 (12.2%)
Gender*	
Female	230 (54.0%)
Male	196 (46.0%)
Grade Level	
Kindergarten	92 (21.5%)
First Grade	113 (26.5%)
Second Grade	118 (27.6%)
Third Grade	104 (24.4%)

Note. Math assessments $N = 427$. *One missing identification.

Dependent Variables

This investigation utilizes the end-of-year student achievement data in mathematics from progress monitoring from STAR Math, which is described below. The scores on these assessments are continuous and are assessed three times per year. Data collection only allowed for one progress monitoring score per year per student, and the researcher chose the end-of-year scores.

Instruments

The results from three principle instruments were examined for this investigation. The instruments provide state summative assessment results as well as formative assessment data from local progress monitoring. The progress monitoring instrument used in the elementary school was Star Math and the state-mandated high stakes assessment was the Michigan Educational Assessment Program (MEAP). Lastly, students also completed Brigance Screens to gauge readiness for kindergarten. The STAR Math assessment was utilized each school year, which made it possible to complete an apple to apples comparison. The Brigance assessment utilized a different scale completely and was a global assessment of student preparation for kindergarten. The score for the Brigance was utilized to help predict student achievement on the STAR Math assessment in conjunction with the other independent variables. Each assessment is described below.

Star Math

The Star Math assessment was used for progress monitoring in the elementary school to assess mathematical operations. The test is computer-adaptive; the program adjusts subsequent questions based on student responses on previous problems. The reported overall reliability is very high ($r = .98$); although there is a range between .89 and .94 provided over different grade levels. The relationship of

student progress measured in slope is very high (r = .93). Validity has been established with moderate to high correlations of predictive validity (rs = .63 to .87) with the Michigan Educational Assessment Program for first through sixth grade from 2002-2016. In general, it is well correlated to other tests of mathematics proficiency, r = .758 (Renaissance Learning, 2022).

MEAP

The Michigan Educational Assessment Program (MEAP) was a mandated high stakes test required for students between third and ninth grade. The results were made public and school quality was judged, in part, based on this assessment. In the third grade, the MEAP had three principle tests: reading, writing, and mathematics. The MEAP was discontinued after the 2013-2014 school year and was replaced with the Michigan Student Test of Educational Progress (M-STEP) and tested third grade students on mathematics and English language arts. The mathematics portion of the test offered high reliability, $\alpha > .90$ (Burns, 1998). Regardless, this was the only statewide assessment of student achievement and was a principle tool used for evaluating school districts in the state.

Brigance

The Brigance Screens that were utilized as a pre-screening assessment prior to beginning kindergarten were administered to students. It is a tool that gauges the child's development and is used to measure and predict students' success in school. There are three areas of developmental testing comprising physical, language, and cognitive/academic domains. The Brigance is used to assess school readiness, as well as to examine potential developmental delays, and is both norm- and criterion-referenced. The academic domain correlates highly with the Mullen Scales of Early Learning composite score, r = .85, whereas the physical and language domains correlate moderately with the Vineland-II, rs = .53 and .68, respectively. The average internal consistency was high, α = .96. The test-retest reliability ranges from rs .84 to .91. Lastly, interrater reliability is high as well, r = .93 for the ages of children entering kindergarten (French, 2013).

Table 3 indicates which assessment was administered to each grade level cohort by grade level. It is important to note that some assessment data is missing depending on the year because it was not required by school or state policy at the time.

Table 3

Assessment Schedule and Variation by Grade Level over Multiple Years

	Mathematics Assessments					
Cohort	Brigance (pre-kinder-garten)	End of kinder-garten	End of 1st grade	End of 2nd grade	End of 3rd grade	MEAP: Beginning of 3rd grade
3rd grade	None	None	SM	SM	SM	Yes
2nd grade	Yes	None	SM	SM		
1st grade	Yes	SM	SM			

Kinder-garten	Yes	SM

Note. SM is the Star Math assessment. Despite classification of an affirmative Brigance assessment being administered in the first grade, only 17% of students had recorded results. Further, up to 2/3 of students missed other end-of-year assessments.

Table 4 reports the cleaned data set for the math assessments. The cleaning also including the removal of extreme *Z* scores (> |+/-3.00|). These included one score in the second grade and kindergarten cohorts, as well as 4 scores from the first-grade cohort.

Table 4

Cleaned Data Math Assessments and Disaggregation by Variables by Grade Level

Classification	3rd Grade	2nd Grade	1st Grade	Kindergarten
Immersion Classification				
Traditional	62 (59.6%)	80 (67.8%)	76 (67.3%)	50 (54.3%)
Immersion	42 (40.4%)	38 (32.2%)	37 (32.7%)	42 (45.7%)
Free/Reduced Lunch Classification*				
No	34 (33.0%)	31(26.3%)	34 (30.1%)	28 (30.4%)
Yes	69 (67.0%)	87 (73.7%)	79 (69.9%)	64 (69.6%)
Special Education Classification*				
General Services	87 (84.5%)	106 (89.8%)	99 (87.6%)	82 (89.1%)
Special Education Services	16 (15.5%)	12 (10.2%)	14 (12.4%)	10 (10.9%)
Gender*				
Female	66 (64.1%)	59 (50.0%)	62 (54.9%)	43 (46.7%)
Male	37 (35.9%)	59 (50.0%)	51 (45.1%)	49 (53.3%)

Note. $N = 427$. *One missing identification, $n = 426$.

Table 5 (following page), on the other hand, reports the standardized values for the different assessments. There is no reporting of mean or standard deviation because the standard mean score is equal to 0 and the standard deviation is +/- 1. Measures of skewness indicate that after the removal of a few students' scores; all remaining scores fell within accepted ranges of +/- 1.000.

Table 5
Standardized Aggregate Assessment Results

Assessments in the Aggregate	*N*	Potential *N*	Missing	Min	Max	Skew	Kurtosis
Brigance	**142**	**323**	**56.0%**	**-1.92**	**1.97**	**-.092**	**-.903**
Kindergarten math assessment	**109**	**205**	**46.8%**	**-2.78**	**1.97**	**-.551**	**.351**
1st grade math assessment	**264**	**335**	**21.2%**	**-2.77**	**2.58**	**-.165**	**-.030**
2nd grade math assessment	**194**	**222**	**12.6%**	**-2.70**	**1.80**	**-.339**	**-.571**
3rd grade MEAP Mathematics	**90**	**104**	**13.5%**	**-1.94**	**2.64**	**.303**	**-.089**
3rd grade math assessment	**92**	**104**	**11.5%**	**-2.03**	**1.68**	**-.196**	**-1.113**

*Note***.** The total dataset includes 427 students. The MEAP data is included between the second and third grade math assessments because that is where it fell chronologically. Each reported assessment for progress monitoring is from the end of year, whereas MEAP was administered in the first semester of third grade.

Analytical Method

Regardless of the question of sample generalizability, student achievement results at this school may differentiate between the groups and describe a fixed effect for program type. One of the purposes of this comparison is to determine the extent to which program type (immersion versus traditional) causes variation in student achievement. Causality cannot be completely determined without randomization of student program enrollment. However, an inspection of the reduction in model deviance and the beta weights will allow program type to be examined with respect to the other variables. Further, other variables could contribute greater influence in student achievement scores than program type. It is important to note that although the dataset is older, the difference in student achievement scores still proves fruitful in the examination of student performance, especially with the ability to control for academic readiness prior to kindergarten.

Hierarchical linear modeling (HLM), also known as multi-level regression, was used with the multiple assessment scores for each student at level one. While it is common in HLM for individual students to be listed at level one with a score and their demographic data, the multiple level one assessment scores are effectively housed within each student. This would be analogous to each student within a nested classroom. Each assessment that a student was administered was nested within each individual student, so effectively there are repeated measures for each student (Woltman et al., 2012). This can indicate a powerful nesting effect because the samples are not independent, rather multiple assessments are housed within each student. The assumption of independent

sample would yield highly inflated effects with a one-level regression. This creates the need for HLM to accommodate the nested data. Therefore, level one test scores act as if they were individual students in typical HLM analyses. Level two consists of the program type, as well as the other demographic predictor variables of each student. In this use of HLM, the individual student would be the equivalent of a grouping, like a classroom would be when testing for teacher-level effects in typical HLM analyses. Even though some students had fewer test scores to examine than others, having a larger number of groups with fewer repeated measure testing points is not a problem with HLM. Woltman et al. (2012) indicated that "it is advantageous to increase the number of groups as opposed to the number of observations per group" (p. 56). HLM, therefore, is a useful tool when there are 427 individual students that yield 891 different test scores.

The hierarchical linear modeling approach is better suited for this analysis over a repeated measures ANOVA or ANCOVA model because the scant data is not lost due to elimination of a student from analysis when there is missing data. This is a primary benefit for an HLM analysis (Snijders, 1996). Another benefit is that there is more flexibility in the use of assessments. Assumptions are not broken when student assessments are held at different times. If progress monitoring dates differed by a week or two between one student and another, which is often the case in educational data in schools, a repeated measures ANOVA would not be appropriate. Further, "HLM can accommodate nonindependence of observations, a lack of sphericity, missing data, small and/or discrepant group sample sizes, and heterogeneity of variance across repeated measures" (Woltman et al., 2012, p. 56). Figure 2 indicates that the variance is similar across each of the testing periods. HLM is a robust analytical tool that accounts for both missing data and uneven testing schedules. One final benefit of multilevel modeling includes the ability to examine how slopes vary among groups, whereas the ANCOVA cannot calculate these effects (Hox, 2010). Within this analysis, each student can have a slope that varies, i.e. individual students learn and achieve at different rates. Therefore, individual growth rates of each student can be taken into account because assessments are nested within individual students. The models that result may then be used with other schools offering similar programs to evaluate those models' generalizability.

Some assumptions that still need to be met to perform a multi-level regression involve normal data and tests for collinearity. Data were standardized to examine extreme *Z* scores and skew. One student's data was removed because it was greater than |+/-3| standard deviations. Additionally, all skew calculations yielded results less than |+/-1|. These two examinations indicated that the data from most measures was normally distributed. Table 7 reports those characteristics. Prior to beginning to conduct multilevel regressions, it was necessary to examine collinearity. Visual inspection of the independent variables proved difficult due to the binary classification noted above. Therefore, Spearman correlations were computed. The results are reported in Table 6 and indicate small but significant correlations between immersion program participation X free-or-reduced lunch status, immersion program status X special education status, and immersion program participation X gender. Students in the immersion program were slightly more likely to come from households with

higher incomes and conversely slightly less likely to receive special education services. The results also demonstrate that female students were slightly more likely to be enrolled in the immersion program relative to male students. There was no significant correlation among free-or-reduced lunch status, special education status, or gender.

Table 6
Correlations among Predictor Variables

	Immersion Program Participation	Free or Reduced Lunch Status	Special Education Status
Free/Reduced Lunch Status	-.169**		
Special Education Status	-.108*	.055	
Gender	-.114*	-.057	.150

Note. Math assessments $N = 426$. *Significant at .05. **Significant at .01.

The dataset was imported into SPSS from Excel for cleaning, but it was also necessary to create a long file from the short file. In the short file, individual rows represent each individual student record with columns for all the testing data and other demographic characteristics. The long file has a row for each individual assessment and an accompanying identifier that matches the corresponding student in the short file. Additionally, the long file includes a new variable labeled test number which indicates assessment order (0, 1, 2, 3). The long file consisted of 891 separate entries for assessment data with identifiers in the short file. The long file and short files were uploaded into HLM 8 to create the multilevel data matrix in order to perform the multilevel regressions (Raudenbush et al., 2019).

Another assumption that needed to be met to determine if HLM was more appropriate an analysis than a one-level multiple regression was that the data was appropriately nested. Therefore, it was necessary to calculate the intraclass correlation coefficient (ICC) before any analyses were performed. The ICC is a measure of nested data and is the group level variance at level two divided by the total variance at level one added to the group-level variance at level two. As long as the nesting effect accounts for more than ten percent of the variance in the data, then HLM is appropriate, although there is not an exact consensus on the required percentage (Woltman et al., 2012). This is an important assumption because "grouped observations from the same group are generally more similar to each other than the observations from different groups, and this violates the assumption of independence of all observations" (Hox, 2010, p. 14). In this analysis, each individual student is considered a group because the assessment data is effectively housed within each student. In this dataset, the ICC = 0.333, indicating that 33.3% of the variance was accounted for by the nesting of individual observations within students. Described another way, one-third of the variance in progress monitoring scores was explained by the same student being assessed multiple times. The remaining two-thirds of the variance in progress monitoring was accounted for by everything else. This ICC supports the use of HLM or multi-level regression over one-level multiple regression due to a nesting effect of scores with students.

Methodological Limitations

Schools represent a difficult situation for rigorous experimental research and are limited by real-world practicality. Although mandatory state assessment data is to be collected from students periodically, students often have absences that conflict with testing schedules. While schools are required to submit student answers for high stakes state assessments as part of the funding formula, building level progress monitoring assessments are not as closely aligned with funding allocations. Because of this, missing data is more likely to appear in progress monitoring. This is unfortunate because there are not also high stakes state assessments in the beginning grades. Therefore, missing data is a principle problem for group comparisons in this investigation. Although the multiple regressions and hierarchical linear modeling handle missing data better than the ANOVA family of tests, missing data is still problematic and detracts from the meaning of group similarities and differences. Significance tests may yield false negative results from the lack of data points. Likewise, results may also suffer from false positive effects due to the limited number of completed pre-kindergarten screens. As described above, only 17% of first grade students were administered the Brigance pre-kindergarten screen, even though it is listed as an assessment for that grade cohort. As evident in Table 7, few students in each grade-level cohort completed each yearly assessment.

Table 7

Missing Data in the Standardized and Cleaned Dataset

Cohort	N	Min	Max	Skew	Kurtosis	Missing
Third-Grade Cohort*	104					
Brigance	0					100.0%
Kindergarten math assessment	0					100.0%
1st grade math assessment	80	-2.36	1.74	-.442	-.522	23.1%
2nd grade math assessment	91	-2.09	1.58	-.153	-1.057	12.5%
3rd grade MEAP-Math assessment	90	-1.94	2.64	.303	-.089	13.5%
3rd grade math assessment	92	-2.03	1.68	-.196	-1.113	11.5%
Completed each math assessment	66					36.5%
Second-Grade Cohort**	118					
Brigance assessment	52	-1.78	1.97	.338	-.775	55.9%
Kindergarten math assessment	0					100.0%
1st grade math assessment	90	-2.77	2.58	-.306	.433	23.7%
2nd grade math assessment	103	-2.66	1.84	-.507	-.114	+12.7%
Completed each math assessment	46					61.0%
First-Grade Cohort	113					
Brigance assessment	19	-1.71	1.69	.013	-.934	83.2%
Kindergarten math assessment	55	-2.32	1.96	-.315	-.414	51.3%

1st grade math assessment	94	-2.69	2.32	.198	.072	+16.8%
Completed each math assessment	9					92.0%
Kindergarten Cohort	92					
Brigance assessment	71	-1.92	1.61	-.430	-.944	22.8%
Kindergarten math assessment	54	-2.78	1.97	-.808	1.285	41.3%
Completed each math assessment	44					52.2%

Note. *The third-grade cohort was administered neither the Brigance nor a first-grade assessment. Neither of these scores were considered in the tabulation of the "Completed each assessment" line for this cohort. **The second-grade cohort was not administered a first-grade assessment. These scores were not considered in the tabulation of the "Completed each assessment" line for this cohort. +One student's score was removed because it was an extreme *Z*.

Results

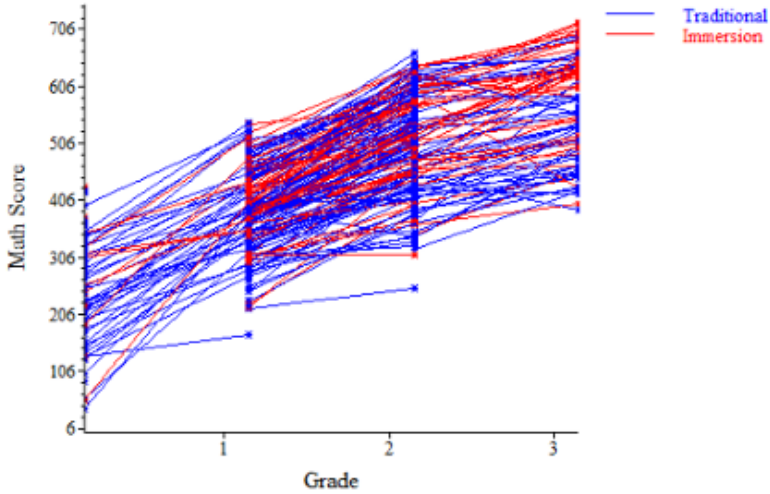
In order to conduct the HLM analysis, the raw scores were utilized in order to examine growth over each year for the student math achievement. This was facilitated because Star Math was the assessment used each year. The researcher decided to not incorporate the MEAP for mathematics assessment data for most analyses because it utilized a completely different scale. Standardized scores would result in the mean of each successive year as zero, indicating no yearly growth in achievement scores.

Mathematics Achievement Models

A null model was created without math scores entered as the outcome variable. No predictors were entered at either level one or level two. The null model was used to calculate the ICC described above, as well as to serve as an initial measure of variance. Test number was entered into the model as a level one predictor. It had been centered previously in SPSS in the creation of the long file. The addition of the test score variable for math assessments at different levels accounted for a large amount of deviance and increased the reliability estimate of the model significantly. Test number was significant, $t = 35.8$, $p < .001$. Visual inspection of the graph of longitudinal math performance indicated a potential small curvilinear trend, with diminishing returns for math achievement with each consecutive year of instruction. The graph of math scores by program type is illustrated in Figure 1.

When the quadratic term was added to the model to account for the curvilinear trend with the test number centered variable as another level one predictor, the decrease in deviance and variance was significant. However, when additional variables were added into level two, the significance of the test number squared term diminished greatly. Although it was a significant addition into the model at level one, $t = -8.36$, $p < 001$, it was then removed from the model for parsimoniousness given its insignificance when level two predictors were added. Lastly, the test number squared variable was entered into the model instead of

Figure 1
Immersion and Traditional Students' Math Scores and Performance Slopes



test number. Although it was a significant level one predictor when accompanying the non-squared predictor, by itself, it reduced the reliability of the model and did not significantly reduce the deviance from the null model. These data are presented in Table 8.

Table 8
Null Model with Predictors at Level One

Model	Parameters	Deviance	Reliability estimate	*T*	*t* with robust standard errors
Null	0	8229.58	.452	--	--
+ Test Number Centered	1	7495.70	.719	38.50***	35.75***
+Test Number Centered + Test Number Squared	2	7428.93	.776	42.56*** -8.36***	38.73*** -7.64***
+ Test Number Squared	1	8222.29	.420	-2.13*	-1.84

Note. method of estimation was restricted maximum likelihood. *$p < .05$, ** $p < .01$, *** $p < .001$

After the inclusion of the test number variable at level one, other predictors were individually entered into the model as fixed effects. Slopes were also allowed to vary in order to assess for random effects. Random slopes did not indicate

model significance with the individual predictor variables, either when level-two predictors were assigned solely to random effects or with both fixed and random effects entered into the model. With the exception of free and/or reduced lunch (FRL) status, random slopes were not significant. As Table 9 indicates, FRL was both significant as a fixed and random effect, indicating an average effect for FRL on student performance and that individual performance slopes varied among students in the FRL population. Program type was a significant level two predictor of math performance, $t = 3.00$, $p < .01$. Student math achievement increased by approximately 23 points on the Star Math assessment for the immersion group relative to students in traditional classrooms. When test number slopes were allowed to vary, the effect of program type was insignificant. Even though immersion was a significant predictor, it did slightly increase the model deviance over the model with test number as the only model predictor.

Even though program type in favor of immersion was a significant predictor at level two, other variables had a stronger effect on math achievement. FRL status was a significant fixed effect, $t = -3.62$, $p < .001$. Students from lower socioeconomic status were expected to score approximately 29 points lower on their math achievement. That estimation was even lower, 32 points, if slope was allowed to vary. Special education status was a stronger predictor than either immersion program or FRL status. A student classified as receiving special education services was a significant predictor, $t = -7.19$, $p < .001$. Students receiving services were expected to score approximately 77 points lower on their math achievement. Lastly, gender was a significant level two predictor as well, $t = 3.69$, $p < .001$. Male students were expected to score 28 points higher on the mathematics assessment than female students.

Table 9 presents interaction effects as individual predictors. Interaction effects for immersion X FRL, immersion X special education, immersion X gender, FRL X special education, FRL X gender, and special education X gender were all performed. Those that reached statistical significance are presented in the chart for main fixed effects.

Table 9

Measure of Level Two Fixed and Random Effects for Math Achievement

Model	Deviance	Reliability estimate	*T*	*T* with robust standard errors	Coefficient
Immersion	7497.08	.711	3.00**	3.05**	22.83
With random test number slope	7488.47	.721	1.26	1.21	3.39

Both fixed and variable test number	7473.87	.713	3.26**	3.22**	25.22
			1.79	1.69	4.89
Free-Reduced Lunch	7466.17	.709	-3.62***	-3.65***	-29.3
With random test number slope	7478.35	.719	-1.67	-1.61	-4.78
Fixed test number and variable test number	7455.52	.707	-3.91***	-3.90***	-31.79
			-2.21*	-2.06*	-6.36
Special education	7430.10	.683	-7.19***	-6.63***	-76.73
With random test number slope	7480.52	.719	-0.17	-0.15	-0.65
Both fixed and variable test number	7422.68	.680	-7.28***	-6.71***	-78.00
			-1.05	-0.98	-3.94
Gender	7465.79	.710	3.69***	3.70***	27.68
With random test number slope	7479.82	.721	1.19	1.14	3.24
Both fixed and variable test number	7456.70	.711	3.96***	3.84***	30.08
			1.84	1.69	5.07
Grade level	7488.53	.719	-0.75	-0.81	-3.24
With random test number slope	7472.23	.747	-4.54***	-4.44***	-6.09
Both fixed and variable test number	7446.13	.767	-4.86***	-4.69***	-26.79
			-6.81***	-6.28***	-11.50

Interaction effects					
Immersion x spec. ed.	7464.94	.720	-2.04*	-1.33	-49.05
Both fixed and variable test number			-1.13	-0.83	-8.34
Immersion x gender	7461.42	.705	4.20***	4.25***	42.59
Both fixed and variable test number	7448.65	.707	4.73***	4.74***	49.55
			2.56*	2.74**	9.37
Free/R lunch x spec. ed.	7436.80	.690	-6.62***	-6.05***	-79.04
Both fixed and variable test number	7428.95	.688	-6.61***	-6.15***	-79.42
			-1.18	-1.20	-4.75
Special ed. x gender	7469.37	.715	-2.90**	-2.88**	-45.37
Both fixed and variable test number	7460.78	.716	-2.62**	-2.62**	-41.71
			1.25	1.38	6.71

Note. method of estimation was restricted maximum likelihood. *$p < .05$, ** $p < .01$, *** $p < .001$

With the identification of significant variables of mathematics achievement identified independently, models were created with multiple predictors entered at level two. The variable time slope was removed from the potential models because it was an insignificant predictor. Likewise, interaction terms were included in analyses, but are not reported because they became insignificant. Both program type and FRL status were significant predictors in the model. Students in the immersion program also demonstrated a significantly higher mathematics achievement when program type was included with special education status and individually with gender. Program type was also significant when FRL and special education status were included in the model. When the interaction term of immersion x FRL was included in the model, program type by itself was no longer a significant level-two predictor. When the interaction term was removed and replaced with gender, all fixed effects (immersion, FRL, special education, and gender) were significant in the model. These effects are reported in Table 10.

Table 10
Multiple Level-Two Fixed Effects Included in the Model for Mathematics Achievement

Model	Deviance	Reliability estimate	*T*	T with robust standard errors	Coefficient
Immersion +	7452.47	.704	2.45*	2.44	18.74
Free/Reduced			-3.19**	-3.14	-25.96
Lunch#					
Immersion +	7417.66	.678	2.20*	2.16*	16.00
Special education#			-6.86***	-6.25***	-73.41
Immersion +	7445.81	.698	5.35***	4.16***	32.83
Gender#			5.47***	4.20***	33.46
Immersion+	7404.74	.671	2.87**	2.10*	17.20
Free/Reduced lunch+			-3.98***	-3.06**	-25.31
Special education			-8.32***	-5.71***	-72.33
Immersion+	7391.89	.669	-0.10	-0.07	-1.00
Free/Reduced lunch+			-4.35***	-3.29**	-38.29
Special education+			-8.41***	-5.79***	-73.02
Immersion x FRL			2.21*	1.63	26.84
Immersion+	7378.65	.653	2.31*	2.25*	16.54
Free/Reduced lunch+			-2.78**	-2.79**	-20.89
Special education+			-6.99***	-6.47***	-72.18
Gender			4.37***	4.29***	30.16

Note. # when the interaction term between the variables was added, it did not reach significance, so it is not reported. The level one variable is test number. Method of estimation was restricted maximum likelihood. *$p < .05$, ** $p < .01$, *** $p < .001$

The equation for the full model is presented below for the four significant predictor variables and the properties of the model are illustrated in Table 11.

Level-1 Model

$$Math\ Score_{ti} = \pi_{0i} + \pi_{1i}{}^{*}Test\ Number_{ti} + e_{ti}$$

Level-2 Model

$$\pi_{0i} = \beta_{00} + \beta_{01}{}^{*}Immersion_{i} + \beta_{02}{}^{*}FRL_{i} + \beta_{03}{}^{*}Special\ Education_{i} + \beta_{04}{}^{*}Gender_{i} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}{}^{*}(Grade\ Level_{i})$$

Mixed Model

$$Math\ Score_{ti} = \beta_{00} + \beta_{01}{}^{*}Imm_{i} + \beta_{02}{}^{*}FRL_{i} + \beta_{03}{}^{*}Sp.Ed._{i} + \beta_{04}{}^{*}Gender_{i} + \beta_{10}{}^{*}Test\ Number_{ti} + \beta_{11}{}^{*}Grade\ Level_{i}{}^{*}Test\ Number_{ti}\ r_{0i} + e_{ti}$$

Table 11
Final Estimation of Fixed Effects and Variance Components for Math Model

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*
For Intercept1, π_0				
Intercept2, β_{00}	450.58	8.45	53.34***	369
Immersion, β_{01}	22.42	7.32	3.06**	369
FRL, β_{02}	-20.82	7.60	-2.74**	369
Special Ed., β_{03}	-69.59	10.44	-6.66***	369
Gender, β_{04}	32.00	6.98	4.59***	369
For Test Number slope, π_1				
Intercept2, β_{10}	66.55	3.07	21.66***	285
Grade level, β_{11}	-6.28	1.30	-4.85	285
Random Effect	Standard Deviation	Variance Component	χ^2	
Intercept1, r_0	55.03	3028.58	1301.03***	369
level-1, e	46.24	2138.08		

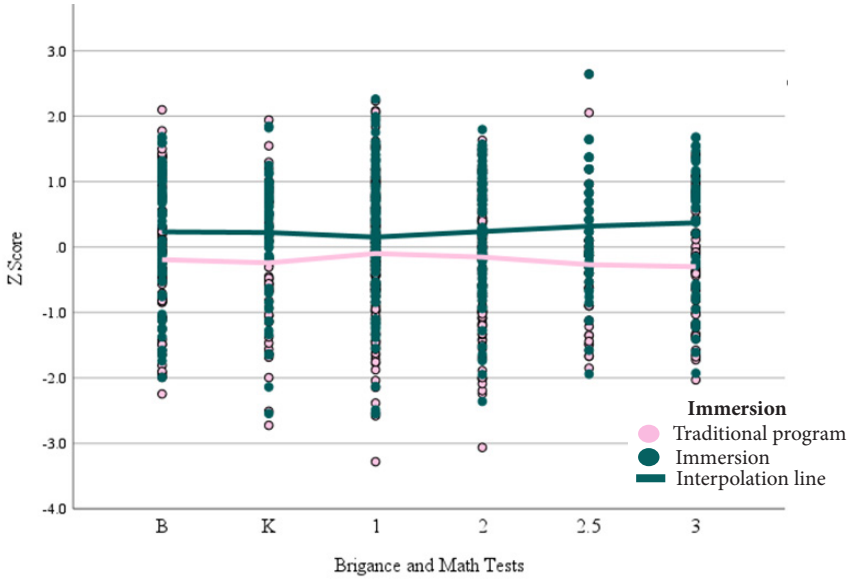
Note. The random level-1 coefficient intercept 1, π_0, reliability estimate is 0.688. The model deviance is 7356.12 with two estimated parameters. *$p < .05$, ** $p < .01$, *** $p < .001$

The Brigance Pre-Kindergarten Assessment as an Important Control for Examining Student Growth and Differences in Math Achievement

Manual inspection of the data revealed that when student scores were separated by program type, the performance on the Brigance screen appeared to be an important determinant of mathematics performance. This is important because students did not have any immersion experience when they were assessed on the Brigance and had not yet begun kindergarten (see Figure 2). For this reason, the Brigance screen was entered into the model as a separate predictor. As Table 12 illustrates, the Brigance screen was a significant predictor at level two, $t = 5.37$, $p < .001$, that greatly reduced model deviance. The Brigance was a significant predictor with both test numbers as fixed and random effects. When both were entered, only test number was significant. Test number was entered in the model as a fixed effect with Brigance, and additional effects were examined. Program type became insignificant, $t = -0.35$, $p > .05$. Although insignificant, it is interesting to note that the performance of the immersion students indicated a negative *t* score with a reduction of four points on the mathematics achievement. The same phenomenon occurred with free and reduced lunch status. FRL lost significance with Brigance screen entered into the model. It is important to note that math performance for students that received a free or reduced lunch was no longer expected to be negative. Similar to the model without the Brigance, both special education status and gender continued to be significant predictors of mathematics performance at level two and in the same direction.

Figure 2

Scatterplot of Standardized Math Performance by Immersion with Interpolation Line



Note. B, K, 1, 2, and 3 indicate the Brigance, kindergarten, first, second, and third grades, respectively. It is important to remember that grade level progress monitoring assessments were administered at the end of the year. The 2.5 indicates performance on the MEAP mathematics achievement, which was assessed in the first semester of the third grade.

Table 12

Inclusion of the Brigance Pre-Kindergarten Screen as a Predictor of Math Achievement

Model	Deviance	Reliability	*T*	*T* with robust standard errors	Coefficient
Brigance	1865.88	.658	5.37***	5.05***	1.33
Brigance with variable slope	1871.38	.636	-5.23***	-4.54***	-0.49
Brigance with both fixed and variable slope	1862.34	.642	2.78** -2.01++	2.54* -1.76	0.90 -0.26
Brigance with additional variables					
Brigance + Immersion	1857.21	.660	5.28*** -0.35	4.83*** -0.34	1.35 -4.04

Brigance +	1857.16	.661	5.30***	4.93***	1.34
Free/reduced lunch			0.24	0.23	2.88
Brigance +	1849.62	.629	5.35***	5.10***	1.29
Special education			-2.62*	-2.58*	-51.16
Brigance +	1844.67	.599	5.76***	5.34***	1.34
Gender			3.72***	3.81***	38.29
Brigance +	1845.54	.680	5.29***	5.01***	1.28
Grade level			-3.60***	-3.65***	-29.23
Brigance +	1807.55	.599	5.66***	5.35***	1.24
Special education +			-3.47***	-3.18**	-61.99
Gender+			3.94***	4.09***	38.13
Grade level			-3.97***	4.09***	-30.94

Note. The level one variable is test number. Method of estimation was restricted maximum likelihood. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, $^{++}p = .05$

Therefore, with the Brigance pre-kindergarten screen entered into the model, the two-level model of student math achievement is represented by the following equations with the three other significant fixed effects of special education, gender, and grade level. Table 13 delineates the model's descriptive characteristics.

Level-1 Model

$$Math\ Score_{ti} = \pi_{0i} + \pi_{1i}{}^{*}Test\ Number_{ti} + e_{ti}$$

Level-2 Model

$$\pi_{0i} = \beta_{00} + \beta_{01}{}^{*}Special\ Education_{i} + \beta_{02}{}^{*}Gender_{i} + \beta_{03}{}^{*}Grade\ Level_{i} + \beta_{04}{}^{*}Brigance_{i} + r_{0i}$$

$$\pi_{1i} = \beta_{10}$$

Mixed Model

$$Math\ Score_{ti} = \beta_{00} + \beta_{01}{}^{*}SpEd_{i} + \beta_{02}{}^{*}Gender_{i} + \beta_{03}{}^{*}Grade\ Level_{i} + \beta_{04}{}^{*}Brigance_{i} + \beta_{10}{}^{*}Test\ Number_{ti} + r_{0i} + e_{ti}$$

Table 13

Final Estimation of Fixed Effects and Variance Components for Math Model with Brigance

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*
For Intercept1, π_0				
Intercept2, β_{00}	427.92	20.08	21.31***	111
Special Ed., β_{01}	-61.99	17.88	-3.47**	111

Gender, β_{02}	38.13	9.67	3.94***	111
Grade Level, β_{03}	-30.94	7.79	-3.97***	
Brigance, β_{04}	1.29	0.22	5.79***	111
For Test Number slope, π_1				
Intercept2, β_{10}	73.83	3.59	20.54***	54
Random Effect	Standard Deviation	Variance Component	χ^2	
Intercept1, r_0	40.21	1616.45	289.05***	111
level-1, e	38.56	1486.97		

Note. The random level-1 coefficient intercept 1, π_0, reliability estimate is 0.599. The model deviance is 1807.55 with two estimated parameters. Grade level was included as a fixed effect instead for the intercept instead of on the test number slope because of improved model reliability and decreased model deviance. Method of estimation was restricted maximum likelihood. *$p < .05$, ** $p < .01$, *** $p < .001$

It is evident that the model with the Brigance test explains more of the variance in student math achievement than the model that did not include the Brigance assessment. Immersion and free-or-reduced lunch status were no longer significant predictors of math performance over the years, although the reliability of the Brigance model is lower. It is also important to note that there was a significant decrease in degrees of freedom in the three-variable model than the four-variable one due to less available testing data. This can account for some of the loss in reliability. In regard to model clarity, the model with Brigance, special education status, gender, and grade level is as parsimonious as the first model with four predictors. Still, there may be variables that play a significant predictive role that are not considered due to the considerable variance that still exists.

Discussion

Findings

This investigation sought to determine if immersion students performed better over time in mathematics relative to their peers from traditional classrooms. Likewise, it was expected that an immersion effect for mathematics achievement was affected by other variables. Indeed, the analyses indicated that students in the immersion program outperformed their peers in traditional classrooms. This provides support for some claims touting that children in immersion programs are able to more quickly excel in mathematics than their peers. In addition, the scatterplot of performance indicates a slight U-shape in the math performance of the immersion students as they enter. Although it is slight, it does appear that there is a small depression in scores relative to peers in traditional classrooms for a period of two years before the immersion students begin to accelerate their growth. Mathematics instruction may be more resilient to the effects of learning through a new language than other content areas like English language arts. The difference in math achievement of the immersion students relative to traditional students appeared muted, however, relative to other variables like lower

socioeconomic status, special education status, or gender. Special education status and gender had much larger effects on student math achievement than enrollment in the Mandarin immersion program. The marker for socioeconomic level, free-and-reduced lunch status, had nearly the same weight on student scores as program type. Despite the statistical significance of the variables in the model, there was still considerable deviance. This indicated that there are other variables that contribute to student performance that were not included.

The second research question focused on the inclusion of pre-kindergarten readiness and its effect on program type and mathematics achievement. The hypothesis that students that performed better on the Brigance screen would demonstrate higher mathematics achievement over the years was supported. However, the pre-kindergarten readiness score also had an unexpected effect. When the Brigance was included, program type was no longer a significant predictor of student performance. In addition, the socioeconomic status of students was no longer significant when controlling for what students were able to do before they entered kindergarten. Regardless of poverty, they experienced similar growth trajectories in mathematics performance relative to wealthier peers. These findings are particularly important and are not often seen in the immersion literature. The model also indicated that despite kindergarten readiness, special education status, gender, and grade level continued to be significant variables in the prediction of student mathematics achievement.

Considering Unequal Groups

Therefore, the pre-kindergarten readiness score allowed the investigation to show that the immersion group differed importantly from students in traditional classrooms. While many studies have examined student performance in immersion education, few studies have examined group differences prior to formal kindergarten education as a factor for student achievement in immersion programs. Within this study, it is apparent that there is a marked group difference both within the elementary school as well as before formal K-12 education. For example, students enrolled in the immersion program were also less likely to receive free or reduced lunch or receive special education services. This is troubling because students living in poverty are also more likely to require special education services (Ryan, 2013; Stansell & McLaughlin, 2013; Sullivan & Bal, 2013). This may explain why there was a strong interaction effect of free-or-reduced lunch status with special education as a separate main effect for math performance. Students that received economic and special education assistance demonstrated marked deficits relative to students that did not receive either form of assistance. Likewise, the financial means of families whose students received meal assistance was unlikely to change over the short time span that the assessment data represent. This is especially apparent given the level of poverty in the surrounding community. Therefore, the difference in rates of students receiving free or reduced lunch and the accompanying rates of special education suggest that there were unequal groups from the beginning of the program, limiting the claims of the benefits of immersion education.

Along with the initial, albeit limited, Brigance pre-kindergarten screenings, it appears that parents of higher means, as evidenced by lower FRL rates, were

more likely to enroll their children in the immersion program. Those children had higher initial academic skills upon entering the formal K-12 education system. This indication favors the conclusion that there existed differences in academic preparation among students prior to enrollment within the immersion program, which has been a criticism of studies of academic achievement of children in immersion programs (Hill, 2018; Steele et al., 2017; Steele et al., 2019; Watzinger-Tharp et al., 2016). These differences, therefore, are of principle concern because conclusions drawn about academic performance related to immersion programming in this population could reflect parents' continued academic investment in their children from early ages and is not an effect of the immersion experience itself. Even the admission into immersion program lotteries requires parents to express their interest in the program to begin with, which can speak to potential inequality among groups of students. In this dataset, the effects of immersion programming disappeared following the inclusion of the Brigance test as a predictor variable. Consequently, the most important factor in determining the long-term success of students in the immersion program relative to students in traditional classrooms was how well the students performed upon entering kindergarten. This pre-kindergarten boost was especially apparent in the tracking of mathematics achievement.

Immersion as Insignificant

The null result of program type is not without consequence. Because immersion education was no longer a significant predictor variable for student success after the inclusion of the Brigance pre-kindergarten screen for mathematics achievement, the oft cited finding that students in immersion programs performed on par with their counterparts in traditional classrooms can be supported. Immersion education does not necessarily provide advantages or disadvantages to student performance by the third grade. Despite spending half their instructional day in a Mandarin classroom, these students experienced no detrimental effects on their mathematics performance. Therefore, immersion programs can be just as effective in ensuring that students progress in their academic skills as programs solely taught in English for English-speaking students. Other variables, such as socioeconomic status, special education status, and gender seemed to be more important explanatory variables for achievement. After accounting for the difference in the pre-kindergarten assessment, free-or-reduced lunch status was no longer a significant predictor of performance for mathematics scores either. This indicates that the elementary school did not increase the disparities between students from different economic backgrounds over the first four years.

It was not surprising that students who received special education services indicated depressed performance on mathematics assessments. However, it was surprising that gender was found to significantly contribute to performance on those assessments. In particular, the mathematics achievement for boys was higher as a single predictor variable, as well as an interaction term with immersion. Another interaction term, gender x special education, was significant as a single predictor variable as well. Likewise, boys outperformed girls even when

accounting for pre-schooling differences, so gender was included in the final model for math achievement. It is unknown whether the boys' experience in the immersion program differed from that of girls. There was also a strong interaction effect between immersion and gender for mathematics achievement prior to the inclusion of the Brigance screen, but immersion disappeared as a significant predictor once the Brigance was included in the model for math. Stereotype threat may potentially play a role in the performance of girls in mathematics at such a young age and may be influenced by gender stereotypes at home and at school (Ganley et al., 2013; Tomasetto et al, 2011).

Overall, the examination of the math dataset presented here tells two different stories. If there is no evaluation of student assessment before entering kindergarten, then immersion is one of multiple factors that is a positive predictor of student achievement. Additionally, math achievement does not indicate a strong, marked initial drop in performance after end-of-year kindergarten progress monitoring assessments. If the data and analyses were taken on face value, then it would appear that immersion education fulfilled the promise of achieving equity with non-immersion students. In addition to the acquisition of another language, the immersion group would then appear to surpass their peers in traditional classrooms. However, once group differences on pre-kindergarten assessments are added into the analyses, the increased benefit of immersion education seems to disappear. Immersion has no significant effect on mathematics achievement and the performance of immersion students is on par with students in traditional programs. This indicates that immersion education is not detrimental to the development of mathematics at the elementary level. The axiom that students in immersion classrooms perform as well as students in traditional classrooms appears to be supported, although immersion education seems to play little role in that conclusion. Rather, it is likely that students that are more academically developed prior to kindergarten with access to continued support in the home are likely to continue to be successful in the formal K-12 educational system as well.

Limitations

Despite the potential implications espoused in this study, there are also many methodological limitations. For example, a primary goal of an HLM analysis was to accommodate missing data on performing the analysis. While HLM can accommodate the missing data and irregular testing periods, it was difficult to model growth over time with substantive missing data for different grade cohorts. Even though progress monitoring was completed with the same STAR Math assessment, it was not possible to include MEAP into the analysis because of different scaling. To examine initial program type performance differences, each of the scores were standardized. While this allowed all available mathematics assessments to be included in the model, it also meant that the standard growth curve became a horizontal line, which limited using test number and grade level as a variable accounting for growth.

Another principle concern with this study is the lack of available testing data from each cohort of students, especially in regard to the pre-Kindergarten

assessment. Very few students had completed the Brigance assessment prior to enrollment in kindergarten and without a complete assessment battery prior to entering the immersion program and continuing over consecutive years, it is difficult to determine if the results are due to the effect of the immersion program or rather to group differences among students as they entered kindergarten and continued parental investment in their children's education. Despite the Brigance screen accounting for most of the reduction in model deviance, the lack of assessment data makes it difficult to determine the extent of those initial group differences.

Similarly, another major limitation of this study relates to group differences and random assignment. Although the census method collects data from each participant, the majority of students in the program will have been self-selected by parents or guardians to participate in the program. This can create a group difference prior to students entering kindergarten and reflect parental investment in education. Other unconsidered variables may exist and intensify the difference in student achievement data over the years. On an individual level, student absences and disciplinary actions are not considered and may play a significant factor in mathematics performance. In addition to individual variables, there are other group-level variables that are absent from the present study. For example, a potential cohort effect that is not accounted for is class size. It is plausible that reduced class sizes in the immersion program for the upper grades helped lead to parity in the test scores. As student numbers in the immersion classrooms declined with advancing grades, the same did not necessarily happen in the traditional classrooms. Class size could be an important explanatory variable and future studies should focus on this issue.

In addition, parents of students in the immersion program may be more likely to seek other academic services such as tutoring in order to counter the expected early achievement losses as the students begin learning through an unfamiliar language. Similarly, immersion students with difficulties in mathematics, for example, may be more or less likely to be designated as needing additional instructional time through an afterschool enrichment program. The many possible academic experiences outside of the typical school day provide opportunities for extraneous variables to confound the results of this investigation.

To continue, some of the variables are problematic. For example, the free or reduced lunch coding is dichotomous. If more accurate data were to be obtained it would be possible to include a category for free, another for reduced, and a last one for full price. While the free or reduced lunch variable is related to family income, families must self-select to receive the benefit. Over the course of the period of years in question, this school was a recipient of a federal grant program that provided free breakfast and lunch to all students. With the school district committed to providing meals, the district free and reduced lunch population dropped. Therefore, the number of students who were classified with regular status may have actually had less family income than would be expected. As a measure that serves as a proxy for familial economic conditions and parental education, this is disconcerting.

To continue, it is also important to consider potential problems associated with the variable of special education status due to its revealed importance in the analyses. For instance, special education encompasses a large variety of services and reflects academic, cognitive, emotional, or physical deficits. Without specification of specific disability categories, the variable is overarching and may lose some of its predictive ability. For example, a reading difficulty may not manifest at all in a student's mathematics achievement, but the student would still have been classified as receiving services. Similarly, students with emotional dysregulation may qualify for special education services but manifest no academic difficulties. Further, special education for a specific learning disability or difficulty with an academic content area may not as of yet been diagnosed for the lower grade levels. Traditionally, an academic deficit is often examined with psychological testing once a student is two years behind normal achievement scores for their grade level. Therefore, students in kindergarten or first grade may potentially have a higher percentage of undiagnosed learning disabilities than students in second or third grade. Students in the immersion group with depressed performance may be falsely attributed to learning mathematics content through Mandarin. The difficulties teaching math content through a developing L2 have been described in the literature (Barwell, 2010); it is plausible it is equally difficult for students learning math through Mandarin. Lastly, students entering kindergarten without any pre-school attendance can further confound and inflate the difference in initial academic achievement as well as potential difficulties in the early diagnosis of potential learning disorders. Within this dataset, it is unknown which students attended any sort of pre-school.

Finally, the conclusions presented within this analysis reflect only one small, rural school in Michigan that had created a Mandarin dual language immersion program over a period of four years. Not only was the program life short, it was not static. This can affect the reliability of assessment scores because some students had left the program, either for a traditional classroom or for a different school district. Other students entered the program in first grade and became part of cohort one. Those changes are not reflected in the data set. Similarly, recruitment of Mandarin teachers was difficult. Within the short life of the program, there were two different kindergarten instructors, two different first grade instructors, one second grade teacher, and two different third grade teachers. This is also not reflected in the dataset, and teacher differences can also play a role in students' academic achievement. Other states with more comprehensive models and programs that better track students and their data have a wealth of information to share regarding the comparison of students in mathematics achievement.

Conclusions

Despite the limitations discussed above, it is hoped that the insights gained from comparisons of students in this school can lead other schools to more carefully track assessment data so that more robust studies can incorporate student readiness for kindergarten in future analyses. Even studies that use propensity matching and examine data from lotteries may not tap into those early differences

among groups. Parental investment in their child's education can be reflected both financially and in time, as well as in the intentions to seek a program that can build social capital beyond traditional school opportunities. Parents may also try to ensure their child does not fall behind and provide tutoring support as well. All of these considerations can lead to a cohort of students in an immersion program that is different from a cohort in a traditional classroom. Although immersion programs have grown in popularity over the last two decades, the assertion that students in these programs perform better may ultimately rest on unequal groups and not reflect language programming at all. Parents who look at the potential of high school students graduating a few classes shy of a minor in a language may be more indicative of their own expectations for their children than the additive bilingualism of a specific language immersion program. The adage that students in immersion programs score on par or better than their peers in traditional programs may solely reflect parental investment in education. Without further examination of factors related to kindergarten readiness and more consistent early assessment data, unaccounted group differences will continue to skew the conclusions regarding the benefits of language immersion on academic achievement.

References

Arthur, G. (2004). Partial Spanish immersion program expands by offering academic excellence for all. *The ACIE Newsletter*, *7*(2). https://carla.umn.edu/immersion/acie/vol7/Feb2004_Partial_Spanish.html

Artzer, M. E. (1990). *The effect of an early partial immersion program on the achievement and academic self-concept of students in grades seven and eight* (9110466). [Doctoral dissertation, Miami University]. ProQuest Dissertations Publishing.

Barwell, R. (2010, May). Tensions in teaching mathematics through a second language. *The ACIE Newsletter*, *13*(3). https://carla.umn.edu/immersion/acie/vol13/no3/may2010_rr.html

Burns, M. (1998). *Interpreting the reliability and validity of the Michigan Educational Assessment Program*. Subcommittee on MEAP of the Standing Committee on Education. https://files.eric.ed.gov/fulltext/ED418138.pdf

Bild, E., & Swain, M. (1989). Minority language students in a French immersion programme: Their French proficiency. *Journal of Multilingual and Multicultural Development*, *10*(3), 255-274. https://doi.org/10.1080/01434632.1989.9994377

Bournot-Trites, M., & Reeder, K. (2001). Interdependence revisited: Mathematics achievement in an intensified French immersion program. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, *58*(1), 27-43. https://doi.org/10.3138/cmlr.58.1.27

Bowman, K., Austin, R., Throndsen, J., & Nielsen, D. (2020). *Dual language immersion program evaluation*. The Utah State Board of Education. https://schools.utah.gov/file/abec99d8-c48b-472b-a34d-683d83a4b68e

Center for Applied Linguistics. (2011). Directory of foreign language immersion programs in U.S schools. Retrieved from http://www.cal.org/resources/immersion/

Chung, M. F. (2020). "I call them my little Chinese kids": Parents' identities and language ideologies in Mandarin-English dual language immersion programs. *Journal of Culture and Values in Education*, *3*(2), 179-195. https://doi.org10.46303/jcve.2020.19/

Cummins, J. (1977). Cognitive factors associated with the attainment of intermediate levels of bilingual skills. *The Modern Language Journal*, *61*(1-2), 3-12. https://doi.org/10.2307/325360

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*, 222-251. https://doi.org/10.2307/1169960

Dorner, L. M., Cervantes-Soon, C. G., Heiman, D. & Palmer, D. "Now it's all upper-class parents who are checking out schools": Gentrification as coloniality in the enactment of two-way bilingual education policies. *Language Policy*, *20*, 323-349. https://doi.org/10.1007/s10993-021-09580-6

Essama, L. (2007). Total Immersion Programs: Assessment Data Demonstrate Achievement in Reading and Math. *The ACIE Newsletter, 11, 1-8*. https://carla.umn.edu/immersion/acie/vol11/BridgeNov07.pdf

Fortune, T., & Tedick, D. J. (2003). What parents want to know about foreign language immersion programs. *ERIC Clearinghouse on Languages and Linguistics*. ED482493. https://files.eric.ed.gov/fulltext/ED482493.pdf

French, B. F. (2013). *Brigance Screens III Technical Manual*. Curriculum Associates. https://oms.brigance.com/Reports/ScreensIII-Tech-Manual-04.pdf

Gándara, P. (2021). The gentrification of two-way dual language programs: A commentary. *Language Policy*, *20*, 525-530. https://doi.org/10.1007/s10993-021-09595-z

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype effects on girls' mathematics performance. *Developmental Psychology*, *49*(10), 1886-1897. https://doi.org/10.1037/a0031412

Genesee, F., & Jared, D. (2008). Literacy development in early French immersion programs. *Canadian Psychology, 49*(2), 140-147. https://doi.org/10.1037/0708-5591.49.2.140

Haj-Broussard, M. (2205). Comparison contexts: African-American students, immersion, and achievement. *The ACIE Newsletter*, *8*(3). https://carla.umn.edu/immersion/acie/vol8/Feb2005_research_africanamerican.html

Higgins, L. (2016, Sept. 2). "Is your school among the worst-performing in Michigan?" *Detroit Free Press*. https://www.freep.com/story/news/education/2016/09/01/priority-list-michigan-failing-schools/89712000/

Hill, S. R. (2018). A meta-analysis of the effect of ten language immersion programs on academic achievement. *Spanish and Portuguese Review*, *4*, 23-50. https://spanportreview.files.wordpress.com/2019/01/5-hill-1.pdf

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed). Routledge.

Jacobsen, S. (2013). *A comprehensive evaluation of a K-5 Chinese language immersion program* (3608844). [Doctoral dissertation, Gardner-Web University]. ProQuest Dissertations Publishing.

Lo, Y. Y., & Lo, E. S. C. (2014). A meta-analysis of the effectiveness of English-medium education in Hong Kong. *Review of Educational Research*, *84*(1), 47-73. https://doi.org/10.3102/00346543134996

MacNamara, J. (1966). *Bilingualism and primary education*. Edinburgh University Press.

Marian, V., Shook, A., & Schroeder, S. R. (2013). Bilingual two-way immersion programs benefit academic achievement. *Bilingual Research Journal*, *36*(2), 167-186. https://doi.org/10.1080/15235882.2013.818075

Mukai, A., Downes, S. M., & Sato, J. (2005). Academic achievement of English-speaking students in Japanese immersion programs. *Educational Technology Research*, 28, 53-58. https://www.researchgate.net/publication/259264385_Academic_achievement_of_English-speaking_students_in_a_Japanese_immersion_program

Oller, D. K., Eilers, R. E., Urbano, R., & Cobo-Lewis, A. B. (1997). Development of precursors to speech in infants exposed to two languages. *Journal of Child Language*, *24*(2), 407-425. https://doi.org/10.1017/S0305000997003097

Padilla, A. M., Fan, L., Xu, X., & Silva, D. (2013). A Mandarin/English two-way immersion program: Language proficiency and academic achievement. *Foreign Language Annals, 46*(4), 661-679. https://doi.org/10.1111/flan.12060

Parkes, J. (2008). Who chooses dual language education for their children and why. *International Journal of Bilingual Education and Bilingualism*, *11*(6), 635-660. https://doi.org/10.1080/13670050802149267

Parks, D. W. (2020). *Effects of partial immersion and one-way immersion models on first and target language acquisition* (27835758). [Doctoral dissertation, Wilmington University (Delaware)]. ProQuest Dissertations Publishing.

Pipia, E. (2013). The history and nature of immersion programs in American context. *Journal of Education*, *2*(2). 65-73. https://jebs.ibsu.edu.ge/jms/index.php/je/article/view/89

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, Jr., R. T., du Toit, M. (2019). *HLM 8: Hierarchical linear and nonlinear modeling.* Scientific Software International. https://www.hearne.software/getattachment/a104646a-d78e-4ccd-88cf-cf04a78b47f2/HLM8-Manual.aspx

Renaissance Learning. (2019). *Star assessments for early literacy: Technical manual.* http://www.floridaearlylearning.com/Content/Uploads/floridaearlylearning.com/files/StarAssessmentsforEarlyLiteracyTechnicalManual.pdf

Renaissance Learning. (2022). *Star assessments for math: Technical manual.* https://help2.renaissance.com/US/PDF/SM/SMRPTechnicalManual.pdf

Ryan, J. E. (2013). Poverty as disability and the future of special education law. *Georgetown Law Journal, 101*(6), 1455-1504. http://ssrn.com/abstract=2152557

Salinger, T., Dumani, S., & Vestal, D. (2019). *Scholastic reading measure: Reliability and validity study*. American Institutes of Research. https://teacher.scholastic.com/education/researchpdf/technical_reports/SCHL_Scholastic_Reading_Measure_Reliability_and_Validity_Study.pdf

Snijders, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality & Quantity*, *30*, 405-426. https://doi.org/10.1007/BF00170145

Stansell, A., & McLaughlin, T. F. (2013). A brief comparison of rural poverty and urban poverty at its consequences for students with special needs. *International Journal of Basics and Applied Science*, *1*(3), 587-593. https://www.researchgate.net/publication/240209351_A_Brief_Comparison_of_Rural_Poverty_and_Urban_Poverty_at_its_Consequences_for_Students_with_Special_Needs

Steele, J. L., Slater, R. O., Zamarro, G., Miller, T., Li Jennifer, Burkhauser, S., & Bacon, M. (2017). Effects of dual-language immersion programs on student achievement: Evidence from lottery data. *American Educational Research Journal*, *54*(Supplement 1), 282S-306S. https://www.jstor.org/stable/44245427

Steele, J. L., Watzinger-Tharp, J; Slater, R., Roberts, G., & Bowman, K. (2019). *Student performance under dual language immersion scale-up in Utah.* Partnership to Study Dual Language Immersion in Utah. https://www.americancouncils.org/sites/default/files/documents/pages/2019-11/Research%20Brief-%20IES%20UTAH%20STUDY%20FINDINGS_0.pdf

Strickland, K., & Hickey, T. M. (2016). Using a national dataset to explore subgroups in Irish immersion education. *Journal of Immersion and Content-Based Language Education*, *4*(1), 3-32. https://doi.org/10.1075/jicb.4.1.01str

Sullivan, A. L., Bal, A. (2013). Disproportionality in special education: Effects of individual and school variables on disability risk. *Exceptional Children*, *79*(4), 475–494. https://doi.org/10.1177/001440291307900406

Swain, M., & Lapkin, S. (1982). *Evaluating Bilingual Education: A Canadian Case Study*. Multilingual Matters.

Thomas, W. P., Collier, V. P., & Abbott, M. (1993). Academic achievement through Japanese, Spanish, or French: The first two years of partial immersion. *The Modern Language Journal, 77*(2), 170-179. https://doi.org/10.2307/328941

Tomasetto, C., Alparone, F. R., & Cadinu, M. (2011). Girls' math performance under stereotype threat: The moderating role of mothers' gender stereotypes. *Developmental Psychology*, *47*(4), 943-949. https://doi.org/10.1037/a0024047

Utah State Board of Education. (2021). 2020-2021 school year. https://schools.utah.gov/file/1c5af035-5431-4231-b4a0-cab79bf7e94a

Valdez, V. E., Freire, J. A., & Delavan, M. G. (2016). The gentrification of dual language education. *Urban Review: Issues and Ideas in Public Education*, *48*(4), 601-627. https://doi.org/:10.1007/s11256-016-0370-0

Wei, L. (2018). Translanguaging as a practical theory of language. *Applied Linguistics*, *39*(1), 9-30. https://doi.org/10.1093/applin/amx039

Watzinger-Tharp, J., Swenson, K., & Mayne, Z. (2016). Academic achievement of students in dual language immersion. *International Journal of Bilingual Education and Bilingualism, 21*(8), 913-928. https://doi.org/10.1080/13670050.2016.1214675

Watzinger-Tharp, J., Tharp, D. S., & Rubio, F. (2021). Sustaining dual language immersion: Partner language outcomes in a statewide program. *The Modern Language Journal, 105*(1), 194-217. https://doi.org/10.1111/modl.12694

Westall, J., Kilbride, T., Utter, A., & Strunk. K. O. (2022). 2022 Preliminary Read by Grade Three Retention Estimates. *Education Policy Innovation Collaborative*. https://epicedpolicy.org/wp-content/uploads/2022/09/RBG3_retention_report_Sept2022.pdf

Williams, C. (2017, December 28). The intrusion of White families into bilingual schools. *The Atlantic.* https://www.theatlantic.com/education/archive/2017/12/the-middle-class-takeover-of-bilingual-schools/549278/

Willig, A. C. (1985). A Meta-analysis of Selected Studies on the Effectiveness of Bilingual Education. *Review of Educational Research, 55*(3), 269-317. https://doi.org/10.3102/00346543055003269

Woltman, H., Feldstain, A. MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(10), 52-69. https://doi.org/10.20982/tqmp.08.1.p052

Woody, A. (2020). "They want the Spanish but they don't want the Mexicans": Whiteness and consumptive contact in an Oregon Spanish immersion school. *Sociology of Race and Ethnicity*, *6*(1), 92-106. https://doi.org/10.1177/2332649218803966

5

Musical Background and French Pronunciation: A Mixed Methods Study

Lisa Brittingham
University of Illinois at Champaign-Urbana

Challenge

What if language teachers could invest in their students' musical backgrounds to help them improve their French pronunciation? Is it more important for language teachers to focus on achieving native-like pronunciation, or is intelligibility of pronunciation more beneficial for building communicative competence? Is there a clearly established relationship between a musical background and French pronunciation?

Abstract

This study focused on two central variables: L2 French pronunciation (intelligibility and nativelikeness) and musical background (musical aptitude and musical training). College level novice learners of French (N = 62) completed three questionnaires and a read-aloud task which were analyzed to investigate if a relationship existed between the two central variables. Additionally, several participants (n = 27) participated in semi-structured interviews as the qualitative element of the study. Multiple regressions were conducted to determine which musical, demographic, and language learning variables were predictors of pronunciation variables. Results indicated no relationship between intelligibility and nativelikeness. Additionally, there was no relationship found between musical aptitude and musical training. However, a positive relationship between the musical variables as well as the language learning variables and intelligibility was discovered. Using K-means cluster analyses, the participants were divided into three different groups: (1) the cluster with the highest intelligibility rate also had the most musical background, (2) the cluster with the highest nativelikeness rate had the lowest intelligibility rate but the highest musical aptitude, and (3) the cluster with the most limited musical background had a moderate intelligibly rate.

Thus, the study was successful in determining the relationship between a musical background and French pronunciation as well as clearly establishing a difference between intelligibility and nativelikeness.

Musical Background and French Pronunciation: A Mixed Methods Study

According to Munro and Derwing (2015), the field of second language acquisition has seen a renewed interest in research on L2 pronunciation in the last decade. More commonly, studies have conceived of pronunciation in terms of strength of foreign accent, which have frequently been assessed through acoustic measures of pronunciation accuracy, using native or near-native production patterns as the acid test for successful learning (Munro & Derwing, 2011). However, in response to the research of Levis (2005), who introduced the intelligibility principle, pronunciation research has taken a new direction. This principle asserts that intelligibility – rather than native-like pronunciation – is most critical for successful communication in second language learning (Munro & Derwing, 2011). Unfortunately, the intelligibility principle has rarely ever been successfully carried both in research and in classroom practices. What is still needed in pedagogy is a conceptualization of intelligibility that can assist teachers to develop a clear understanding of what it is and how it can be achieved (Munro & Derwing, 2011). Further research ought to provide practical information for teaching pronunciation (Mroz, 2020).

In previous research of pronunciation, it has been argued that the processes at stake in language learning and music learning share several similarities at the sound and the structure level (Fedorenko et al., 2009), such as rhythm, stress, and pitch, all of which can be characterized as elements of little intrinsic meaning which, once combined, make structures with a great difference in meaning (Lerdahl & Jackendofe, 1983). Considerable research has been conducted that links musical training with first language acquisition. As early as 1994, Douglas and Willatts showed a link between musical ability and reading ability, and more notably, that training in music skills can lead to an improvement in reading scores. Koelsch et al. (2005) also found a strong overlap of neural resources involved in the processing of syntax in language and music. However, the effects of musical training on second language acquisition have not yet been fully explored (Zeromskaite, 2014). Some studies have been conducted to investigate the effects of musical training and musical aptitude on phonology and pitch perception and production (Kempe et al., 2015; Marques, et al., 2007; Mokari & Werner, 2018). However, little research has been conducted that examines the effects of musical training and aptitude on intelligibility or nativelikeness of second language learners. This study thus focused on novice learners of French in higher education and sought to determine whether there was a relationship between musical background (comprised of both musical training and musical aptitude, each measured differently), and pronunciation (a clear distinction was made between intelligibility and nativelikeness) at different levels of instruction.

Literature Review

Overview of Theories

This study is mainly framed within sociocognitive theory, which asserts that both language and language acquisition are simultaneously occurring and interactively constructed both in the head (cognitively) and in the world (socially). From a social angle of a sociocognitive approach, the primary purpose of learning is a fluent, environmentally adaptive action in the world and not the construction of abstract, decontextualized systems or competences (Churchill et al., 2010). In other words, language and its acquisition are fully integrated into other activities, people, and things. In fact, language use and language acquisition are seen in terms of action and participation within the social world. Intelligibility defined as the extent to which speech is understood verbatim by a listener is a well-suited construct to investigate pronunciation from a social angle.

While language is happening within the social world, Atkinson (2002) further asserts that language use and language acquisition are cognitive phenomena whereby language is stored in, comprehended by, and produced in the brain, and therefore reflects the basic design features of the human brain. Up until recently, it was believed that Broca's area was the part of the brain that was language specific. However, recent results show that it is not exclusively language specific (Magne et al., 2006). Rather, this brain area was activated not only by syntactic processing of linguistic phrases, but also by syntactic processing of musical phrases (Maess et al., 2001). More precisely, there is an overlap of brain structures involved in the processing of music with the same brain structures known to be involved in the processing of language (Koelsch et al., 2005). Thus, within this study, nativelikeness defined as acoustic measures compared to native benchmark scores were examined to investigate the cognitive aspect of pronunciation.

Pitch Perception and Processing

The beneficial link that exists cognitively between language and music development has long been established (Koelsch et al., 2002, 2005), notably concerning musical ability and reading ability (Douglas & Willatts, 1994), as well as syntactic processing and harmonic processing (Besson & Schön, 2001). Schön et al. (2004) examined whether extensive musical training facilitates pitch contour processing not only in music but also in language. The results indicated that similar cognitive computations and neural systems are involved in the integration of pitch processing in both language and music. Regarding pitch more specifically, Magne et al. (2006) concluded that there is a common processing mechanism in both language and music perception. Perrachione et al. (2013) concluded that cognitive mechanisms for pitch processing in language and music are shared beyond simple reliance on overlapping auditory sensory pathways or domain-general working memory and attention. Marques et al. (2007) conducted a study that examined pitch violation in a foreign language. These event-related potential (ERP) latency analyses revealed that musical expertise had a large influence on the time course of pitch processing.

Ong et al. (2017) conducted another study that compared pitch experts (tone language learners and non-tone musicians) and Australian English-speaking non-musicians. The researchers concluded that extensive experience with pitch through either the linguistic or the musical domain did enable the participants to perceive lexical tones more readily. However, they concluded that distributional learning of lexical tones is only facilitated by domain-specific pitch experience. Chang et al. (2016) conducted a study of 55 participants. The results indicated that the musical experience and tone language experience were mutually transferable to facilitate cross-domain tone categorization but not discrimination of fine-grained tonal changes. Chen et al. (2020) also conducted a study to determine whether musical experience affects tone perception efficiency. The researchers concluded that non-tonal musicians showed stronger categorical perception of pitch directions than tonal musicians. Delogu et al. (2006) studied the effects of musical ability on learning tonal languages. The researchers concluded that the subjects performed significantly better in identifying phonological variations rather than tonal ones. In another study conducted by Delogu et al. (2010), two different experiments were conducted which examined the musical transfer effect on linguistic processing both in adults and children. Both children and adults showed poorer performance in tonal discrimination compared to phonological discrimination. They concluded that musical ability did enhance Mandarin Chinese lexical tone discrimination. Thus, each of these studies show a relationship between a musical background and pitch perception and processing.

Pronunciation and Music

For this study, the intelligibility principle coincides with the sociocognitive theory as it focuses on language as communication, which is a social phenomenon. The intelligibility principle is concerned with the production of utterances that are understandable by listeners, thus placing an emphasis on listener comprehension. As part of pronunciation, intelligibility can be clearly defined as the extent to which the verbatim transcript of an L2 learner's speech matches the target text they intended to produce (Munro & Derwing, 2015). According to the intelligibility principle, different features of a language have different effects on understanding, so instruction should focus on those features that are most helpful for understanding and should deemphasize those that are relatively unhelpful (Levis, 2005). The mere fact that a phonological structure poses difficulty for a learner says nothing about whether it is actually worth teaching or whether it can even be taught (Munro & Derwing, 2011). The goal of the intelligibility principle is to teach L2 learners to be easily understood, even if an accent remains in their speech (Mroz, 2018). On the other hand, the nativeness principle encourages the learners to pursue native-like or near native-like pronunciation. Nativeness is indeed a complex issue as it is difficult to determine which native speech within the Francophone world is "standard" French. In order to examine this construct, learners' speech must be acoustically analyzed and compared to a benchmark standard for nativelikeness[1]. For this study, the nativeness principle coincides with the sociocognitive theory

1. This benchmark standard is based on the vowel target frequencies of Paillereau (2016).

as it focused on language and language acquisition as cognitive phenomena. These two constructs (intelligibility and nativelikeness) were further examined to determine whether nativelikeness and intelligibility share much in common or are even related to each other. This concept is very important pedagogically in ascertaining whether certain well-known nativelike contrastive patterns are indeed worth teaching if our goal is intelligibility.

As this study intended to investigate the link between pronunciation (whether intelligibility or nativelikeness) and musical background (musical training and musical aptitude), it is necessary to define and differentiate between musicians and non-musicians. More specifically, it is important to differentiate between musical training and musical aptitude in order to gain a better understanding of which variable, if any, shows a greater relationship to language learning. Previous studies have employed various methods for determining levels of musical training and musical aptitude. Many studies (Chang et al., 2016; Li & DeKeyser, 2017; Ong et al., 2017; Perrachione et al., 2013; Schön et al., 2004, Wong et al., 2007) have relied on self-reporting of the participants or musical background questionnaires to determine musical training. Other studies (Delogu et al., 2010; Kempe et al., 2015; Milovanov et al., 2008; Mokari & Werner, 2018) have utilized various musical aptitude tests. Shabani and Torkeh (2014) employed Gardner's (1993) multiple intelligences test to determine musical aptitude. The participants for the present study took a musical background questionnaire and Gardner's (1993) multiple intelligences questionnaire. In order to focus primarily on music and, specifically, to determine which students have a musical aptitude, the musical-rhythmic questions (Q17-20) were the focus of this questionnaire: (Q17) After hearing a tune once or twice, I am able to sing or whistle it quite accurately; (Q18) When listening to music, I am able to discern instruments or recognize melodies; (Q19) I can easily keep the rhythm when drumming a melody; (Q20) I notice immediately if a melody is out of tune. The purpose of these questionnaires was to establish different levels of formal musical training as well as identify students that had an aptitude in music.

Speech Perception and Production in Musicians

In order to be well informed, it is imperative that previous studies on speech perception of musicians be investigated. Baskent and Gaudrain (2016) conducted a study with 38 participants, 18 musicians and 20 non-musicians, all native speakers of Dutch. The results indicated that musicians showed overall better intelligibility than non-musicians. Connaghan and Patel (2019) conducted a study to examine whether the musicianship advantage for speech perception translates to enhanced processing of degraded speech produced by individuals with dysarthria. Overall transcription scores did not differ between the groups of listeners with and without musical training experience. However, listeners without musical training had higher rates of deletion errors. Additionally, listeners with musical training showed a higher rate of substitution errors. A study conducted by MacCutcheon et al. (2020) investigated the effect of one year of learning to play a musical instrument on speech-in-noise perception and phonological short-term memory

in 5-to-7-year-old children. They concluded that musical training improved neither phonological short-term memory, nor speech-in-noise perception in any of the listening conditions. A similar study by Fleming et al. (2019) examined the effects of short-term musical training on the neural processing of speech-in-noise in older adults. The researchers concluded that six months of musical training appeared to increase the response to speech in bilateral frontal (left Middle Frontal Gyrus and right Medial Frontal Gyrus), left parietal (left Supramarginal Gyrus), and right temporal (Superior/Middle Temporal Gyrus) cortical regions.

It is also necessary to examine previous literature that studies speech production. Li and DeKeyser (2017) investigated perception practice, production, practice, and musical ability in L2 Mandarin tone-word learning. Musical tonal ability was found to correlate with accuracy performance in both tone-word perception and production. Coumel et al. (2019) investigated whether musical expertise and working memory capacities relate to phonological awareness. The correlational analysis showed that pronunciation achievement as measured on an accent faking task correlates with musical expertise but not with working memory capacities. Delogu and Zheng (2020) examined the beneficial effects of musicality on the development of productive phonology skills in second language acquisition. They assessed three different aspects: (1) intelligibility, (2) accuracy, and (3) foreign accent. The researchers concluded that musicality is associated with greater accuracy and more intelligibility in production. The role of musicality on the amount of foreign accent was not clear.

In their study involving 140 Persian speakers learning English, Shabani and Torkeh (2014) attempted to find the relationship between musical intelligence and foreign language proficiency. A Pearson correlation was used, and results found a correlation coefficient index ($r = 0.82$) which indicated a statistically significant correlation between musical intelligence and foreign language proficiency. In a study conducted by Mokari and Werner (2018), 40 Azerbaijani learners of Standard Southern British English participated in a pre-test/training/post-test setting. Results indicated that the experimental group's productions were significantly improved after training. However, no significant correlation between overall musical ability scores and L2 vowel production was found. Results revealed, however, a correlation between discrimination improvements and tonal memory. Thus, previous research has established a correlation between a musical background and certain aspects of second language acquisition.

Multilingualism

It is possible that a musical background combined with other covariables such as multilingualism could play a role in successful pronunciation. According to Roehr-Brackin et al. (2021), current research indicates that multilinguals differ from monolinguals due to the M-factor. Early studies (Papagano & Vallar, 1995) compared adult multilinguals to bilingual speakers of Italian and only one other language on two learning tasks. The results indicated significantly higher recall rates for multilinguals on the word-learning task, but not on the paired associate learning task. Van Hell and Candia Mahn (1997) compared results of multilinguals to monolinguals in learning vocabulary and observed a definite learning advantage for the more experienced

language learners. More recent studies include a study by Keshavarz and Astaneh (2004), which examined the effects of multilingualism on varying literacy levels. They determined that both bilingual groups outperformed the monolingual group and the biliterate group outperformed the monoliterate group. In two different studies, Kaushanskaya and Marian (2009a, 2009b) demonstrated that multilinguals outperformed monolinguals in novel word learning. Additionally, Kaushanskaya and Rechtzigel (2012) compared adult English-Spanish bilinguals to English monolinguals. The results indicated a multilingual advantage when novel words were associated with concrete concepts but not when novel words were associated with abstract concepts. Thus, it is clear from these studies that multilingualism can be a factor in successful pronunciation.

Learners' Beliefs

The moment that L2 learners step into the classroom, they also bring all their personality traits along as well as their beliefs, attitudes, and language styles (Sharajabian & Hashemian, 2015). In an exploratory study of 62 French university students conducted by Gabillon (2007), the researcher concluded that the learners' beliefs concerning course expectations, course content, and goals for studying English were all factors that influenced their attitudes towards learning as well as their levels of motivations and motivational orientations. In a study conducted by Sato and Storch (2022), the researchers concluded that the learners' motivation was largely substantiated by their immediate learning goals. Sharajabian and Hashemian (2015) further assert that attitude is considered one of the most important components of language learning. They divide attitude into three different aspects: (1) behavioral aspect, (2) cognitive aspect, and (3) emotional aspect. In their study with Iranian EFL learners, they concluded that learners' attitudes, the type of task at hand, and cultural background and previous experiences all contributed to their ability to learn. According to Derwing and Munro (2015), pedagogical practices in the last twenty years have clearly moved towards an emphasis on intelligibility in the classroom. However, learners' beliefs on the importance of intelligibility and nativelikeness are definitely important factors to consider. Thus, according to these studies, learners' beliefs are another factor that contributes to successful pronunciation.

Research Questions

Informed by the previous review of literature, the present study addressed the following research questions:

(RQ1) Is there a relationship between (a) intelligibility and nativelikeness and is there a relationship between (b) musical aptitude and musical training in college level novice learners of French?

(RQ2) To what extent do musical variables (musical aptitude, musical training), demographic variables (age, gender), and language learning variables (level of French instruction, status of French as a foreign language) contribute to predicting (a) intelligibility and (b) nativelikeness?

(RQ3) How do learners with high vs. low intelligibility, nativelikeness, musical aptitude, and musical training believe their musical background influences their acquisition of French pronunciation?

Methods

This study is based on an explanatory sequential mixed method design. The qualitative phase of the study (the semi-structured interviews) was used to help explain the initial quantitative results.

Participants

Sixty-two (N = 62) voluntary participants took part in the study. They came from American universities that had a four-semester basic language program in French and were recruited in 2021. The majority of the participants were enrolled in the basic language program at a midwestern university which is offered to meet the four-semester language requirement for various degree programs. In an effort to compare four-semester basic language programs across universities in the United States, 104 different universities were contacted across the country. Unfortunately, the response was minimal. There were six (n = 6) students from a southern university and one (n = 1) student from a western university among the sixty-two (N = 62) voluntary participants. The courses at the midwestern university, which comprised the majority of the participants, met four times a week for 50 minutes. According to the syllabi, they aim to achieve Novice High proficiency after semesters one and two and Intermediate Mid proficiency after semesters three and four in oral expression, listening comprehension, reading, writing, and cultural understanding (ACTFL, 2012). Although pronunciation was integrated within the instruction, it was not the primary focus of these courses.

Participants varied in gender. Three (n = 3) participants identified as gender non-binary, thirty-five (n = 35) identified as female, and twenty-four (n = 24) identified as male. Participants had a mean age of 26 years (range: 18-77). The majority of the participants identified English as their first language (n = 48). Other first languages were listed as Chinese (n = 4), Spanish (n = 4), Russian (n = 2), Greek (n = 1), Polish (n = 1), Hindi (n = 1), and Urdu (n = 1). The onset age for learning French had a mean of 18 years (range: 1-66). There were fifteen (n = 15) participants who indicated that they were multilingual prior to studying French.

Research Instruments and Procedure

Participants received a link to complete three online questionnaires and one audio recording. The first questionnaire was a language background questionnaire collecting demographic information as well as language use during childhood and language use in education.

The second questionnaire was a musical background questionnaire, designed specifically for this study. This questionnaire collected information regarding the amount of formal musical training that participants had received. It also aimed at comparing the amount of training received across various musical instruments, including voice. As previously noted in the literature review, many studies (Chang et al., 2016; Li & DeKeyser, 2017; Ong et al., 2017; Perrachione et al., 2013; Schön et al., 2004, Wong et al., 2007) have relied on self-reporting of the participants or musical background questionnaires to determine musical training.

The third questionnaire was Gardner's (1993) Multiple Intelligence Questionnaire, similar to what was done by Shabani and Torkeh (2014). This

questionnaire distinguishes between eight areas of intelligence, including verbal-linguistic, visual-spatial, mathematical-logical, body-kinesthetic, interpersonal, intrapersonal, naturalistic, and musical-rhythmic and was administered to focus on the musical-rhythmic section.

Next, the students underwent an oral task on *Phonic.* The read-aloud task was an excerpt entitled "*La bise et le soleil*," which comes from Aesop's *Fables*. This excerpt was chosen because it is widely used by phoneticians. The oral task was submitted as an audio recording.

In addition to the questionnaires and recordings, students were also given the option to do a semi-structured interview in English with the researcher (Appendix A). After the survey was completed, those participants that indicated a willingness to participate received a link to schedule a Zoom interview. Twenty-seven ($n = 27$) of the participants completed the Zoom interview, which was audio-recorded, to discuss their processes of learning.

Analyses

First, data were retrieved from the language background questionnaire. Since there were three participants that identified as gender non-binary, variables were created for each of these participants using both female and male genders. Moreover, the questionnaire allowed to establish each participant's level of French instruction, whether English was their L1, onset age of learning French, and total years of French studied. Second, the musical intelligence questionnaire was based on a 4-point Likert scale, with 1 used for "Does not describe me at all" and 4 used for "Describes me extremely well." Musical aptitude was determined by adding the total of the four scores in the multiple intelligence questionnaire that pertained to music. Third, musical training was established by computing total years of instrumental training, total years of instrumental and voice training, and number of instruments played.

Next, to determine each participant's intelligibility rate, data were handled following Mroz's (2020) protocol. Three raters were used to determine participants' intelligibility rate: two near-native speakers and one native speaker of French. As the read-aloud task was identical for all participants, asking raters to perform verbatim transcriptions would have been inappropriate as the raters would have become increasingly familiar with the text. Instead, the raters were given a copy of the text read by the participants and instructed to only play the recording one time. They were asked to circle the parts where there was a discrepancy between the written script and the pronunciation that they heard. The number of unintelligible words circled by the raters served to establish a match rate at the word-level for each recording in order to determine participants' intelligibility rate.

The Interclass Correlation Coefficients (ICC) were computed (Table 1). When analysis for Intraclass Correlation Coefficient was run in *SPSS* using absolute agreement with all three raters, the result was 67.6%. When ICC analysis was run using consistency for all 3 raters, the result was 80.1%, which shows that the raters were consistent in their rating. Even though Rater 3's ratings were consistent with the other raters, the decision was made to exclude Rater 3, who was one

of the near-native speakers of French, as including Rater 3 greatly decreased the interrater reliability percentage. Finally, the ratings of Raters 1 and 2 were averaged to establish the intelligibility rates for each participant.

Table 1
Intraclass Correlation Coefficient

Raters	Absolute agreement	Consistency agreement
Raters 1, 2, 3	.676	.801
Raters 1 and 2	.907	.913
Raters 1 and 3	.523	.675
Raters 2 and 3	.511	.684

In order to analyze nativelikeness of the read-aloud task, acoustic measures were taken for all occurrences of the contrastive vowels of [e] and [ə] (schwa). These particular vowels were chosen because they were the most frequent pair of contrastive vowels in the read-aloud task. (Other phonemic variables could be used as venues for future research). This transcription of vowels was based on Fougeron and Smith (1993), resulting in 23 occurrences of [ə] and 13 occurrences of [e]. After acoustic measurements were taken in *Praat* for each of the participants in the F1 and F2 formants, these measurements were then converted to a Bark scale using the formula used by Nguyen and Fagyal (2008). The benchmark scores used to calculate the frequency differences for each participant were taken from Paillereau (2016). These numbers were also converted using the Bark scale. Finally, frequency differences between the benchmark scores and the scores of the participants were calculated for each of the four phonetic cases ([ə] for F1 and F2 formants, and [e] for F1 and F2 formants).

Lastly, the semi-structured interviews were transcribed and then coded by the researcher based on the following four emerging influences as reported by the participants: influence of listening to music and/or musical training, influence of other audio resources, influence of speakers of French, and influence of multilingualism.

In order to address RQ1a regarding the relationship between intelligibility and nativelikeness in college level novice learners of French, and RQ1b, which examined the relationship between musical aptitude and musical training in college level novice learners of French, Spearman's rank-order correlations were conducted in *SPSS*. For RQ2, which examined the extent to which musical variables (musical aptitude, musical training), demographic variables (age, gender), and language learning variables (level of French instruction, status of French as a foreign language) contributed to predict (a) intelligibility and (b) nativelikeness in college level novice learners of French, a series of multiple linear regressions were conducted in *SPSS*. Lastly, in order to address RQ3, which examined how college level novice learners of French with high vs. low intelligibility, nativelikeness, musical aptitude, and musical training believed their musical background

influenced their acquisition of French pronunciation, a standardization of all scores into z-scores was conducted in order to run K-means cluster analyses. This was followed by Analyses of Variance (ANOVAs) to determine possible groupings, as well as which variables, if any, could significantly distinguish participants across clusters. Within these cluster groupings, 26 of the 27 interviews were included. The qualitative data collected from the interviews were triangulated with these groupings to answer RQ3.

Results

RQ1 – Intelligibility / Nativelikeness and Musical Aptitude / Musical Training

The Spearman's rank-order correlations yielded no statistically significant correlation between intelligibility and nativelikeness, $r_s(59) = -.104$, $p = .424$ or between musical aptitude and musical training, $r_s(60) = .113$, $p = .383$.

RQ2a – Predictors of Intelligibility

The multiple regression run to predict intelligibility from musical aptitude, musical training, age, gender, level of French language learning, and status of French as a foreign language (Table 2) significantly predicted intelligibility, $F(6,51) = 6.085$, $p < .001$, adj. $R^2 = .349$. Musical aptitude ($p = .008$), musical training ($p = .023$), level of French language learning ($p < .001$), and status of French as a foreign language ($p = .018$) statistically significantly contributed to the prediction. Age ($p = .108$) and gender ($p = .669$) did not.

Table 2
Predictors of Intelligibility

Intelligibility	***B***	**95% CI for *B***		***SE B***	**ß**	R^2	ΔR^2
		LL	***UL***				
Model						.42	.35***
Constant	77.294***	67.40	87.19	4.93			
Gender	.554	-2.04	3.15	1.29	.05		
Age	.193	-.04	.43	.12	.22		
Status of French	3.577*	.63	6.52	1.47	.28*		
Language level	2.454***	1.25	3.66	.60	.47***		
Musical aptitude	-.290**	-.50	-.08	.11	-.38**		
Musical training	.205*	.03	.38	.09	.27*		

*$p < .05$. ** $p < .01$. *** $p < .001$

RQ2b – Predictors of Nativelikeness

The multiple regression did not statistically significantly predict nativelikeness of schwa F1 formant, $F(6,51) = 1.688$, $p = .143$, adj. $R^2 = .067$.

The multiple regression (Table 3) did statistically significantly predict nativelikeness of schwa F2 formant, $F(6,51) = 2.953$, $p = .015$, adj. $R^2 = .171$. Gender ($p < .001$) statistically significantly contributed to the prediction. Musical aptitude ($p = .508$), musical training ($p = .816$), level of French language learning ($p = .871$), status of French as a foreign language ($p = .331$), and age ($p = .490$) did not.

Table 3
Predictors of Nativelikeness of Schwa F2 Formant

Nativelikeness of schwa F2 formant	***B***	**95% CI for *B***		***SE B***	**ß**	**R^2**	**ΔR^2**
		LL	***UL***				
Model						.26	.17*
Constant	.319*	-1.20	1.84	.76			
Gender	.769***	.37	1.17	.20	.50***		
Age	-.013	-.05	.02	.02	-.11		
Status of French	-.221	-.67	.23	.23	-.13		
Language level	.015	-.17	.20	.09	.02		
Musical aptitude	.011	-.02	.04	.02	.10		
Musical training	-.003	-.03	.02	.01	-.03		

*$p < .05$. ** $p < .01$. *** $p < .001$

The multiple regression (Table 4) did statistically significantly predict nativelikeness of [e] F1 formant, $F(6,51) = 2.980$, $p = .014$, adj. $R^2 = .172$. Gender ($p < .001$) statistically significantly contributed to the prediction. Musical aptitude ($p = .575$), musical training ($p = .547$), level of French language learning ($p = .845$), status of French as a foreign language ($p = .918$), and age ($p = .841$) did not.

Table 4
Predictors of Nativelikeness of [e] F1 Formant

Nativelikeness of [e] F1 formant	***B***	**95% CI for *B***		***SE B***	**ß**	**R^2**	**ΔR^2**
		LL	***UL***				
Model						.26	.17*
Constant	-.123*	-1.18	.94	.53			
Gender	.547***	.27	.82	.14	.51***		
Age	-.003	-.03	.02	.01	-.03		
Status of French	-.016	-.33	.30	.16	-.01		
Language level	.013	-.12	.14	.06	.03		
Musical aptitude	.006	-.02	.03	.01	.09		
Musical training	.006	-.01	.03	.01	.08		

*$p < .05$. ** $p < .01$. *** $p < .001$

The last multiple regression (Table 5) did statistically significantly predict nativelikeness of [e] F2 formant, $F(6,51) = 4.642$, $p < .001$, adj. $R^2 = .277$. Gender ($p = .001$), age ($p = .027$), and musical aptitude ($p = .026$) statistically significantly contributed to the prediction. Musical training ($p = .396$), level of French language learning ($p = .321$), and status of French as a foreign language ($p = .342$) did not.

Table 5
Predictors of nativelikeness of [e] F2 formant

Nativelikeness of [e] F2 formant	*B*	95% CI for *B*		*SE B*	ß	R^2	ΔR^2
		LL	*UL*				
Model						**.35**	**.28*** **
Constant	**2.677*** **	**1.25**	**4.10**	**.71**			
Gender	**-.633****	**-1.01**	**-.26**	**.19**	**-.41****		
Age	**-.039***	**-.07**	**-.01**	**.02**	**-.33***		
Status of French	**.202**	**-.22**	**.63**	**.21**	**.12**		
Language level	**-.086**	**-.26**	**.09**	**.09**	**-.12**		
Musical aptitude	**.035***	**.00**	**.07**	**.02**	**.33***		
Musical training	**-.011**	**-.04**	**.01**	**.01**	**-.10**		

*$p < .05$. ** $p < .01$. *** $p < .001$

RQ3 – Influences of Musical Background

After running various K-means clusters analyses, it was concluded that the best model was three clusters (Table 6, next page). There were no significant difference

Table 6
K-Means Clusters

Clusters	***C1***		***C2***		***C3***	
Size	*n* = 15		*n* = 25		*n* = 18	
	M	*SD*	*M*	*SD*	*M*	*SD*
Musical Aptitude (*p* = .025)	20.13	10.26	14.28	6.127	14.28	3.953
Total Instruments (*p* < .001)	1.47	1.125	1.36	.810	2.78	.878
Musical Training (*p* < .001)	4.37	6.119	2.40	2.52	14.11	7.227
Intelligibility (*p* = .027)	89.97	7.945	92.80	4.81	95.14	2.67
Schwa F1 Formant (*p* = .006)	.19	.391	.52	.413	.68	.446
Schwa F2 Formant (*p* < .001)	.36	.528	1.35	.718	.94	.664
[e] F1 Formant (*p* < .001)	.40	.267	.87	.453	1.13	.581
[e] F2 Formant (*p* < .001)	2.33	.545	1.31	.620	1.28	.686
Denomination	High nativelikeness		Limited musical background		Highest intelligibility rate	

between the clusters in terms of multilingualism[2]. Descriptive statistics indicated that Cluster 3 (n = 18) was characterized by the highest intelligibility rate as well as the most years of musical training. Cluster 1 (n = 15) was characterized by the highest nativelikeness and the highest musical aptitude. Cluster 2 (n = 25) was characterized by a limited musical background but a moderate intelligibility rate. It is also important to mention that seven (n = 7) participants were excluded from the K-means clusters because they did not readily fit into any of the groups.

Using the groupings created by the K-means cluster analysis in *SPSS*, the interviews were divided into the various groupings. There were a total of fifteen (n = 15) participants in Cluster 1, which included seven (n = 7) participants who completed the interview. There were twenty-five (n = 25) participants in Cluster 2, which included nine (n = 9) participants who completed the interview, and eighteen (n = 18) participants in Cluster 3, of which ten (n = 10) participants completed the interview. There were four distinct categories of emerging influences noted on the participants' learning in the interviews: (1) influence of listening to music and/or musical training, (2) influence of other audio resources, (3) influence of speakers of French, and (4) influence of multilingualism.

Cluster 1 (C1), which was characterized as the cluster with the highest nativelikeness and the highest musical aptitude, mentioned that their learning was influenced by "the in-person feedback from [their] instructors and teachers" (participant #1021, echoed by #1032, 1037, and 1055). Compared to Clusters 2 (C2) and 3 (C3), which will be described below, C1 participants only mentioned the influence of their teachers and did not mention other native French speakers. Musical influence was almost totally absent, except for one participant (#1008) who stated that they would listen to "a good amount of French music." There was also minimal influence expressed from multilingualism. Only one participant (#1038) mentioned the benefit of "knowing another language." Finally, other audio resources such as YouTube, CDs, or the online workbook were mentioned as an important influence in learning (#1008, 1032, 1037, and 1055). When asked to choose the most important influence, only participant #1055 chose other audio resources as more beneficial than French speakers. Thus, C1 can be characterized as the cluster most greatly influenced by French speakers, especially their instructors.

Cluster 2 (C2), which was characterized as the cluster with a limited musical background and a moderate rate of intelligibility, had the most diversity in responses. Several participants mentioned the influence of French speakers (#1031, #1035, #1042, #1053), including sources outside of their classes. For example, participant #1031 mentioned that they have heard their family "speak [French] since [they were] young." Two participants (#1031, #10071/10072) mentioned the influence of multilingualism in their learning. Several participants also said, "sometimes I'll listen to French music" (#1035, echoed by #1052, #10051/10052, #10071/10072). This is an interesting factor as their musical training is very limited. However, when asked what the most important influence in their learning was, the overwhelming response was "listening to recordings" (#1035, echoed by #1042, #1053, #10461/10462). Thus, C2 can be characterized as the cluster most greatly influenced by other audio resources.

2. A Kruskal-Wallis H Test was run to determine significance for multilingualism, p = .875.

Cluster 3 (C3), which was characterized as the cluster with the highest intelligibility rate and the most musical training, also mentioned French speakers as an important influence. However, unlike C1, which only mentioned the influence of their instructors, C3 participants mentioned that "hearing a native speaker just speak French" (#1014, echoed by #1001, #1036, and #1051) was an important influence in addition to their instructors. Several participants from C3 also mentioned the importance of "listening to [French] music" (#1012, echoed by #1001, #1033, #1045, #1051) as influential in their language learning. One participant mentioned their formal training in their undergraduate degree in French diction and commented on the benefit of that to their language learning: "I've been singing French for so long now that the language is just very familiar, and it doesn't take much for me to just go and start speaking it" (#1051). Two participants also mentioned that "knowing another language helps too" (#1013, echoed by #1001). Finally, five of the ten participants mentioned that other audio sources like Duolingo, *Portails*, and French videos and television shows were helpful in their French learning (#1001, #1002, #1012, #1014, #1036, #1045). When asked what the most important influence was in their learning, the overwhelming answer was the influence of French speakers (#1001, #1003, #1012, #1013, #1014, #1033, #1036). However, unlike C1, which only mentioned their instructors as the French speakers, several participants specifically mentioned native French speakers. The most interesting finding in this cluster was the response to the question, "What, if any role, has music had as an impact on your French pronunciation?" Several participants readily agreed that music had an impact on their learning, and several others were not sure, but the majority of the participants (#1001, #1002, #1013, #1036, #1003, #1014, #1033, #1051) mentioned that they did have a background either in playing a musical instrument or singing. Thus, C3 can be characterized as the cluster most influenced by native French speakers and a musical background.

Discussion

Determining Learning Goals

RQ1a sought to determine if there was a relationship between intelligibility and nativelikeness in college level novice learners of French. There was no significant relationship found. The (in)ability to correctly pronounce the [e] and the [ə] did not affect the intelligibility rate of the learners. This has implications in terms of pedagogy. First of all, the learning goals must be determined. As instructors, is it more important to strive for native-like pronunciation or is intelligibility a more realistic goal? If the goal is intelligibility, is it necessary to spend significant classroom time attempting to differentiate between these two phonemes? For example, a Spanish speaker of French oftentimes mispronounces the [ə] as an [e] and yet is fully understandable. According to Munro and Derwing (2011), if intelligibility is the goal, a phonological structure that poses difficulty for a learner is not necessarily a valid reason for teaching it. This further confirms previous research about the fact that scapegoating of accent should not be happening (Derwing, 2010).

Benefits of Musical Training

A positive relationship between musical variables and intelligibility was confirmed by the results. This has important implications for pedagogy. The benefits of a musical background in learning a second language are clearly established by this study. The majority of the participants of C3, which was the cluster with the highest intelligibility rate, mentioned that they did have a background either in playing a musical instrument or singing. Language teachers can gain insight in pedagogical expectations by learning the musical background of their students. They can also inform their students that their background can be advantageous in learning the language. These results can also lend support to both music programs and language programs within schools in demonstrating that they can be mutually beneficial to the students.

Level of Language Learning

The greater the level of language learning was, the higher the rate of intelligibility was. This seems like a logical conclusion, but again offers implications for pedagogy. As language teachers, it is encouraging to have documented data that indicates that pronunciation improves as the language level advances. Even when pronunciation is not the primary focus of language classes, as were the classes in this study, the rate of intelligibility increased. This fact provides confirmation that continued language instruction will also benefit pronunciation.

Multilingualism

Finally, the more languages learners had learned prior to studying French, the higher their intelligibility rate was. As previously mentioned, this concept is known as the M-factor. Previous research has proven the advantages of multilingualism, and this study confirms this. These results are very beneficial to language programs within schools because they demonstrate the advantages of language learning.

Conclusion

There are several limitations of this study. For example, although recruitment to other universities did take place, the overwhelming majority of the participants were from one midwestern university. Thus, this study is effectively limited to the learners of this university and cannot be generalized across universities with similar programs in the United States. Furthermore, the inter-rater reliability between the three raters was such that one of the raters was excluded from the analysis in order to increase the reliability of the study. Additionally, the results indicated that the reliability of the musical intelligence instrument to determine musical aptitude was also limited. Future venues of research include, but are not limited to, further pursuit of the study of whether musical training or aptitude is a better predictor of intelligibility, the relationship between musical aptitude and nativelikeness, as well as other influences on intelligibility and nativelikeness, such as a theatrical background.

The main purpose of this study was to determine if there was a relationship between a musical background and French pronunciation for adult novice

learners. Previous literature had clearly established this relationship between a musical background and speech perception, but there was minimal research establishing a relationship between a musical background and speech production. Furthermore, there has also been discussion differentiating the difference between intelligibility and nativelikeness. This study was successful in determining the relationship between a musical background and French pronunciation as well as clearly establishing a difference between intelligibility and nativelikeness. For pedagogical purposes, determining the musical backgrounds of students could be helpful in determining whether they have a higher potential for intelligibility. It could also be helpful to make students aware of additional assets that they bring to the language learning experience.

References

ACTFL. (2012). "ACTFL Proficiency Guidelines 2012." ACTFL Central. https://www.actfl.org/resources/actfl-proficiency-guidelines-2012

Atkinson, D. (2002). Toward a sociocognitive approach to second language acquisition. *The Modern Language Journal, 86*(4), 525-545. https://doi.org/10.1111/1540-4781.00159

Baskent, D., & Gaudrain, E. (2016). Musician advantage for speech-on-speech perception. *Journal of the Acoustical Society of America, 139*(3), EL51-EL56. https://doi.org/10.1121/1.4942628

Besson, M., & Schön, D. (2001). Comparison between language and music. *Annals of the New York Academy of Sciences, 930*(1), 232-259. https://doi.org/10.1111/j.1749-6632.2001.tb05736.x

Chang, D., Hedberg, N., & Wang, Y. (2016). Effects of musical and linguistic experience on categorization of lexical and melodic tones. *Journal of the Acoustical Society of America, 139*(5), 2432-2447. https://doi.org/10.1121/1.4947497

Chen, S., Yang, Y., Zhu, Y., & Wayland, R. (2020). How musical experience affects tone perception efficiency by musicians of tonal and non-tonal speakers? *PloS one, 15*(5), 1-2. https://doi.org/10.1371/journal.pone.0232514

Churchill, E., Okada, H., Nishino, T., & Atkinson, D. (2010). Symbiotic gesture and the sociocognitive visibility of grammar in second language acquisition. *The Modern Language Journal, 94*(2), 234-253. https://doi.org/10.1111/j.1540-4781.2010.01019.x

Connaghan, K.P., Fisk, D., & Patel, R. (2019). The impact of musical training on understanding dysarthric speech: A preliminary study of transcription errors. *Communication Disorders Quarterly,* 1-8. https://doi.org/10.1177/1525740119886541

Coumel, M., Christiner, M., & Reiterer, S.M. (2019). Second language accent faking ability depends on musical abilities, not on working memory. *Frontiers in Psychology, 10*(257), 1-10. https://doi.org/10.3389/fpsyg.2019.00257

Delogu, F., Belardinelli, M.O., & Lampis, G. (2006). Music-to-language transfer effect: May melodic ability improve learning of tonal languages by native nontonal speakers? *Cognitive Processing, 7*(3), 203-207. https://doi.org/10.1007/s10339-006-0146-7

Delogu, F., Belardinelli, M.O., & Lampis, G. (2010). From melody to lexical tone: Musical ability enhances specific aspects of foreign language perception. *European Journal of Cognitive Psychology, 22*(1), 46-61. https://doi.org/10.10180/09541440802708136

Delogu, F., & Zheng, Y. (2020). Beneficial effects of musicality on the development of productive phonology skills in second language acquisition. *Frontiers in Neuroscience, 14*(618), 1-8. https://doi.org/10.3389/fnins.2020.00618

Derwing, T.M., & Munro, M.J (2015). Pronunciation fundamentals. Evidence-based perspectives for L2 teaching and research. *Language Learning and Language Teaching, 42*(1), 1-222. https://doi.org/10.1075/lllt.42

Douglas, S., & Willatts, P. (1994). The relationship between musical ability and literacy skills. *Journal of Research in Reading, 17*(2), 99–107. https://doi-org.proxy2.library.illinois.edu/10.1111/j.1467-9817.1994.tb00057.x

Fedorenko, E., Gibson, E., Patel, A., Casasanto, D., & Winawer, J. (2009). Structural integration in language and music: Evidence for a shared system. *Memory & Cognition*, *37*(1), 1-9. https://doi:10.3758/MC.37.1.1

Fleming, D., Zendel, B.R., Peretz, I., Belleville, S., West, G. (2019). The effects of short-term musical training on the neural processing of speech-in-noise in older adults. *Brain and Cognition, 136,* 1-12. https://doi.org/10.1016/j.bandc.2019.103592

Fougeron, C., & Smith, C. (1993). Illustrations of the IPA: French. *Journal of the InternationalPhonetic Association, 23*(2), 73-76. https://doi:10.1017/S0025100300004874

Gabillon, Z. (2007). Learner beliefs on L2 attitudes and motivation: An exploratory study. *Lingua et Linguistica, 1*(1), 68-90. https://www.researchgate.net/publication/259803418_Learner_Beliefs_on_L2_Attitudes_and_Motivation_An_Exploratory_Study

Gardner, H. (1993). *Frames of Mind: The theory of Multiple Intelligences.* Basic Books Publisher.

Kaushanskaya, M., & Marian, V. (2009a). The bilingual advantage in novel word learning. *Psychonomic Bulletin & Review: A Journal of the Psychonomic Society, Inc., 16*(4), 705-710. https://doi.org/10.3758/PBR.16.4.705

Kaushanskaya, M., & Marian, V. (2009b). Bilingualism reduces native-language interference during novel-word learning. *Journal of Experimental Psychology: Learning Memory, and Cognition, 35*(3), 829-835. http://dx.doi.org.proxy2.library.illinois.edu/10.1037/a0015275

Kaushanskaya, M., & Rechtzigel, K. (2012). Concreteness effects in bilingual and monolingualword learning. *Psychonomic Bulletin & Review: A Journal of the Psychonomic Society, Inc., 19*(5), 9350941. https://doi.org/10.3758/s13423-012-0271-5

Kempe, V., Bublitz, D., & Brooks, P.J. (2015). Musical ability and non-native speech-sound processing are linked through sensitivity to pitch and spectral information. *British Journal of Psychology, 106*(2), 349-366. https://doi-10.1111/bjop.12092

Keshavarz, M.H., & Astaneh, H. (2004). The impact of bilinguality on the learning of English vocabulary as a foreign language (L3). *International Journal of Bilingual Education & Bilingualism, 7*(4), 295-302. http://www.multilingual-matters.net.proxy2.library.illionois.edu/beb/007/4/default.htm

Koelsch, S., Gunter, T., vonCramon, D.Y., Zysset, St., Lohmann, G., & Friederici, A.D. (2002). Bach speaks: A cortical "language-network" serves the processing of music. *NeoroImage, 17*(2), 956-966. https://doi-10.1016/S1053-8119(02)91154-7

Koelsch, S., Gunter, T., Wittfoth, M., & Sammler, D. (2005). Interaction between syntax processing in language and in music: An ERP study. *Journal of Cognitive Neuroscience, 17*(10), 1565-1577. https://doi-10.1162/089892905774597290

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music.* MIT Press.

Levis, J.M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly 39*(3), 369–377. https://doi.org/10.2307/3588485

Li, M., & DeKeyser, R. (2017). Perception practice, production, practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition, 39*(4), 593-620. https://doi.org/10/101/S0272263116000358

MacCutcheon, D., Ljung, R., Eccles, R., van der Linde, J., Panebianco, C., & Füllgrabe, C. (2020). Investigating the effect of one year of learning to play a musical instrument on speech-in-noise perception and phonological short-term memory in 5-to-7-year-old children. *Frontiers in Psychology, 10*(2865), 1-9. https://doi.org/10.3389/fpsyg.2019.02865

Maess, B., Koelsch, S., Gunter, T.C., & Friederici, A.D. (2001). "Musical syntax" is processed in the area of Broca: An MEG-study. *Nature Neuroscience, 4*(5), 540-545. https://doi.org/10.1038/87502

Magne, C., Schön, D., & Besson, M. (2006). Musician children detect pitch violations in both music and language better than non-musician children: Behavioral and electro-physiological approaches. *Journal of Cognitive Neuroscience. 18*(2), 199-211. https://doi.org/10.1162/jocn.2006.18.2.199

Marques, C., Moreno, S., Castro, S.J., & Besson, M. (2007). Musicians detect pitch violation in a foreign language better than non-musicians: Behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience 19*(9), 1453-1463. https://doi-10.1162/jocn.2007.19.9.1453

Milovanov, R., Huotilainen, M., Termaniemi, M., Valimaki, V., & Esquef, P.A. (2008). Musical aptitude and second language pronunciation skills in school-aged children: Neural and behavioral evidence. *Brain research, 1194*(1), 81-89. https://doi.org/10.1016/j.brainres.2007.11.042

Mokari, P., & Werner, S. (2018). Perceptual training of second-language vowels: Does musical ability play a role? *Journal of Psycholinguistic Research, 47*(1), 95-112. https://doi.org/10.1007/s10936-017-9517-8

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals 51*(3), 617-637. https://doi.org/10.1111/flan.12348

Mroz, A. (2020). Aiming for advanced intelligibility and proficiency using mobile ASR. *Journal of Second Language Pronunciation 6*(1), 12-38. https://doi.org/10.1075/jslp.18030.mro

Munro, M. J., & Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching, 44*(1), 316-327. https://doi.org/10/1017/S0261444811000103

Munro, M.J., & Derwing, T.M. (2015). A prospectus for pronunciation research in the 21st century. A point of view. *Journal of Second Language Pronunciation, 1*(1), 11–42. https://doi.org/10.1075/jslp.1.1.01mun

Nguyen, N., & Fagyal, Z. (2008). Acoustic Aspects of Vowel Harmony in French. *Journal of Phonetics, 36*(1), 1-27. https://doi.org/10.1016/j.wocn.2007.07.001

Ong, J.H., Burnham, D., Escudero, P., & Stevens, C.J. (2017). Effect of linguistic and musical Experience on distributional learning of nonnative lexical tones. *Journal of Speech,Language, and Hearing Research, 60*(10), 2769-2780. https://doi.org/10.1944/2016_JSLHR-S-16-0080

Paillereau, N. M. (2016, July 4). *Do isolated vowels represent vowel targets in French? an acoustic study on Coarticulation*. SHS Web of Conferences. Retrieved July 21, 2022, from https://www.shs-conferences.org/articles/shsconf/abs/2016/05/shsconf_cmlf2016_09003/shsconf_cmlf2016_09003.html

Papagano, C., & Vallar, G. (1995). Verbal short-term memory and vocabulary learning in polyglots. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 48*(1), 98-107. https://doi.org/10.1080/14640749508401378

Perrachione, T.K., Fedorenko, E.G., Gibson, E., Vinke, L., & Dilley, L.C. (2013). Evidence for shared cognitive processing of pitch in music and language. *PLoS one, 8*(8), 1-10. https://doi.org/10.1371/journal.pone.0073372

Roehr-Brakin, K., Gánem-Gutiérrez, G.A., Olivera-Smith, L., & Torres-Marin, M.T. (2021). Are individual differences in cognitive abilities and stylistic preferences related to multilingual adults' performance in explicit learning conditions? *Language Awareness, 30*(4), 391-412. https://doi.org/10.1080/09658416.2021.1969404

Sato, M., & Storch, N. (2022). Context matters: Learner beliefs and interactional behaviorsin an EFL vs. ESL context. *Language Teaching Research, 26*(5), 919-942. https://doi.org/10.1177/1362168820923582

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology 41*(3), 341-349. https://doi.org/10.1111/1469-8986.00172.x

Shabani, M., & Torkeh, M. (2014). The relationship between musical intelligence and foreign language learning: The case of Iranian learners of English. *International Journal of Applied Linguistics & English Literature, 3*(3), 26-32. https://doi.org/10.7575/aiac.ijalel.v.3n.3p.26

Sharajabian, M., & Hashemian, M. (2015). On the relationship between Iranian L2 teachers' pedagogical beliefs and L2 learners' attitudes. *English Language Teaching, 8*(11), 56-68. https://eric.ed.gov/?id=EJ1080299

Van Hell, J.G., & Candia Mahn, A. (1997). Keyword mnemonics versus rote rehearsal in learning concrete and abstract foreign words by experienced and inexperienced foreign language learners. *Language learning, 47*(1), 507-546. https://doi.org/10.1111/0023-8333.00018

Wing, H.D. (1968). *Tests of musical ability and appreciation: An investigation into the measurement, distribution, and development of musical capacity.* Cambridge University Press.

Wong, P.C., Skoe, E., Russo, N.M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10*(4), 420-422. https://doi.org/10.1038/nn1872

Zeromskaite, I. (2014). The potential role of music in second language learning: A review article. *Journal of European Psychology Students, 5*(3), 78-88. https://dxdoi.org/10.5334/jeps.c

Appendix A

Semi-structured Interview

1. What are some tools that have helped you in French pronunciation?
2. Which of these tools have been the most effective and why? Give me a specific example.
3. Which of these tools have been the least effective and why? Give me a specific example.
4. What, if any role, has music had as an impact on your French pronunciation?
5. Please explain this impact a little further.
6. Do you think that a musical background can give someone an advantage in French pronunciation? Why or why not?
7. What do you think of when I say that something is musical?
8. What makes something sound French?

6

Empathy, Equity, and Empowerment in the Spanish Classroom

Clara Burgo
Loyola University Chicago

Challenge

Many world language departments' missions focus on helping our students to become global citizens. Therefore, there is the dire need to contest hegemonic ideologies through the development of emotional and social values such as empathy, empowerment, equity, and resilience for the sake of their success and emotional well-being.

Abstract

World language (WL) educators must explore the role of language to help students express their emotions and opinions on social topics that may ultimately contribute to their emotional well-being while fighting injustice. This article shows how the use of *testimonios* or Latinx literature in the Spanish classroom can be used in higher education to teach them values such as empathy, equity, and inclusion for their own personal development. *Testimonios* are a genre with a pedagogical approach oriented towards social justice and based on an oral Latin American tradition (Delgado Bernal et al., 2012). Latinx faculty can serve as mentors and role models of success, while also positively impacting their communities. These values are presented to create a critical awareness that allows students to take action. This can only take place by making the classroom a safe space where everyone feels included. Finally, teaching resilience based on ethics of care where their local language varieties are validated will provide students with the tools they need to create a resilient community that can become an active agent of change.

Empathy, Equity, and Empowerment in the Spanish Classroom

Emotions are related to well-being and psychological health. World language (WL) instruction contributes to global citizenship by educating students to be empathetic with other cultures, people, and languages (Byram et al., 2013). WL teaching also sensitizes students about human suffering and cultivates empathy and inclusion (Zembylas, 2019). It empowers students to challenge the status quo. However, historically some populations of students in the US have been marginalized from participating in the WL classroom, and they may have felt there was no room for criticality or social justice topics (Glynn & Spenader, 2020) which may, in turn, lead to a lower enrollment in these courses. Teaching social justice involves attention to four aspects: the way structures are affected by social inequity, our roles containing unequal social power, the importance of thinking critically, and the action that is needed to achieve justice (Sensoy & DiAngelo, 2017). Emotional competence is related to higher retention (Conley, 2015), which is particularly important in WL programs where the student enrollment has decreased over the last decades. Therefore, this article aims to provide examples of what we can do in the Spanish classroom, whether in traditional second language (L2) or heritage language (HL) settings; that is, with L2 learners (or additional language learners) who primarily learned Spanish in a classroom setting, and with HL learners who mainly learned Spanish at home. As US higher education instructors, we must align our learning goals with our institutions' mission goals of social justice values, so that our students can develop the needed tools to become global citizens.

Background

Many educators are committed to changing the curriculum to include social justice topics; WL educators should explore the role of language in discrimination or ideology (Osborn, 2006). As educators, we need to create assignments to promote students' voices so that they can express their emotions. *Testimonios* (testimonials) validate students' marginalizing experiences and incite action (Delgado Bernal et al., 2012). They are a tool for reclaiming their Latinx identities through their own writing (Mendoza et al., 2019). It is a method typically used by first-generation Latinx college students as they navigate through higher education (González Herrera, 2020). However, the use of *testimonios* by Latinx could also be implemented in the L2 classroom. It is crucial that L2 students be exposed to authentic materials such as these to be immersed into the Latinx cultures and life narratives such as Rigoberta Menchú's *testimonio*. This *testimonio* advocates for the rights of indigenous communities in Latin America; it focuses on identifying the roots of discrimination in Guatemala.

Latinx literature is also key to promoting empathy. Students can be encouraged to take the role of a specific character to understand their perspective (Floyd & Castek, 2020). This can be easily accomplished by incorporating local writers and artists into the curriculum such as Sandra Cisneros, a Chicago-born Latinx writer, so that they can connect with literature set in a familiar context. One example is Cisneros' critically acclaimed *The House on Mango Street* which takes place in

the Latinx neighborhood Humboldt Park. Students can engage meaningfully with these authentic resources by completing projects that require fieldwork in local communities such as compiling oral histories, engaging in cultural projects, or conducting sociolinguistic interviews. What follows here are descriptions of the roles of empathy, equity, empowerment, and inclusion alongside examples of how Spanish language teachers in higher education settings can incorporate *testimonios* and Latinx literature to build those capacities.

Empathy

Being empathetic implies being understanding of others, which gives an emotional direction to learning through solidarity with other cultures and peoples as global citizens (Ros i Solé, 2013). Empathy is not automatically generated within WL instruction, rather it needs to be nurtured (Finney, 2018). Therefore, higher education should focus on teaching empathy for personal development (Kemmis et al., 2014). WL curriculum should be designed in hopes of developing empathy and understanding the complex issues that many communities are facing. Critical content-based instruction allows students to examine issues of power and equity, develop critical consciousness, and engage in action (Glynn & Spenader, 2020). An example of how to teach empathy is described in Sellers (2021). She presented a project-based learning unit on empathy in a Latin American literature course at the university level. Learners had to articulate how texts elicit empathy, how they themselves experience empathy, and how to make connections with the time period in the literature. Another example of empathy development can be found in Di Stefano et al. (2021) through the design of a vlog activity that exposed students to the Latinx community. The recording of this vlog allowed students to engage with authentic materials and to advocate for Latinx individuals in the US by focusing on a social justice topic, reacting to it, and reflecting upon aspects that impacted them.

Empathy and compassion are key in a caring instructor (Emeric Csaszar et al., 2018). Not all Latinx speak Spanish, and linguistic assimilation is sometimes seen as an attack on Spanish and a form of ethnic erasure (Davis & Moore, 2014). The silencing of Spanish disconnects Latinx from their cultural identities since language is an important marker of Latinx culture. Added to feelings of loss and shame, this loss could influence work with their communities (Mendoza et al., 2019). Therefore, teachers need to be empathetic to develop a positive learning environment for all and recognize that teacher stress may have negative consequences and may undermine their relationship with their students. As instructors, we need to advocate for encouraging Spanish language learning while also honoring Latinx identities of individuals who do not speak Spanish as well.

Equity

In order to teach our students about empathy towards local communities, a deep understanding of ongoing fights towards equity is needed. The theoretical framework of Critical Latinx Indigeneities (CLI) would help in designing a Latinx educational agenda and a curriculum accounting for the hybrid hegemonies of

Latinx communities; an agenda that faces educational challenges to fight inequity. That is, CLI encourages us to envision teaching approaches that recognize the diversity of Latinx experiences (Calderón & Urrieta, 2019). Moreover, CLI reflects on how to improve the understanding of Indigenous immigrant experiences (Barillas Chón, 2019).

One way to fight for equity is through mentorship. Mentorship is especially important when working with US-born Latinx students who report higher levels of depression than foreign-born Latinx. (Garcini et al., 2020). As these researchers claim, one hypothesis to explain the mental health disparity between these two groups is that US Latinx experience higher risks of depression when facing discrimination that is difficult to manage, despite their higher levels of education and the fact that they are often English dominant. Therefore, offering social support may be a significant factor to help US Latinx deal with discrimination. Social support could be encouraged through critical service-learning opportunities where students need to reflect on their service to offer help in community centers in the form of translations or mentoring.

Another way of fighting for equity is by reviewing our current textbooks. Research has shown that Spanish-language textbooks offer a reductionist portrayal of indigenous communities and Afr-Latinx, reinforcement of stereotypes, and understatement of US local varieties of Spanish (Padilla & Vana, 2019). These ideologies influence instructors' attitudes and opinions that are passed on their students (Quan, 2021). In fact, Latinx heritage teachers of Spanish in the US have experienced negative judgments regarding their Spanish varieties and linguistic practices (Bustamante & Novella, 2019). These are racialized discourses that influence teachers' and learners' beliefs and attitudes with radical consequences such as leaving the profession (Kissau et al., 2019). It is critically important to focus on sociolinguistic variation and for language instructors to fight for equity.

Empowerment

Once students are informed on empathy and equity, empowerment can emerge. Academic language is a raciolinguistic ideology that frames local varieties as deficient (Flores & Rosa, 2015). This raciolinguistic ideology is based on the premise that the origin of racial inequalities is in the deficient linguistic varieties of racialized communities and the solution is to change their varieties (Rosa & Flores, 2017). Therefore, we need to center local dialects in our curriculum. As bilinguals, students will engage in translanguaging practices to communicate, and instructors need to be informed about ways to support them. Hence the importance of including authentic bilingual materials, resources, and texts and the study of bilingual writers and artists.

Latinx faculty need to be represented on campus so that they can serve as role models and interpreters. With these relationships there should be an ambience of solidarity against microaggressions and any kind of discrimination Latinx individuals may face in higher education. Along this line, Latinx faculty provide the students with emotional support and motivation, supporting student retention and completion of their degrees (Mendoza et al., 2019). Spanish is used to maintain

social bonds among the Latinx communities. Therefore, Latinx who do not speak Spanish may feel excluded by those who do (Reyes, 2018). There are challenging ideologies that identify being Latinx with speaking Spanish and acknowledging a national right for their identity (Lado & Del Valle, 2022). However, even if some Latinx lose the command of Spanish, they do not lose their Latinx identity.

In the same vein, there is a contrast between the social stigma attached to the language and the strong identity marker that the language means for many young Latinx individuals, which accounts for theories of resistance (Alvero & Pattichis, 2022). The classroom should be a safe space that incorporates a critical Latinx pedagogy based on an ethics of care and creating a sense of belonging. Critical Latinx pedagogy empowers students and provides opportunities for them to get a deeper understanding of Latin Americans in the US. This pedagogy makes students feel welcomed in the classroom while challenging assumptions about controversial topics such as immigration, race, or culture. We should encourage students to see their histories as a source of pride, validating their presence in higher education to see themselves as producers of academic knowledge (Casavantes Bradford & Morales, 2021). Teaching for social justice is based on the idea that education is liberatory and the classroom emancipatory. As instructors, we are responsible for creating these spaces and facilitating critical consciousness (Andrzejewski & Carson Baggett, 2020).

One example of an effort to share information and enact social change related to Latinx empowerment is the podcast series "Block Chronicles", produced by the University of North Carolina, Chapel Hill. This podcast series is used to disseminate knowledge and to enact political agency. It is a form of public pedagogy to push back against systematic violence and silencing (Carrillo & Mendez, 2019). Teachers should use resources like this podcast as a model for empowerment and develop tasks in which students can write and dialogue, with the ultimate goal of empowering students to foster their critical language awareness as they use their HL. Students are given opportunities to reflect on their experiences as bilingual and bicultural individuals as well as the ideologies that have been imposed on them (Alamilla, 2021). HL learners could interview their family members or Latinx leaders of their community and record a podcast to reflect on linguistic ideologies about their Spanish. L2 learners could also interview HL learners in the classroom or Latinx students from their university and record a podcast to reflect on their linguistic experiences in the US. In the same way that a cultural component is always expected in a WL course, we should expect a HL course to utilize critical pedagogy to empower students to learn about their heritage languages and cultures.

Inclusion

Inclusive education favors the development of all students and the cohesion of all members of the community, reduces the negative impact on their well-being, and improves their coexistence in the classroom (Navarro-Mateu et al., 2019). It is an asset and an indicator of quality in higher education (Baltaru, 2020). Inclusive pedagogy involves all students and inclusive instructors foster relationships

of trust with their students and promote interest in learning (Echeita et al., 2016). Inclusive pedagogy operates on three premises: all students bring value, difference is valued, as is working with students and their communities (Gale & Mills, 2013). Critical Language Awareness facilitates the discussion of linguistic ideologies and their impact on social change, including resistance to change (Alim, 2010). As educators, we must validate all Spanish varieties and cultures of Latinx individuals in order to create an inclusive environment. If we include Latinx literature in our curriculum, students will feel validated and empowered to challenge the status quo (Valdez, 2019). We can model inclusion and engagement of students' literacy centered on their heritage and on their own experiences as Latinx. In doing so, we are acknowledging and honoring their backgrounds (Foulis & Barajas, 2019). An example of this would be the works by Gloria Anzaldúa: the short story *How to Tame a Wild Tongue* (Anzaldúa, 1987a) and the poem *To Live in the Borderlands* (Anzaldúa, 1987b). The short story is about language rights in the US and the poem about the struggle that marginalized groups go through. In order to build inclusion using these two works, students could write a list of suggestions on how to protect linguistic and identity rights in the US and consider how these actions could be implemented in their communities.

As mentioned earlier, historically, some US minorities have historically been marginalized from participating in the language classroom for a variety of reasons including a Eurocentric curriculum, a curriculum with a focus on language, or a curriculum that does not prioritize social topics such as race or social justice in lower-level language courses (Kubota et al., 2003). However, there is a natural connection between teaching language and teaching social justice. Social justice promotes critical awareness to create an inclusive environment for students and to provide transformative experiences for them (Dover, 2013). A critical approach using culturally sustaining pedagogy is needed within language teaching to offer students a more inclusive space to address issues of belonging and attrition and to promote criticism (Paris & Alim, 2017).

Resilience

Resilience can become a protective factor. Students with resilience can transform situations of stress into opportunities for learning and can better adapt to the university (Gómez- Molinero et al., 2018). Unfortunately, Latinx individuals and other minorities have often experienced resilience in silence (González, 2020). *Familismo* is a source of strength, a Latinx cultural value related to the importance and support of the family (Jiménez et al., 2021). It becomes a protective factor in Latinx college students. Higher levels of *familismo* correlates with lower levels of depression (Corona et al., 2017). Resilience must be fostered in both Latinx faculty and their students. Oral history projects consisting of interviews with family members about a relevant topic such as immigration, could be used in instruction. Students should create the interview questions to gain an understanding of their family members' perspectives, conduct the interview, and write a final report with the findings and a reflection on the topic focusing on how these stories are a sign of resilience. Bringing students' family stories to the classroom as part of the curriculum is a way to embrace *familismo*.

To foster resilience, networks among faculty members should be built for faculty of color by creating safe online spaces to learn about leadership and connect with existing

Latinx leaders (Marín et al., 2018). Considering that dominant language ideologies affect Latinx perceptions of their abilities, we must remember how these ideologies generate feelings of alienation and insecurities in Spanish courses in higher education contexts (Holguín Mendoza, 2018).

New instructors must develop resilience and persistence of their own in order to teach these concepts to students, who are the targets of negative stereotypes. Additionally, moving from a deficit perspective to an asset-rich perspective means recognizing the cultural capital of the students (Villegas-Reimers, 2021). Programs that strengthen individual resilience may lead to more resilient communities if participants know how to do so. Resilient Latinx can be agents of change by helping their Latinx peers to cope with difficulties, identify their resilience, and maintain their ties to their communities (Tellez Lieberman et al., 2019).

Conclusion

WL education has the potential to teach students how to develop feelings of empathy to become global citizens while empowering students to challenge the status quo. Unfortunately, some populations of students in the US have been marginalized from participating in the WL classroom. Social justice topics are more important than ever and must be included in the curriculum from novice levels of proficiency so that our students can become global citizens and agents of change. Hence the importance of designing assignments that incorporate *testimonios,* readings from US Latinx literature such as those by Anzaldúa, and podcasts that can serve as tools to teach equity, develop empathy, and build inclusion, ensuring that students feel they have a voice in the classroom. While it can be challenging to find enough authentic materials that are directly connected to local communities, the suggestions here provide examples for instructors. Furthermore, Latinx faculty are needed as well to teach values of empowerment and resilience as successful role models and to make a real change in the communities. In many places there are not enough Latinx faculty in leadership roles. Future research should explore and address these issues in more depth, especially theoretical frameworks such CLI with a Latinx education agenda in mind, to guide instructors in this direction.

References

Alamilla, R. (2021). Using classroom conversation to embrace ethnolinguistic identity and heritage language. *Ianua. Revista Philologica Romanica, 21,* 1- 35.

Alim, H. (2010). Critical language awareness. In N. Hornberger (Ed.), *Sociolinguistics and language education* (pp. 205– 230). Bristol, UK: Multilingual Matters.

Alvero, A.J., & Pattichis, R. (2022). Linguistic and cultural strategies: Identification and analysis of Spanish language usage in college admissions essays. *SocArXiv Papers,* 1-50. https://doi.org/10.31235/osf.io/wmsre

Andrzejewski, C.E., & Carson Baggett, H. (2020). Magic and hocus pocus: Teaching for social justice in a qualitative methods course. *The Qualitative Report, 25*(3), 864-875. https://nsuworks.nova.edu/tqr/vol25/iss3/19

Anzaldúa, G. (1987a). *How to Tame a Wild Tongue. Borderlands/La frontera: The new mestiza* (pp. 75–86). San Francisco, CA: Aunt Lute Books.

Anzaldúa, G. (1987b). *To live in the Borderlands. Borderlands, La frontera: The New Mestiza*: San Francisco, CA: Aunt Lute Books.

Baltaru, R. D. (2020). The rise of agentic inclusion in the UK universities: Maintaining reputation through (formal) diversification. *Studies in Higher Education.* https://doi.org/10.1080/03075079.2020.1739015

Barillas Chón, D. W. (2019). Indigenous immigrant youth's understandings of power: Race, labor, and language. *Association of Mexican American Educators Journal, 13*(2), 15-41.

Bustamante, C., & Novella, M.A. (2019). When a heritage speaker wants to be a Spanish teacher: Educational experiences and challenges. *Foreign Language Annals*, *52*, 184–98.

Byram, M., Holmes, P., & Savvides, N. (2013). Intercultural communicative competence in foreign language education: questions of theory, practice and research. *The Language Learning Journal, 41*(3), 251–253.

Calderón, D., & Urrieta, L. (2019). Studying in relation: Critical Latinx indigeneities and education. *Equity & Excellence in Education*, *52*(2-3), 219-238. DOI:10.1 080/10665684.2019.1672591

Carrillo, J.F., & Mendez, J. (2019). Inner work, public acts: Making a case for public pedagogy and spatial justice within Latinx communities. *Urban Review, 51,* 444-456. https://doi.org/10.1007/s11256-018-00495-x

Casavantes Bradford, A., & Morales, A.E. (2021), Toward a critical Latinx pedagogy: A multi-generational reflection on teaching and learning in the U.S. Latinx history survey and beyond, *Journal of Latinos and Education,* 1-11. https://doi.org/10.1080/15348431.2021.1910035

Cisneros, S. (1984). *The house on Mango Street*. Arte Público Press.

Conley, C. S. (2015). SEL in higher education. In J.A. Durlak, C.E. Domitrovich, R.P. Weissberg, & T.P. Gullotta (Eds.), *Handbook of social and emotional learning: Research and practice.* (pp. 197-212). New York: The Guilford Press.

Corona, K., Campos, B., & Chen, C. (2017). Familism is associated with psychological well- being and physical health: Main effects and stress-buffering effects. *Hispanic Journal of Behavioral Sciences*, *39*(1), 46–65. https://doi.org/10.1177/0739986316671297

Davis, T. Y., & Moore, W. L. (2014). Spanish not spoken here: Accounting for the racialization of the Spanish language in the experiences of Mexican migrants in the United States. *Ethnicities, 14,* 676–697. doi:10. 1177/1468796814523740

Delgado Bernal, D., Burciaga, R., & Flores Carmona, J. (2012). Chicana/Latina testimonios: Mapping the methodological, pedagogical, and political. *Equity & Excellence in Education, 45*(3), 363–372. https://doi.org/10.1080/10665684 .2012.698149

Di Stefano, M., Almazán-Vázquez, A., Yepes-Amaya, W., & Britton, E. R. (2021). "Tu lucha es mi lucha": Teaching Spanish through an equity and social justice lens. *Foreign Language Annals, 54,* 753–775. https://doi.org/10.1111/ flan.12578

Dover, A. (2013). Teaching for social justice: From conceptual frameworks to classroom practices. *Multicultural Perspectives, 15*(1), 3-11.

Echeita, G., Sandoval, M., & Simón, C. (2016) Notas para una pedagogía inclusiva en las aulas [Notes for an inclusive pedagogy in the classrooms]. In Actas IV Congreso Iberoamericano de Síndrome de Down, Salamanca, Spain, 16–17 March 2016. Available at: http://www.psicologiauam.es/gerardo.echeita/DOCUMENTOS,%20ARTICULOS,%20PONENECIAS,/Echeita,%20Sandoval%20y%20Simon.%202013.%20Congreso%20SD.pdf

Emeric Csaszar, I., Curry, J.R., & Lastrapes, R.E. (2018). Effects of loving kindness meditation on student teachers' reported levels of stress and empathy. *Teacher Education Quarterly, 45*(4), 93-116.

Finney, S. (2018). Developing intercultural empathy through a strategy-based simulation in intermediate Spanish. *The Language Learning Journal, 48*(6), 754-767.

Flores, N., & Rosa, J. (2015). Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review, 85*, 149–171. https://doi.org/10.17763/0017- 8055.85.2.149

Floyd, R., & Castek, J. (2020). Academic, emotional, and social growth in the second language classroom: A study of multimodality. In McCray, C.(Ed.), *Multifaceted strategies for social-emotional learning and whole learner education,* (pp. 163-188). IGI Global: Hershey, PA.

Foulis, E., & Barajas, J. (2019). Weaving our histories: Latin@ ethnography in the heritage language classroom. *Journal of Folklore and Education, 6,* 98-107.

Gale, T., & Mills, C. (2013) Creating spaces in higher education for marginalised Australians: Principles for socially inclusive pedagogies. *Enhancing Learning in the Social Sciences, 5*(2), 7–19.

Garcini, L.M., Chen, M.A., Brown, R., LeRoy, A.S., Cano, M.A., Peek, K., & Fagundes, C. (2020). Abrázame que ayuda (Hug me, it helps): Social support and the effect of perceived discrimination on depression among US-and foreign born Latinx in the US. *Journal of Racial and Ethnic Health Disparities, 7*(3), 481-487.

Glynn, C., & Spenader, A. (2020). Critical content-based instruction for the transformation of world language classrooms. *L2 Journal, 12*(2), 72-93.

Gómez-Molinero, R., Zayas, A., Ruiz-González, P., & Guil, R. (2018). Optimism and resilience among university students. *International Journal of Developmental and Educational Psychology, 1*(1), 147-154.

González, E. (2020). Foreword: Understanding Latina/o resilience. *International Journal of Qualitative Studies in Education, 33*(8), 791-795. https://doi.org/10.1080/09518398.2020.1783016

González Herrera, M.A. (2020). Amor y apoyo: Lecciones de Latinx families in nourishing resilience to first and second-generation college students. *Challenger: A McNair Scholar Paper Series*, 19-29.

Holguín Mendoza, C. (2018). Critical Language Awareness (CLA) for Spanish heritage language programs: Implementing a complete curriculum. *International Journal of Multilingualism*, *12*(2), 65–79. https://doi.org/10.1080/19313152.2017.1401445

Jiménez, A., Piña-Watson, B., & Manzo, G. (2021). Resilience through family: Family support as an academic and psychological protective resource for Mexican descent first-generation college students. *Journal of Hispanic Higher Education,* 1-12.

Kemmis, S., Wilkinson, J., Edwards-Groves, C., Hardy, I., Grootenboer, P., & Bristol, L. (2014). *Changing practices, changing education.* Singapore: Springer Science & Business Media.

Kissau, S., Davin, K.J., & Wang, C. (2019). Aspiring world language teachers: Their influences, perceptions, and commitment to teaching. *Teaching and Teacher Education, 78,* 174-82.

Kubota, R., Austin, T., & Saito-Abbott, Y. (2003). Diversity and inclusion of sociopolitical issues in foreign language classrooms: An exploratory survey. *Foreign Language Annals, 36*(1), 12-24.

Lado, B., & Del Valle, J. (2022). For critical language awareness and against the "exclusive-use-of-the-target-language" myth: The effects of sociolinguistic content in English in an elementary Spanish classroom. *L2 Journal, 14*(3), 1-24.

Marín, A.S., Romero Marín, M., & Vázquez, L. (2018). Educational leadership development: Moving from a deficit model to an ecologically strengths-based model for Latinx. In Rodríguez, C., Valle, F., & Martínez, M.A. (Eds.), *Latino educational* leadership: Serving Latino communities and preparing Latina/o Leaders across the *P-20 pipeline,* (pp. 77-96). Charlotte, NC: Information Age Publishing.

Mendoza, N.S., Lechuga-Peña, S., López, K., & Jackson, K.F. (2019). "Mi'jita, what for? Exploring bicultural identity of Latina/Chicana faculty and implications for supporting Latinx students in social work. *Journal of Women and Social Work, 34*(2), 259-276.

Navarro-Mateu, D., Franco-Ochoa, J., Valero-Moreno, S., & Prado-Gascó, V. (2019). To be or not to be an inclusive teacher. Are empathy and social dominance relevant factors to positive attitudes towards inclusive education? *PLOS ONE, 14*(12): e0225993. https://doi.org/10.1371/journal.pone.0225993

Osborn, T. (2006). *Teaching world languages for social justice: A sourcebook of principles and practices.* Mahway, NJ: Lawrence Erlbaum Associates.

Padilla, L.V., & Vana, R. (2019). Ideologies in the foreign language curriculum: Insights from textbooks and instructor interviews. *Language Awareness, 28*(1), 15-30.

Paris, D., & Alim, H. (2017). *Culturally sustaining pedagogies: Teaching and Learning for justice in a changing world.* New York, NY: Teachers College Press.

Quan, T. (2021). Critical approaches to Spanish language teacher education: Challenging raciolinguistic ideologies and fostering critical language awareness. *Hispania, 104*(3), 447-459.

Reyes, D. (2018). *Learning to be Latino*. New Brunswick, NJ: Rutgers University Press.

Ros i Solé, C. (2013). Cosmopolitan speakers and their cultural cartographies. *The Language Learning Journal, 41*(3), 326–339.

Rosa, J., & Flores, N. (2017). Unsettling race and language: Toward a raciolinguistic perspective. *Language & Society, 46,* 621–647. https://doi.org/10.1017/S0047404517000562

Sellers, J. A. (2021). Elicit empathy through the study of literature in the world language classroom. *Hispania, 104*(3), 363- 367.

Sensoy, O. & DiAngelo, R. (2017). *Is everyone really equal? An introduction to key concepts in social justice education* (2nd ed.). New York, NY: Teachers College Press.

Tellez-Lieberman, J. Lobban, K., Flores, Z., Giordano, K., Nolasco-Barrientos, E., Yamasaki, Y., & Martínez-Donate, A.P. (2019). "We all have strengths": A retrospective qualitative evaluation of a resilience training for Latino immigrants in Philadelphia, PA. *Health Equity, 3*(1), 548-556. https://doi.org/10.1089/heq.2019.0070

Valdez, A. (2019). Benefits of culturally relevant literature on Latino students. *Capstone Projects and Masters' Theses, 536.* https://digitalcommons.csumb.edu/caps_thes_all/536

Villegas- Reimers, E. (2021). Educating teachers to work with Latinx children and families in challenging sociopolitical contexts and times. In Bempechat, J., & Jiménez-Silva, M. (Eds.), *Latinx experiences in US Schools: Voices of students, teachers, teacher educators, and education allies in challenging sociopolitical times,* (pp. 131- 151). Lexington Books.

Zembylas, M. (2019). From the ethic of hospitality to affective hospitality: Ethical, political and pedagogical implications of theorizing hospitality through the lens of affect theory. *Studies in Philosophy and Education. 39*(2), 37-50. https://doi.org/10.1007/s11217-019-09666-z