# WORLD LANGUAGE TEACHERS' ASSESSMENT BELIEFS AND GRADING PRACTICES

Melinda Cotton

Submitted to the faculty of the School of Education
in partial fulfillment of the requirements
for the degree
Doctor of Education
in the Department of Curriculum and Instruction
Indiana University
March 2025

Accepted by the School of Education Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Education.

Doctoral Committee

_____

Karen Wohlwend, Ph.D.

_____

Sharon Daley, Ph.D.

_____

James Damico, Ph.D.

Date of Defense
February 20, 2025

**Acknowledgements**

I would like to thank several people for their support throughout my doctoral journey. Starting coursework in August 2020, six months into the pandemic, was incredibly challenging, and I could never have made it through without my Cohort Five colleagues from all over the world. I appreciate all the laughter and tears, the commiseration and compassion of this wonderful group.

My professors every semester pushed me to think more deeply about my profession than ever before. All of them impacted me greatly, with special thanks to Dr. Wohlwend, Dr. Damico, Dr. Daley, Dr. Nyikos, Dr. Coronel-Molina, Dr. Pawan, Dr. Lester, Dr. Hickey, and Dr. BrckaLorenz. My dissertation committee was amazing, and I appreciate their input, patience, and advice.

I can never express enough gratitude to my participants, who are my work family. I learned so much from them through this process. Their insights increased the respect that I hold for each one of them. I am so thankful to work alongside them every day, sharing in the highs and lows of public-school education.

And finally, I want to thank my children, who have cheered me on for the past five years. They have listened to all my crazy dissertation ideas, my frustrations, my triumphs, and my failures. I love them with all my heart. They are my rock.

Melinda Cotton

# WORLD LANGUAGE TEACHERS' ASSESSMENT BELIEFS AND GRADING PRACTICES

Gradebooks take center stage in educational settings and are used to determine a student's eligibility to pass a class, graduate, or attend college. Grading practices are often private and vary greatly, even in the same department. This study examined one world language department in a U.S. public high school to compare the assessments that were entered into the teachers' gradebooks, why and how those assessments were administered, and subsequently how they were weighted in the gradebook. I conducted interviews with seven world language teachers to explore the range of teachers' beliefs about student achievement in world language, since this achievement could be assessed in many different ways, such as speaking, writing, reading, and listening. The qualitative research was conducted through individual interviews, to inquire about the teachers' gradebook setups, their beliefs about student achievement, and each of the assessments found in their gradebook reports. The interviews were recorded on Zoom, transcribed, analyzed, and coded for themes through the lens of Bakhtin's authoritative discourse. Assessment and gradebook practices in the interviews revealed diverse, patchwork views about world language achievement. The beliefs, ranging from macro to micro levels, related student world language achievement to understanding concepts, accumulating knowledge, communicating, negotiating meaning, participating in communities of practice, reciting and reading aloud, and connecting to cultures. The teachers also believed to varying degrees that non-academic factors like effort should be reflected in students' grades. These results indicated that teachers' beliefs may contribute to teachers' avoidance of certain kinds of important assessments. Self-reflection along with individual and collaborative gradebook

evaluations could aid teachers in finding gaps in their assessment practices. In departments that require the implementation of common assessments, holding diverse achievement beliefs can cause conflict, but collaboration and mutual training might improve the departments' overall assessment and grading practices.

Keywords: world language, assessment, grading, gradebooks, teacher beliefs, contact zone

Karen Wohlwend, Ph.D.

Sharon Daley, Ph.D.

James Damico, Ph.D.

## Table of Contents

# Chapter 1 Introduction

## Statement of Problem of Practice

Gradebooks and report cards are central to our current educational system. There are many stakeholders invested in the reliability and validity of teachers' grading systems, including students, parents, local, district, state, and federal school officials, colleges, and even the community at large, as students leave high school with a diploma and a transcript that reflect their basic competencies in specific professional and educational domains. However, grading is a seemingly private matter, not often shared or standardized with other teachers even in the same school (McMillan & Nash, 2000). Even in schools with explicit grading policies and guidelines, gradebooks show significant variation among teachers at the same grade level or academic department (Brookhart, 1993; Guskey & Link, 2019: McMillan, 2001). Many studies (e.g., Guskey & Link, 2019; Stiggens et al., 1989; Sun & Chen, 2014) have shown that teachers ignore expert recommendations for grading, and often include grading components that are not related to academic achievement, like motivation, effort, and behavior. If grades are to be recognized as reliable and valid, then arbitrary values placed on irrelevant grading components can render those grades impossible to interpret as evidence of students' achievement.

This variation occurs not just in grade calculation, but also in the practices and perceptions of what, how, and why to assess students, even when curricular assessment expectations are standardized and published (Leung, 2004). In the field of world language, this variation comes from the complexity of the nature of language learning and the approaches that can be used to measure it. Language learning can be seen as the correct response to stimulus (behaviorist theory), understanding and problem-solving (constructivist theory), or it can be seen as the interaction between the learner and his environment (sociocultural theory) (Leung, 2013).

Language learning can also be conceptualized in different ways– grammar-oriented as a set of lexical and syntactical rules or as discourse-oriented, as a means of negotiating meaning through social and power relationships (Leung, 2013). World language assessment approaches have evolved quickly since the 1960's, with focuses on skills and elements (traits) testing, performance assessments, pragmatic language testing, communicative language testing, interaction-ability testing, task-based performance assessments, and interactional approaches to language testing (Bachman, 2007). "Language attainment has been expressed in formulations of progression that are often a patchwork of different ideological, institutional, and societal aspirations and requirements that are not necessarily conceptually or educationally coherent" (Leung & Scarino, 2016, p. 88).

Assessments can be used to measure a student's level of language acquisition after instruction has occurred as summative assessments, or they can be used to guide the teacher's instruction by identifying students' strengths and weaknesses along the way in formative assessments. Teachers can also allow the students to self-assess or assess their peers in their learning process. This myriad of choices in the "what," "why," and "how" of language assessment leads to high levels of variation in teacher assessment and grading practices.

## Purpose

In this research project, I explored the grading and assessment practices of one U.S public school world language department. I attempted to answer Leung's (2004) call to discover "What do teachers look for when they are assessing? What theory or "standards" do teachers use when they make judgments and decisions?" (p. 880). I wanted to understand how the teachers valued these assessment priorities in terms of grade weight calculations in their gradebook.

## Background and Significance

Variation in grading practices and teachers' knowledge about what and how to grade has been studied extensively since the 1950's, with a flurry of studies occurring in the 1980's and 1990's (e.g., Brookhart, 1993; Frary et al., 1993; Gullickson, 1985; Stiggins, 1989; Wise, 1991). Even more recent studies, like Guskey and Link's (2019) self-reported teacher survey, have shown that teachers still use a wide variety of components and procedures in their grading and assessing, despite new understandings of best practices based on empirical evidence produced in the studies of the past forty years. In many of these studies on grading, foreign language teachers were underrepresented in the sampling (e.g. Frary et al., 1993; Kunnath, 2017; Randall & Engellhard, 2009). Only in the past twenty years have researchers looked at grading specifically in foreign language classrooms (e.g. Cheng et al., 2004; Cheng & Wang, 2007; Nowruzi, 2021; Sun & Cheng, 2014). Both recent and older studies continue to show that teachers have low Assessment Literacy, a term originally used by Richard Stiggins (1991) to describe assessors' ability to describe their students' achievement objectives in specific terms, collect quality student achievement data by designing appropriate, clear assessments, and use that data to give the students feedback about their learning.

### Grading Issues in World Languages

One complication for grading in foreign language classrooms is the nature of the school subject itself. Larsen-Freeman and Freeman (2008) explained that language has always been hard to pin down as a school subject because of its lack of disciplinary home and its vastness. They traced language as a school subject matter throughout recent history, starting at the beginning of the 20th century, when language was studied as a means to support the arts (literature, art, music, etc.). They described how language was continually redefined through its

teaching methodologies throughout history: grammar-translation, audio-lingual, creative-constructivist, natural approaches, humanistic, communicative, interactional, and critical. Goals for proficiency and competency have changed from multilingualism (native-like proficiency in more than one language) to plurilingualism (partial contextualized usage-based proficiency in more than one language). “The problem [with language as a school subject] lies in attempts to corral it, atomize it, and fix its units to make them learnable as subject matter,” (p. 162). Larsen-Freeman and Freeman believed that, because of rapid globalization and simultaneous localization and hybridization of languages, there is no agreed upon content to be taught in school language classes, and that we have reached a point of post-disciplinarity. They postulated that the destabilization of languages as content, with the surrounding availability of constantly changing language-in-use, leads to tension and even to the decomposition of the field itself.

The language learning field may not have yet reached the extreme point of decomposition or post-disciplinarity, but problems certainly persist in measuring foreign language achievement. Bachman (2007) explained that these issues arise because language is both the object to assess, as well as the instrument of measurement. Many other researchers believe that problems in measuring foreign language achievement revolve around the assessment construct (e.g., Elder et al., 2017; Jamieson, 2014; McNamara, 2007). Generally, assessment constructs guide teachers in what should be tested, and are based on standards, research, theory, and teachers’ experience and practice (Purpura, 2016). But Bachman (2007) suggested that language testing practitioners, including classroom teachers, base their assessment construct definitions, subsequent observations, and interpretations, on their own personal perspective of knowledge, their own purposes, their view of the world, and how they view assessment constructs, and the contexts of the constructs.

### *Defining the Construct of World Language*

In 1996, The American Council on the Teaching of Foreign Languages (ACTFL) published the first version of the *Standards for Foreign Language Learning in the 21st Century* (The National Standards Collaborative Board, 2015) to provide a standardized framework for the teaching of languages. These standards included the 5 C's of Communication, Culture, Comparisons, Communities, and Connections. However, most of the focus in operationalizing the assessment of learner outcomes has been on the Communication standard, with Connections being called "the lost C" (Troyan, 2012), since its implementation in the classroom has been studied the least. Campbell and Duncan (2007) explained that while these standards attempt to provide the content of what is to be assessed in language as a subject matter, they do not solve the problem of the specific performance expectations for each standard. For example, while the Culture Standard focuses on the products, practices, and perspectives of a culture, practitioners still do not agree on what specific products, practices, and perspectives to teach or assess (Troyan, 2012). Alderson et al. (2017) believe that there are still big questions regarding the multidimensionality of language skills, the boundaries of language abilities, language variability by context, and the values that define the language assessment construct.

### *World Language Learning Theories and Assessment*

Adding to the complication of the construct of the language to be learned, is the nature of language learning, which is tied to the three major theories of learning—behaviorist, cognitive/constructivist, and situated/sociocultural. Mary James (2006), reminded us that none of the three theories are methodologies for instruction or assessment; they are simply theories about how learning occurs. Behaviorism still strongly influences contemporary trends in instructional and assessment practices, like differentiated learning in homogenous ability groups, that help

learners accumulate more and more complicated skills as students progress through a hierarchy of abilities in the domain. Behaviorist-oriented assessments track what knowledge students possess, not what they understand, or how they interact with their environment or their resources. Alternatively, summative assessments aligned with cognitive/constructivist theory might take the form of concept mapping and problem solving, while formative assessments might involve accessing students' prior knowledge to add to their understanding. Situated/socioculturally-aligned assessments could include ethnographic observation and inference, possibly using portfolios as artifacts. James suggested that teachers often blend the three approaches to form eclectic approaches, and sometimes synthesis approaches, but almost never purist approaches to learning theories. This could mean that teachers are using assessments that do not align with the language learning theories that they espouse or the methodologies that they are using in their instruction.

With all of the changing constructs of language as a subject matter, the differing and contradictory views on language learning, and the commonly "broken rules" of grade calculations, it is no wonder that the everyday language teacher struggles to keep up with best practices. Veteran and new teachers alike used similar grading and assessment procedures, according to a recent survey by Guskey and Link (2019), so assessment literacy is not linked to years of experience. They speculated that even though new teachers are trained in modern, researched-based methods of grading and assessment in their pre-service training, they become overwhelmed by all the demands and limitations of actual classroom practice and conform and adapt to veteran methods that are outdated or less backed by research. Veteran teachers are often less trained in current research-based grading policies and often continue to follow traditional practices that may not be as effective.

### *World Language Assessment in Different Contexts*

Many of the recent studies about world language teachers' assessment and grading practices have taken place in Europe and Asia (e.g., Bahtiar & Purnawarman, 2020; Berry et al., 2019 Brumen & Cagran, 2011; Lam, 2015; Lam, 2019; Mede & Atay, 2017; Nowruzi, 2021; Sun & Cheng, 2014; Tsagari & Vogt, 2017). Those studies found that grading practices were context-driven, depending on the cultural practices and traditions in the educational institutions. Many of them also explored world language teachers of English as a Foreign Language. English is a dominant language in social media culture and world business, which makes it highly desirable as an acquired skill. Demand for learning English in other countries is high because students and their parents see it as valuable to their future careers (Yan et al., 2018). European and Asian students are also more likely to speak or be exposed to multiple languages as children, so learning languages is already a part of their cultural upbringing. English language learning often starts in the primary grades, and those students have at least some other language proficiency by the time they reach secondary grades (Brumen & Cagran, 2011). Educational institutions recognize the value of language acquisition, and often prioritize high stakes testing as a part of students' educational assessment (Cheng et al., 2004). This institutional priority for passing high stakes language tests for university entrance or for furthering their careers trickles down to world language teachers in these countries, who must prepare the students to be successful (Chen et al., 2004; Wang et al., 2020; Yan et al., 2018). Student and teacher language learning motivation is high because of these external factors, and this contextualizes the teachers' grading and assessment practices.

However, in the United States, students already speak the dominant world language and thus have less motivation to learn a second minority language (Stein-Smith, 2019). Languages

are often elective classes, depending on the state, and not required for graduation from high school. Language learning typically begins in secondary school rather than primary school. During state budget shortages, language courses and teachers often find themselves as first contenders for elimination. Many university programs are reducing or eliminating their world language programs (Quinn, 2023). World language teachers often feel forced to recruit students to their language programs in order to keep program numbers high enough to justify their jobs. The ACTFL website has a whole webpage entitled, “Save a Language Program” (ACTFL, 2025b). The school where this current study took place holds several events during the year where the language teachers display posters, student work, food, and cultural samples, to entice future students to sign up for language classes. The lesser taught languages feel the pressure to recruit the most, since their job positions might be reduced to part-time from full-time, or be eliminated altogether,  if they cannot maintain full class rosters. Very few U.S. institutions require high stakes language testing as part of the educational pathway (Devlin, 2015). Students see languages as a diversion, rather than a core academic subject that they will need for their future careers. Most American students and parents are monolingual and have no experience with the advantages of speaking more than one language. All of these external factors differentiate the American language learning climate from European or Asian contexts. This study aimed to look at grading and assessment practices in a context where languages are less highly valued.

## Research Questions

In world language departments, teachers often carry complex, changing views on learning and language, and their assessments might align with those varied epistemological beliefs (Leung, 2013). Grading practices are also strongly correlated to teachers’ philosophies

(Kunnath, 2017; McMillan & Nash, 2000). Sometimes teachers explicitly verbalize those beliefs, and sometimes teachers' values must be inferred from their teaching practices (Leung, 2013). Because of the power that teachers wield over students' grades in the classroom, teacher assessment constructs and criteria need to be made explicit (Leung, 2004). This research seeks to answer one of Richard Stiggins' (1991) questions, "What does this assessment tell students about the achievement outcomes we value?" (p. 535).

The research questions of this study are the following:

- How do teachers talk about their grading and assessment practices?
- How do they explain their rationale for the grade weighting on their report cards?
- What do teachers' actual assessment and grade weighting practices reveal about their beliefs and priorities about language learning and student success?

## Local Educational Context

This study took place in a large, diverse, suburban district in the southeastern part of the United States. Seven world language teacher volunteers from one public high school were invited to participate. All the participants had at least ten years of teaching experience. Each teacher represented one of five of the six different languages taught at the school, and all but two teachers were born and raised elsewhere besides the continental United States. Five of these seven teachers have been working together in the same department for at least ten years, and have built a rapport of trust and support as a professional learning community.

The expectation at this school was that all the teachers in each professional learning community include the same categories and category weighting on common class syllabi, in an attempt to provide fairness to the students who land in one teacher's class as opposed to another by way of random scheduling. This school used the term "teacher lottery" to describe this

phenomenon, where one student might land in the "easy" Spanish 1 teacher's class, as opposed to the dreaded "hard" Spanish 1 teacher's class.

These teachers were asked to open up their gradebooks to better understand the kinds of assessments, categories, and weighting that the individual teachers used, and to gain more insight about these teachers' beliefs about learning languages. This research was intended to benefit both the participants themselves, by allowing them to better understand their assessment and grading practices, and also to add to the body of research about world language grading practices and beliefs.

## Framing the Problem

The ever-changing, complicated field of foreign language grading and assessment points to what Bakhtin (1981) described as a contact zone, where teachers struggle internally as they accept, assimilate, or reject the disparate internally persuasive and authoritative discourses in their field. Authoritative discourses embody "the authoritativeness of tradition, of generally acknowledged truths, of the official line and other similar authorities" (p. 344). Internally persuasive discourse is "gradually and slowly wrought out of others' words that have been acknowledged and assimilated, and the boundaries between the two are at first scarcely perceptible" (p. 345). These two discourses mix and mingle in the individual's ideological consciousness in an intense struggle among "various available verbal and ideological points of view, approaches, directions and values" (p. 346). This struggle occurs in the zone of contact. In this zone of contact, the authoritative discourses are the expert grading and assessment recommendations and top-down mandates from national, state, district, and local institutions. The internally persuasive discourses about grading and assessment include the teachers' own previous experiences both as students and teachers, as well as the internally persuasive

discourses coming from multiple sources like other colleagues in on-line and in-person professional learning communities, and from other stakeholders like parents and students.

According to Bakhtin (1981), authoritative discourse carries historicity, tradition and power already fused to itself. It is the word of the fathers, the teachers, the experts, or the moral, political, or religious leaders. It is connected to the past and exists in a hierarchically higher, distant zone. It carries more weight than internally persuasive discourse. It is the official line from the authority and is generally acknowledged to be the truth. It may be upheld by other supporting discourses, but it remains separate from those discourses, demarcated by distance and demanding allegiance or rejection. It is not flexible and is not subject to being creatively stylized. It comes as a whole and is not divisible into smaller units that can be partially accepted or rejected. The authoritative discourses related to language assessment have evolved dramatically over time, sometimes in contradictory directions, as newer experts emerge onto the academic scene of language acquisition theory, pushing older mandates and ideas to the sides. Authoritative discourses regarding grading have remained more consistent over the years. For example, experts have long insisted that grading should only reflect student achievement, rather than effort, motivation, or ability (Bonner & Chen, 2019; Stiggins et al., 1989).

On the other hand, internally persuasive discourse is highly individualized, from one person to the other (Bakhtin, 1981). This discourse half-belongs to the individual, and half-belongs to the original source. It is others' discourse that is creatively assimilated, re-organized, and freely developed within the individual to adapt to new materials and context. It is not isolated or static. It is an awakening of new and independent discourse, while maintaining aspects of the original source's discourse, whether it was authoritative discourse, or another's internally persuasive discourse. World language teachers base their assessment and grading

practices on these varied internally persuasive discourses that they have assimilated into their individualized contexts. A common internally persuasive discourse that continues to defy the dominant authoritative discourse about grading is the inclusion of effort as a prominent factor in calculating grades (e.g., Frary et al., 1993; Guskey & Link, 2019; McMillan & Nash, 2000; Nowruzi, 2021; Randall & Engelhard, 2009; Stiggins et al., 1989; Sun & Cheng, 2014; Tierney et al., 2011).

The zone of contact is where authoritative discourse and various internally persuasive discourses collide (Bakhtin, 1981). There is a presumed amount of distance between and among internally persuasive discourses, just as there is hierarchical distance between internally persuasive discourses and authoritative discourses. This is where the internal struggle occurs among internally persuasive discourses, where context, influence, distance, and questioning interact to give birth to new insights and a rebirthing, reframing, and reclaiming of another's internally persuasive discourses as one's own. These discourses are not reproduced, rather they are developed in "a new context under new conditions" (Bakhtin, 1981, p. 347). This struggle is essential for an individual's coming into ideological consciousness, by "selectively assimilating the words of others" (Bakhtin, 1981, p. 341). World language teachers stay in this zone of contact as they are bombarded by authoritative discourse from real or presumed experts in the form of professional development or in their studies toward advanced degrees, and by others' internally persuasive discourses in professional learning communities in their individual contexts.

## Situating Myself in the Problem of Practice

As a world language teacher of French and Spanish, I also exist in this zone of contact. I started teaching on a provisional certificate in 1997, without an education degree or any

pedagogical training. For over twenty years in the profession, I only ever took a handful of courses on pedagogy. So my grading and assessment beliefs and practices have evolved over the years through trial and error as I have been in the trenches of the classroom. I have heard some of the authoritative discourses about grading and assessment in professional development, but most of my internally persuasive discourse on the subject has been adapted by borrowing, trying on, and adapting or rejecting my colleagues' ideas over the years. Some of the expert recommendations that I found in my literature review in the next chapter were new to me, even though I have many years of experience. This research has already helped me be more reflexive about my own grading and assessment practices, by uncovering some of my beliefs and priorities about student achievement. I hope the findings of this study can encourage this same reflexivity for the participants and my readers.

I accidentally fell into teaching after college despite having always stated that I would NEVER be a teacher. I majored in Spanish language and culture in college and minored in French language and culture. My original goal was to use my languages in some kind of international business context. When my life path set limitations to my original goals, I took a teaching opportunity that was literally dropped into my lap. Language teachers were in high demand in my state at the time, and I applied for a provisional certificate, with the expectation to complete the streamlined teacher education program requirements within five years. The program allowed me to substitute traditional student teaching by using my current teaching position as an internship. I also took two teaching methods courses, one second language acquisition theory course, and one special education course. I completed the program requirements and received my renewable certificate by 1999. A few years later, I pursued a master's degree in Spanish literature.

I present all that long personal history to highlight the lack of formal education training that characterized my career. When I started my current Ed.D. program in 2020, I had extremely limited knowledge of the grand learning theories, despite having over twenty years of experience. I had never heard of (or did not remember) Vygotsky or sociocultural theory or any of the other language acquisition theories or traditional methodologies. I had hit-or-miss professional development classes throughout my two decades of teaching and attended only two professional conferences early in my career, and that was only because my district at the time had extra funds to send the world language teachers and pay for all the conference expenses. The state's professional development requirement for certificate renewal of ten hours every five years became embedded into the schoolwide professional development program so that teachers would not have to seek continuing education credits at their own cost. Most of this professional development was geared towards core academic subjects, with very little training specifically targeted for world language teachers.

Therefore, my own understanding of language assessment and grading was obtained on the job, through doing what everyone else was doing, with no particular rhyme or reason besides what made sense to me personally from my own experiences in my own context, and what my local school required for compliance to the current and ever-changing district expectations. My classroom contact zones varied throughout my career as I taught elementary, middle, and high school, both private and public, and also Pre-K, university, and adult education. The internally persuasive discourses of my peers in each of these contexts influenced my own beliefs about language learning, assessment, and grading.

I also share my personal history to highlight the time period of my career and professional development. The bulk of my limited pedagogical training occurred before the year

2000. The ACTFL *Standards for Foreign Language Learning: Preparing for the 21st Century* (The National Standards Collaborative Board, 2015) were originally published in 1996, the year before I began my teaching career. The standards were later revised in 1998 and 2006. The *ACTFL Proficiency Guidelines* were first published in 1986, right before my high school language learning experience, and they were revised in 1999, 2001, 2012, and 2024 (ACTFL, 2025a). The Integrated Performance Assessment was piloted and studied from 1997-2000 (Adair-Hauck et al., 2006). All of these events demonstrate the monumental changes occurring in the authoritative discourses in the contact zone of language teaching during the past quarter of a century as I was developing my own internally persuasive discourses about assessment and grading.

These philosophical changes at the national level took time to trickle down to the everyday world language teacher who was likely educated according to the old-school philosophies and methods. Throughout my career, I have had to adapt to new ways of teaching and assessing language, to stay current with the national standards, even as the standards themselves have been simultaneously undergoing change and development. As a fellow language teacher and assessor, I understand the confusion that this constant change can place on the public school teacher who is trying to survive the everyday context of teaching and assessing in the classroom, while processing the continual transformation of the teaching profession. In my research process, I believe this insight helped me listen closely and with empathy to what these world language teachers were saying about their assessing and grading practices to uncover new understandings about their beliefs.

## Key Concepts and Definitions

This study investigated world language teachers' beliefs and practices in assessment and grading. "Language assessment is a broad term referring to a systematic procedure for eliciting test and nontest data (e.g., a teacher checklist of student performance) for the purpose of making inferences or claims about certain language-related characteristics of an individual" (Purpura, 2016, p. 191). Language assessment constructs form the basis of assessments, and come from "research, theory, standards, accumulated experience, or principled practice" (p. 191). Assessment constructs are the "what" of the assessment, not the "how." Language assessment constructs demonstrate students' skills and abilities in isolated or integrated aspects of language and culture.

This study examined classroom-based assessments, rather than large-scale, nationally normed language proficiency assessments. Hill and McNamara (2011) defined classroom-based assessments as "any reflection by teachers (and/or learners) on the qualities of a learner's (or group of learners') work and the use of that information by teachers (and/or learners) for teaching, learning (feedback), reporting, management or socialization purposes" (p. 396). Classroom-based assessments, including language assessments, can be tools used by teachers to drive future instruction (formative), or they can be used to evaluate a student's learning (summative). Classroom-based language assessments can be traditional, using "multiple-choice questions (MCQs), true/false tests, short answer questions, and essays" (Tian et al., 2023, p. 3) or alternative, such as "storytelling, journals, reflection, portfolios, and teacher observations" (p. 3).

Grades are "the symbols assigned to individual pieces of student work or to composite measures of student performance created for report cards and other summary documents" (Guskey & Brookhart, 2019, p. 215). Grading constructs are meant to measure student

achievement, but other elements like effort, motivation, personality, attitude, and interest have also been found to be part of grading constructs (Stiggins, 1989). Grades are meant to provide feedback for students and parents, but are often used to rank, select, and categorize students (Brookhart, 1993).

## Organization of the Research Study

This research study collected and analyzed interviews with seven world language teachers working at one southeastern U.S. suburban public high school. Prior to the interviews, consenting participants were asked to provide a report that listed their grading categories and category weighting. They were also asked to provide a semester-long "Course Content" report from the online grading system that the district used. This report showed the names of all the assessments that appeared in student gradebooks, along with their point values and categories, without showing any student data. In preparation for the interviews, a preliminary analysis of the Course Content report was conducted to calculate the percentage weight of each of the individual assessments by category.

During the interviews, which were conducted and recorded over Zoom, I first asked about the participant's overall beliefs about what high, average, and low student achievement looked like in a world language course and how the teachers determined each student's achievement level. Secondly, the participants were asked to explain their overall grading categories and weighting, and the reasoning behind choosing their system for calculating students' grades. Finally, I shared my screen to go through the participants' Course Content report as a guide for my questions. The participants were asked about each assessment item in the Course Content report—what the assessment measured, what category it fit into, and why they decided to include it in the calculation of the students' overall grade.

After each interview, I re-watched the video recording, while comparing the preliminary transcription that Zoom provided to the participants' words. I corrected wording and punctuation to create an accurate transcript. I then extracted my words, except where necessary to preserve meaning, from the transcripts to convert the dialogue into monologue form. From the monologues, I began creating initial codes, which were then organized into categories and later collapsed into themes regarding what, how, and why the teachers used assessment practices. These themes were organized into findings using the theoretical framework of Bakhtin's (1981) authoritative discourses, internally persuasive discourses, and contact zone. To increase the validity of my findings, I used my participants' words whenever possible, to make their voices heard. I also conducted member checks by sharing the results of my findings via email to give my participants the opportunity to provide feedback and further insights.

I organized my dissertation into seven chapters. This introductory chapter is followed by the traditional literature review in Chapter Two and a deeper look into my methodology in Chapter Three. Chapter Four examines the contextual factors of the teachers' contact zone that helped shape their assessment and grading practices. Chapter Five reveals the teachers' internally persuasive discourses about assessment and grading based on how they talked about their practices. Chapter Six discusses the possible authoritative discourses that the teachers adopted or modified to form their own internally persuasive discourses. Chapter Seven considers the implications of the findings, recommendations for further research, and limitations to this study.

# Chapter 2 Literature Review

## Introduction

Grades and assessments are an essential part of the educational system, and many stakeholders depend on the reliability of the data that they represent. Teachers have many options regarding the "what," the "why," and the "how" when assessing students and reporting these assessments in their gradebooks. Teachers make a wide variety of choices, based on their internally persuasive discourses about their assessment priorities, within the contact zones of their individual classroom contexts.

## Organization of the Literature Review

This literature review will examine the empirical research that shows the conflicts between the authoritative discourses and the internally persuasive discourses that world language teachers experience in their assessment and grading. Three key internal conflicts will be discussed—the "what," the "why," and the "how" of language assessment and grading, that is to say, the constructs, the purposes, and the methods. Afterwards, I will discuss the external factors in the contact zone where the authoritative and internally persuasive discourses interact. These factors influence the ways that language teachers adopt, reject, or adapt authoritative discourses and others' internally persuasive discourses to develop and evolve their own individual internally persuasive discourse. Assessment research studies will relate strictly to second language assessment, because of the nature of languages as a subject. However, studies about grading will come from the more general field of education, since there are fewer studies in the field of second language that address grading practices.

## "What": The Assessment Construct

According to the National Academies of Sciences, Engineering, and Medicine (2020), an assessment construct is an abstract, theoretical concept that describes and specifies "the knowledge, skills, abilities, and other characteristics" (p. 27) to be tested. The construct is formed by mixing "theory, research, and experience" (p. 27). Constructs in language assessment have changed immensely over the past 100 years, as language acquisition theory has evolved and language assessment has been researched.

### The Menu of Assessment Construct Options

Purpura (2016) described four historical approaches to constructs: trait-based, task-based, interactionist, and sociointeractional. Trait-based constructs assess discrete components of language proficiency, particularly in environments where simple-to-complex methodologies are used. Another trait-based model assesses grammatical, sociolinguistic, discourse, and strategic competencies. Task-based approaches to constructs assess performance. Strong tasks are defined as those that mimic real-life situations. Interactionist constructs approach assessments with goal-oriented, relevant tasks to activate the students' knowledge of the language, along with the cognitive processes needed to perform the tasks. In task-based approaches, language knowledge and strategies are both assessed in the context of the task. Sociointeractional approaches assess interactional competencies, which are the product of both interlocutors in a local, moment-by-moment, contextualized event. Current language assessment textbooks (e.g., Brown & Abeywickrama, 2019) recommend language assessment constructs that are integrative, communicative, alternative, authentic, performance-based, or task-based, as opposed to the original behaviorist, discrete-point, trait-based constructs, such as translations, definitions of vocabulary, and multiple choice.

## Making Choices from the Menu of Construct Options

However, when classroom teachers use published tests and quizzes from the materials provided by their institutions, they are not thinking of the constructs that are being assessed. A study in Chile showed that even language teacher-educators were unsure about the constructs that they were assessing, with instances of construct-irrelevance and construct-underrepresentation (Villa-Larenas & Braunfot, 2022). Sometimes the constructs used by the teacher-educators in the study were context-driven, depending on their institution, particularly if the institution had a strong grammar approach to language learning. Arias and Maturana (2005) found no consensus among the teacher-educators at the same institutions about the constructs of their assessments or their grades. This confusion about what to assess would likely be passed down to the new classroom language teachers who were coming into contact with the authoritative discourses about assessment for the first time in their pre-service classes.

Some studies highlighted instances of language teachers using more communicative assessment constructs as the experts recommended (Arias & Maturana, 2005; Berry et al., 2019; Frodden et al., 2004; Giraldo, 2019). And many teachers believed that assessments should measure communicative abilities, even if they used traditional traits-based constructs in their classrooms (Díaz et al., 2012; Tsagari & Vogt, 2017). Arias & Maturana (2005) found that the teachers in their study used authentic assessment activities but found no clear construct for what was being measured, nor clear criteria for how the construct was to be measured. The researchers found that the teachers graded students arbitrarily and that instructional and assessment activities were often intermingled. Their interview participants talked about grading the four abilities, meaning speaking, writing, reading, and listening.

Most studies showed that classroom language teachers assessed trait-based constructs of grammar and vocabulary or the four skills of speaking, reading, writing, and listening (Davison, 2004; Díaz et al., 2012; Frodden et al., 2004; Giraldo, 2019; Vogt & Tsagari, 2014). Mede and Atay (2017) found that Turkish teachers were more comfortable assessing the microlinguistic aspects of language like vocabulary and grammar, and not as comfortable with assessing receptive or productive language or integrated skills. Microlinguistic aspects of language were simpler to assess and easier to evaluate and explain to the students and other stakeholders.

Differing contexts prioritized assessing different traits and skills. Young language learners were assessed mostly on their listening and speaking skills (Brumen & Cagran, 2011). But in other studies on primary and secondary learners, speaking was assessed the least (Tsagari & Vogt, 2017). Some studies found a prioritization on assessing the productive skills of writing and speaking (Vogt et al., 2020b), while others focused mainly on receptive skills of reading and listening (Tsagari & Vogt, 2017). Frodden et al. (2004) found that Colombian teachers used quizzes and drills to assess skills, aspects of language, or units of vocabulary in isolation, one at a time. But their tests assessed multiple differing skills and multiple aspects of language and multiple units of vocabulary. This study looked specifically at language assessment constructs. The results showed that using "hard" assessments, which were objective, product-oriented, summative, and quantitative, with strong practical and reliable qualities, were only able to demonstrate students' knowledge of vocabulary and grammar, and not their communicative competence. The teachers seldom used "soft," interactive, authentic, process-oriented, communicative, formative assessments, but when they did, the assessments lacked construct clarity or valid criteria for evaluation.

Writing assessments showed up in many studies, where teachers differed in their beliefs about whether content should be given more weight than grammar (Crusan et al., 2016). Davison (2004) found that in Hong Kong, teachers used a wide variety of constructs to evaluate writing ability, including grammar, content, organization, and flow of ideas. In speaking assessments, Giraldo (2019) found that Colombian teachers included non-linguistic constructs such as confidence, physical performance, and the students' design of their PowerPoint slides. However, the teachers in that study had strongly articulated reasoning for including those constructs in their context, explaining that these elements were part of communicative competence or "illocutionary abilities" (p. 46). Other assessment constructs that teachers used included the content and objectives of the courses, grammar, vocabulary, thematic knowledge, pronunciation, fluency, but also the ability to state a point of view and discuss cultural components (Arias & Maturana, 2005).

**The American Menu of Construct Options**

The majority of the studies mentioned above occurred in international contexts. English was not the only foreign language taught by the participants, but it dominated the studies about language assessment constructs. Language assessment studies in the United States typically focused on English for speakers of other languages (e.g., Jia et al., 2006; Llosa, 2011), assessments in dual language programs of English and another foreign language (e.g., Downer et al., 2012; Guzman-Orth et al., 2019), or assessments for heritage speakers of Spanish taking language courses to better their Spanish (e.g., Carreira, 2012; MacGregor-Mendoza, P., 2012). These studies were not included in this literature review, since they did not center on teachers who were administering assessments in a language that was not the dominant language of the culture where they were living and teaching. Freed et al. (2004) showed that achieving advanced

proficiency in a foreign language that was not dominant or readily available in the culture was much more difficult for students than being fully or partially immersed in the second language both in the classroom and in the world outside the classroom. This lack of foreign language assessment studies in an American context was striking, and likely related to the drop in interest in learning languages in the U.S. over the past twenty years (Devlin, 2015; Quinn, 2023; Stein-Smith, 2019). This current study was undertaken to fill the gap in literature on non-dominant language learning assessment constructs in an American context. U.S. world language teachers should not fall behind their international counterparts regarding classroom-based assessment, just because of lack of interest in the field. If anything, the U.S. needs speakers of other languages now more than ever in these changing political climates (Stein-Smith, 2019). Improving U.S. language programs by examining assessment constructs could be a step in the right direction towards keeping up with the rest of the world.

## "What": The Grading Construct

Timothy Quinn (2013) defined grades as "labels placed upon student work that are determined through a process of evaluation" (p. 4) or "a piece of information that attempts to report something about a student's education" (p. 5). He further explained that a grade could be a quantification of a student's learning, simplified into a symbol. It could be a description, an evaluation, or a ranking. It could measure a student's skill, or it could measure a student's performance, which might include complicating external factors, like mood, interest, health, time of day, distractions, etc. It could measure a student's skill level or performance on one particular task, or it could be a measure of performance on multiple tasks. It could be a frozen snapshot of student performance on one particular task or it could measure the student's progress in skill over time. A grade could measure a student's skill or performance as compared to a standard, as

compared to a particular starting point, or as compared to peers. Quinn emphasized that no matter how a grade was defined, its meaning should be communicated transparently to stakeholders. Zoeckler (2007) explained that the grading system represented a shared understanding between the teacher and the students.

**Battle Lines Drawn**

However, Cox (2011) presented qualitative data that showed there was a large range of individualized differences in teacher beliefs about the meanings of grades, what should be included in grades, and grading policies like common assessments, homework, late work, and 0's. When Stiggins et al. (1989) conducted their research comparing teachers' actual grading practices to experts' recommendations, they used current measurement textbooks that were used to educate pre-service teachers at that time. Those recommendations represented the authoritative discourse of that era. That discourse discouraged the incorporation of non-achievement elements like cognitive ability, attitude, motivation, effort, interest, and personality when determining student grades. Instead, academic experts at the time declared that grades should only reflect academic achievement. Their study showed that teachers only partially adopted this authoritative discourse as their internally persuasive discourse. They found that while all the teachers in the study believed that achievement should be the primary factor in determining grades, most of them also included effort as an important element.

**Let the Battle Begin**

Over the next thirty years, study after study corroborated the findings by Stiggins et al., with little change in belief or practice during this time despite the experts' continued push for grading on achievement only. A recent study (Mulyani, 2023) showed the same resistance to the authoritative discourse, while the experts continued to double and triple down on their

recommendations in measurement textbooks (e.g., Guskey & Brookhart, 2019). In the research, this noncompliance was almost universal, regardless of demographic factors. For example, Cheikem (2015) showed no differences in grading practices between male and female teachers or rural and urban teachers. Guskey & Link (2019) found that differences in grading practices were not related to years of classroom experience. Simon et al. (2010) discovered that pre-service teachers held similar beliefs to those of in-service teachers about the inclusion of noncognitive factors in determining grades. Research was conducted across many cultures with the same results, including the U.S. (Frary et al., 1993; Guskey & Link, 2019; Randall & Engelhard, 2009; Randall & Engelhard, 2010; Stiggins et al., 1989; Zoeckler, 2007), Canada (Duncan & Noonan, 2007; Tierney et al., 2011), China (Sun & Cheng, 2014), Colombia (Arias & Maturana, 2005), Iran (Mulyani et al., 2023; Nowruzi, 2021), and Nigeria (Chiekem, 2015). Two studies (Nowruzi, 2012; Sun & Cheng, 2014) showed that some cultures valued academic enabling elements like effort, improvement, ability, study habits, and participation, even more than achievement.

The only demographic differences in grading practices had to do with grade level and subject matter. Guskey & Link (2019) found that teachers across elementary, middle, and high school only weighted noncognitive factors between 10-20%. But early elementary teachers weighted effort less than late elementary, middle and high school teachers. Duncan and Noonan (2007) found that secondary math teachers included nonacademic factors in grading the least, with science teachers coming in a close second. Performing arts, fine arts, and religious studies teachers included nonacademic factors the most, with social studies and English teachers ranking in the middle.

Some of the reasons for the continuing conflicting beliefs about the inclusion of non-cognitive elements in grading related to the development of students' moral character, as revealed in a study by Mulyani (2023). In another recent study by Nowruzi (2021), effort, improvement, ability, participation, and behavior were considered academic enablers that could increase future learning and motivation. On the other hand, McMillan and Nash (2000) speculated that grading for effort might dilute student achievement, in that students might conflate working hard, and not ability, with achievement.

Across all the grading construct studies, American world language teachers were rarely included, with notable exceptions like Frary et al. (1993), Kunnath (2017), and Randall and Engelhard (2009). Even in these three studies, the percentage of world language teachers was much lower than teachers of math, science, language arts, and social studies. Duncan and Noonan's (2007) Canadian study of the differences in inclusion of nonacademic factors in grading across subject matter was significant in showing that a closer look at American world language classes was warranted in the body of research. Just like the Canadian study, in a world language classroom in the U.S., effort and participation and how they are graded might look different than in other class subjects like math or language arts. Or they might not be included at all in student report cards, as the authoritative discourse mandated. This study attempted to look at what a small group of world language teachers included in their grading constructs, and the rationale behind those inclusions. The studies showing differences in constructs and rationale across other subject matter and grade levels indicated that varied practices and beliefs might also be uncovered within the smaller context of an American world language department.

## "Why": The Purposes of Assessment

In current educational research, assessment has been categorized as one of three types (Schellekens, 2021). Assessment *of* learning (summative assessment) is an evaluation that occurs at the end of a unit of instruction that is used to determine students' knowledge and skills after instruction has been completed. Assessment *for* learning (formative assessment) is used during the course of instruction to provide feedback to both the student and the teacher about how much the student is learning, and to help guide the direction and pace of the teacher's future instruction. Assessment *as* learning (a subcategory of formative assessment) is used to provide students with feedback during the process of assessment so that they can self-reflect to guide their future learning.

### Alphabet Soup: A*o*L, A*f*L, A*a*L

The authoritative discourse coming from academic experts like Black and William (1998) saw using more formative assessments as a way to increase student academic self-esteem, raise student engagement in the assessment process, and improve the efficacy of teacher instruction. However, Frey and Schmitt (2010) found that teachers used formative assessments only 12% of the time and that three out of every four assessments were given after instruction was complete. Wang et al. (2020) learned from interviews that teachers perceived Assessment *for* Learning to be more important than Assessment *of* Learning, but their practice showed the opposite. Another study in 2020 by Widiastuti showed strong dissonance between teachers' beliefs about formative assessment and their implementation in practice. Teachers with more continuous professional development in formative assessment had more positive beliefs about its potential, but this belief did not translate to actual classroom practice. Cheng et al. (2004) found many differences across the contexts of Hong Kong, Canada, and Beijing related to student-centered purposes,

instructional purposes, and preparing their students for high stakes standardized tests. Deluca et al. (2019) found that pre-service teachers had mixed views on assessment purposes. The results of their questionnaire showed that 40.4% of teachers identified most closely with an assessment *for* learning approach, 27.3% viewed assessment *as* learning, 19.2% upheld an assessment *of* learning approach, and 13.1% formed a hybrid view of *for* learning and *as* learning, but not *of* learning. However, sometimes, neither the students nor the teachers recognized instances of assessment embedded into the teaching and learning processes because they were so closely tied together (Vogt et al., 2020b).

Still other teachers used classroom assessment as a means of encouraging further learning, rather than as a measurement of what had been learned (Nowruzi, 2021). Brumen & Cagran (2011) learned that teachers of young language learners used assessment to motivate, give feedback, observe student progress, promote future learning, and to encourage students to take language learning more seriously. The teachers in their study believed that without assessment, the students would not learn anything at all. These teachers of young learners also used assessments to think about and change their teaching practices and spark discussions with colleagues. Another study showed that pre-service teachers regarded assessment as a means of classroom management, student motivation, and social justice, rather than to support learning (Simon et al., 2010).

Some of the reasons behind all these differences in assessment purposes were national testing culture (Lam, 2015), lack of training in Assessment *for* Learning or Assessment *of* Learning (Lam, 2019), and teaching experience, student characteristics, school culture, training, and school logistics (Wang et al., 2020). Widiastuti (2020) found that beliefs could be changed by continuous professional development, but not practice.

These studies showed that many of these formative or summative uses of assessment related to the individual contexts of each classroom teacher. In this current study, these individual differences in the participants' training and teaching experience, along with the composition of the students sitting in their classrooms, could reveal similar differences in teacher implementation of formative and summative assessments.

## "Why": The Purposes of Grading

According to the authoritative discourse of the academic experts, grades should be used to show where students are on their learning journey, not to define them (Guskey & Brookhart, 2019). Grades should not be used to rank or select students (Quinn, 2013). Brookhart (1993) found that teachers knew that, unfortunately, grades would be used in the future to rank and select their students, despite expert recommendations. Teachers took this knowledge into consideration when developing their grading systems (Brookhart, 1993; McMillan & Nash, 2000; Sun & Cheng, 2014).

### Show Me the Money

Experts also agree that grades should not be used to motivate or control behavior, based on the psychological principles of intrinsic and external motivation (Quinn, 2013). Using grades as rewards and punishments to increase short term motivation to learn could backfire, as those external rewards and punishments have not been shown to increase intrinsic motivation in the long term. However, Brookhart (1993) found that teachers treated grades as an economic system, as academic payment for classroom work. Almost twenty years later, Simon et al. (2010) found that this view of grading as a currency of classroom management, student motivation, and social justice, remained prevalent in the pre-service teachers that they studied. Sun and Cheng (2014)

found that teachers assigned grades as rewards for student work, but that their grades also served as encouragement and reminders.

These studies showed that teachers never fully adopted the authoritative discourse regarding the purposes of grades. It's possible that the long history of paying students with grades for their work reached the "backfire" stage caused by constant extrinsic motivation that Quinn (2013) warned about, and students no longer derived pleasure and pride in their learning achievements. Perhaps both teachers and students became so accustomed to this capitalist vision of education that they no longer even wrestled internally with the authoritative discourse about the purposes of grading.

## "How": The Methods of Assessment

Expert recommendations for classroom-based language assessments have changed significantly over the past century. Teachers now encounter a myriad of ways to measure students' abilities. The authoritative discourse of the experts (e.g., Brown & Abeywickrama, 2019) now encourages the use of alternative assessments, rather than the traditional pen-and-paper tests of the past. Traditional assessments are typically pen-and-paper tests that incorporate multiple choice, true-false, short-answer questions, and essays, but they have been accused of being "indirect, decontextualized, product-oriented, and one-shot without constructive feedback" (Tian et al., 2023, p. 3). Alternative assessments include "storytelling, journals, reflection, portfolios, and teacher observations," (Tian et al., 2023, p. 3) as well as games, projects, singing, plays, self-assessment, peer-assessment, and projects (Tsagari & Vogt, 2017). Alternative assessments have been proposed as a way to increase student agency and motivation, to incorporate assessment into the instructional process, and to encourage students' synthesis of knowledge, rather than regurgitation of memorized facts (Mede & Atay, 2017).

**There is No School Like the "Old School"**

Despite the push towards more alternative assessments, traditional pen and paper tests seemed to remain the most common way to measure language students' abilities in empirical studies that examined teacher practice. Tian et al. (2023) found traditional pen and paper tests to be widely used by current teachers, regardless of years of teaching experience. A study of teachers of young learners also indicated that the preference for traditional testing held true irrespective of grade level (Brumen and Cagran, 2011). This tendency was found across cultures as well, in studies involving teachers in Canada, Hong Kong, and Beijing (Rogers et al., 2004); Colombia (Muñoz et al., 2012); Croatia, the Czech Republic, and Slovenia (Brumen & Cagran, 2011); Cyprus, Germany, Greece, and Hungary (Tsagari & Vogt, 2017; Vogt et al., 2020); Hong Kong (Lam, 2019); and Vietnam (Tian et al., 2023).

Some studies did show evidence of adaptation of the experts' recommendations for alternative assessment. In a survey of 1,788 learners and 658 teachers of English in Cyprus, Germany, Greece and Hungary, the Greek learners reported more diverse communicative assessment methods, in comparison to the Greek teachers or the other countries included in the survey (Vogt et al., 2020b). Two studies in South America, one in Colombia (Giraldo, 2019) and one in Chile (Villa-Larenas & Braunfot, 2022), showed that teachers used a variety of assessment methods, both traditional and alternative. A Chinese study (Sun & Cheng, 2014) reported varied assessment practices, including performance-based and project-based assessments, along with traditional tests. In a study of 702 tertiary teachers from 41 countries across five continents, Crusan et al. (2016) found that 62% of their participants had used portfolios in their assessment practices.

**It's a Great Idea, But . . .**

Some studies indicated a conflict between teachers' beliefs about alternative and traditional assessment. Teachers in Canada, Hong Kong and Beijing demonstrated this dissonance in their opinions that assessments simulating real life experiences were better than traditional tests, but that traditional tests were more likely to show students' actual learning (Rogers et al., 2007). A study by Berry et al. (2019) in the UK, Spain, and France, indicated that teachers used a variety of excellent assessments, but considered them to be part of good teaching, rather than actual assessments. While those teachers had positive beliefs about assessment activities incorporated into good instructional practices, they generally had negative beliefs about the term "assessment." A study in Hong Kong (Lam, 2019) showed that teachers had positive conceptions of alternative writing assessments, although many were skeptical about logistical implementation. And teachers of English in Germany, Cyprus, and Greece used few forms of alternative assessment despite holding positive beliefs about them, with isolated examples of games, projects, singing, plays, and portfolios (Tsagari & Vogt, 2017).

These differences between assessment belief and practice demonstrated the interplay between the authoritative discourse of the academic experts and the internally persuasive discourses of the individual language teachers across many cultures in their zones of contact. Several studies explored the reasons why teachers chose to use traditional assessment methods instead of alternative methods, despite having positive beliefs about the more modern methods. In Vietnam, the conflict zone for the teachers was related to their tightly held belief system, Confucianism, which valued a hierarchical relationship between teachers and students (Tian et al., 2023). This belief dissuaded them from using alternative assessments, where the power of assessment would have been shared by both teachers and students. Other inhibiting factors in this

culture were large class sizes, workload demands on teachers, lack of teacher agency, and low salaries.

**It's a Juggling Act**

Educational culture also exerted influence on teachers grappling with the decision to use alternative assessment. In a cross-cultural study of assessment practices in Hong Kong, Canada, and Beijing, teachers in Canada and Beijing reported fewer alternative assessments due to their focus on preparing their students to pass high stakes English tests for college entrance (Cheng et al., 2004). Local school assessment culture also had an effect on teachers' assessment methodologies, as found in a study by Vogt et al. (2000). Other institutional barriers, such as administrative duties, teaching duties, time spent grading, and curriculum restrictions deterred teachers from attempting alternative assessment (Lam, 2019). Rogers et al. (2004) speculated that alternative assessments required more time to grade and were more subject to students questioning the grade given, which caused teachers to hesitate in their adoption.

Lack of training in alternative assessment was another factor that hindered full implementation in classrooms (Vogt & Tsagari, 2014; Vogt et al., 2020). Tsagari and Vogt (2017) found that teachers had been trained in microlinguistic assessment methodologies and needed more training using alternative assessments. In a study by Villa-Larenas and Braunfot (2022) teachers expressed a lack of knowledge in evaluating their language tests' validity or reliability.

Vogt et al. (2000) found that teachers' practices were materials-based. They used whatever materials were available in the form of published textbooks and tests. Villa-Larenas & Braunfot (2022) discovered that teachers borrowed from colleagues and used published tests as their assessments, even if they did not feel that the tests were a great fit for their context, due to

time constraints and lack of confidence in producing their own assessment materials. They also used the same kinds of assessments over and over, throughout the years, so as not to spend time reinventing the wheel each year.

These studies showed that teachers' assessment methods were very individualized and limited by the constraints of their environmental resources like materials, time, and training. Both external forces, like institutional culture, and internal factors, like beliefs about the effectiveness of certain types of assessments, collided in the contact zone where teachers meshed out their internally persuasive discourses about their methodologies. The teachers in this current study might have experienced similar environmental and external forces in their contact zone that influenced the internally persuasive discourses that they formed regarding world language assessment.

## "How": The Methods of Grading

In 2025, educators find themselves in the middle of another evolution when it comes to grading. Starting in the 1990's, standards-based grading has been proposed as a way forward to clarify students' grades as a reflection of their academic proficiency and achievement. Standards-based grading is a method of reporting students' achievement on specific content standards using scale descriptors like "modest, intermediate, proficient, or exemplary" (Welsh, 2019, p. 113), rather than traditional numerical or alphabetical reporting. This system has the potential to increase consistency, validity, and transparency. However, many questions and issues remain in implementing this ambitious system (Link & Guskey, 2022). And standards-based grading has not been widely implemented in the U.S. (Young, 2023). Therefore, for the purposes of this study, standards-based grading will be abandoned, and traditional grading recommendations and actual practices will be examined.

**Pick a Number, Any Number**

Ken O'Connor (2008) discussed fifteen fixes for broken grades in a publication by the Educational Testing Service (ETS) Assessment Training Institute, a traditional authority in education. He recommended that grades should be organized by learning goals, not by assessment category. However, Kunnath (2017) found that teachers used weighted categories, specifically for assessments, which they considered to be objective, and work, which included more subjective assessments. This was almost thirty years after Stiggins et al. (1989) discovered that none of the teachers that they surveyed understood the principles of grade weighting or organization in assigning report card grades. Another leading set of grading experts, Guskey and Brookhart (2019), recommended using only a few grade categories to reduce subjectivity. Laurence Zoeckler's (2007) study revealed an example of this subjectivity. The teachers interviewed reported that grade weights could be manipulated to offset a "hard" grade to keep failure rates in line with administrators' expectations.

Experts like Guskey and Brookhart (2019) suggested that multiple varied assessment methods added clarity and meaning to overall grades. In a study by Stiggins et al. (1989), teachers collected a sufficient quantity of grades, according to the then-recommended minimum of two or three per grading period. However, those teachers actually collected too many grades, and intermingled daily assignments with large assessments in their data collection, which did not follow the recommendations of the era. According to O'Connor (2008), these daily formative assessments were still not recommended to be included in the tabulation of the final grade.

Guskey and Brookhart (2019) recommended that grades should include information about student process, product, and progress. Following those best practices recommendations, Tierney et al. (2011) found that a majority of teachers practiced self-referenced grading criteria

by comparing students' improvement over time as part of their grading calculations. Sun & Cheng (2014) also found that sometimes grades were assigned based on student progress.

O'Connor (2008) argued against grading students by comparing them on a curve. This method had been criticized for decades, insomuch that Stiggins et al. only found one teacher that used this method back in 1989. However, Tierney et al. (2011) found that some teachers still graded on a bell curve in their practice, even though all the teachers in the study agreed that grades should be indicative of students' individual achievement of learning objectives, rather than their ranking when compared to a group.

O'Connor (2008) disapproved of using an average of scores to calculate a final grade. Instead of using strict calculation methods, he suggested using a central tendency of scores, where the teacher determined the final grade using good judgment based on the central tendency of the student, ignoring outlier scores. None of the studies mentioned using a tendency of scores, although Tierney et al. (2011) reported that teachers considered a variation of this concept when they struggled with whether or not they should drop inconsistent grades or lowest grades, or whether to include all assessments in the calculation.

A number of different grading systems were reported in empirical studies. Stiggins et al. (1989) found that half of the teachers used raw-point accumulation to calculate grades, with arbitrary letter grade cut-off points, which went against textbook recommendations at the time. Zoeckler (2007) found that teachers used a variety of grading methods, including points or weights or converting points to percentages. Tierney et al. (2011) reported that almost all teachers agreed that some assessments should be weighted more than others, but there was little consensus about whether to weight newest grades more or the same as older grades. Cheng and Wang (2007) found differences across cultures in how final grades were reported. In Canada,

teachers were more likely to use percentage scores. In Beijing, teachers used scores out of 100 based on the accumulation of students' points. Teachers in Hong Kong used letter grades derived from percentages.

In grading individual assignments, O'Connor (2008) also opposed using zeros as punishment for missing or incomplete work. However, Tierney et al. (2011) reported that teachers disagreed with their province's discouragement of using 0's for late work, stating that this was not preparing students for future academic careers or life in general. McMillan & Nash (2000) found that teachers used 0's to motivate students, while also allowing late work and second attempts. Some teachers left the 0's in the gradebook, especially if they had a lot of grades to average together, and some teachers refused to put 0's in the gradebook because of the detrimental effects to their grade.

O'Connor (2008) believed that reporting student achievement on individual assignments could be communicated most clearly by using a combination of percentage correct scores and rubric scores. Rubrics were used often by foreign language teachers (Cheng & Wang, 2007; Crusan et al., 2016; Villa-Larenas & Braunfot, 2022). However, teachers also used a mix of intuitive grading and a deficit approach in marking papers, focusing on errors (Tsagari & Vogt, 2017). Cheng & Wang (2007) found that English teachers in Beijing primarily used holistic scoring on individual assignments, and rarely used rubric scoring. Teachers in Canada and Hong Kong used a combination of analytical and rubric scoring. The teachers in this study based their scoring decisions on what the assessment was, how it was to be used as part of instruction, the validity and reliability of the assessment, and the external context of the assessment. Wang et al. (2020) found that Chinese English teachers believed that written comments were more important than scores, but that the students often ignored the written comments and only perceived the

scores as actual feedback. Brumen and Cagran (2011) reported that teachers of young language learners in Eastern Europe believed that students should be graded with descriptions rather than numerical grades.

**The Right Way, the Wrong Way, or My Way**

Interestingly, after thirty years since the landmark study of teachers' grading practices by Stiggens et al. (1989), Thomas Guskey (2020) stated that *"There are no best practices in grading and reporting, and there is no perfect grading and reporting system* [author's emphasis]" (p. 220). Zoeckler (2007) explained that teachers tailored their grades in very individual ways, to express which lessons they considered more important than others, even as they seemed to be uniform. He also did not believe that the complex moral decision-making and weighting process of determining grades on a daily basis could be summarized in simple terms like teacher personality. Davison (2004) proposed that teachers had different assessment grading orientations, according to their beliefs. These grading orientations included the assessment technician—someone who ticked the boxes without personal judgements; the legalistic interpreter of the law, who was somewhat willing to make accommodations in a detached way; the arbiter of community values, whose more personal grading decisions were based on the future outcomes for the student; and the assessor of God, who used intuitive, inarticulate grading criteria.

Regardless of their grading orientation, most teachers agreed that they were able to justify their grade calculations to convey information both to their students and the parents (Tierney et al., 2011). McMillan and Nash (2000) revealed some of the reasons that teachers gave for their grading decision-making, which were the requirements of the learning objectives themselves, achieving fairness by using a variety of assessments to meet student needs, promoting the use of

constructed-response assessments, supporting their own personal views about homework, and learning from their own professional experience.

It was no wonder that teachers used "hodgepodge" grades as labeled by Brookhart (1991, p. 36), considering that the authoritative discourse of the grading experts acknowledged that there were some "wrong" practices, but none of the experts insisted on a "right" way to calculate students' grades. In these studies, the teachers' internally persuasive discourses adapted and assimilated through context and contact with other internally persuaded discourses became stronger than the authoritative discourse coming from the academic experts. McMillan and Nash (2000) found that grade calculation practices and beliefs were highly individualized and private, and that teachers, particularly veteran teachers, had a difficult time articulating their rationale. This current study attempted to create a safe space for individual teachers to express their beliefs about how they privately implemented grading in their practices.

## Factors in the Contact Zone

According to Bakhtin (1981), the zone of contact is where "discourse becomes more reified, more concrete, more filled with everyday elements" (p. 345). In each new changing context, novel discourses are encountered and in this new zone of contact, the individual must "wrest new answers from it, new insights into its meaning" (p. 346). The following section will look at some of the contextual factors in the contact zone that influenced teachers' assessment and grading beliefs and practices.

## The Contact Zone of Assessment

While teachers may have held certain internal beliefs about what, why, and how to assess languages, their everyday work constraints might have caused them to implement assessments in their practice that did not correspond to those internal beliefs. They might have subscribed to the

authoritative discourse on best assessment practices in theory, but the practical issues regarding their workload (Lam, 2019; Yan et al., 2018) as well as the quantity of their syllabus content (Wang et al., 2020) put pressure on their assessment practices. The time required to develop, administer, and evaluate performance assessments in the language classroom was mentioned by many teachers in empirical studies (Rogers et al., 2004; Wang et al., 2020). Class size was another mitigating factor in the language classroom (Cheng et al., 2004; Wang et al., 2020), as well as the availability of assessment materials (Cheng et al., 2004; Villa-Larenas & Braunfot, 2022; Vogt et al., 2000).

Another main influence on language assessment beliefs and practices was institutional culture. Villa-Larenas and Braunfot (2022) found that 50% of the teachers were conflicted between their personal beliefs about language assessment and their university's internal culture. Other studies uncovered specific institutional factors like exam culture, where instruction was driven by required high stakes language testing that affected both teachers and students (Cheng et al., 2004; Wang et al., 2020; Vogt et al., 2000; Yan et al., 2018). The stakeholder demands at the institution, from parents, students, principals, and the board of education, also caused dissonance between teachers' assessment beliefs and practices (Yan et al., 2018). The characteristics of the students at each institution factored heavily in assessment practices, depending on their language proficiency, motivation, and engagement (Wang et al., 2020).

According to these studies, context mattered when deciding what, how, and why to assess the students. The teachers in this current study may have worked in the same institutional culture, but their class sizes, class levels, student engagement and proficiency levels, material resource availability, and workloads may have been very different. And their assessment beliefs and practices possibly reflected those differences.

## The Contact Zone of Grading

### Beliefs About Grading

Part of the struggle with accepting the authoritative discourse about grading was that teachers already had well-established internally persuasive discourses that they implemented in their practice. Kunnath (2017) found grading practices to be strongly correlated with teacher philosophy, which was broken down into four components: "desire to promote student understanding, student motivation and engagement, desire for student success, and desire to accommodate student differences" (p. 81). McMillan and Nash (2000) included this desire for student success in their five central principles of teachers' grading beliefs, but they called it "pulling for students" (p. 10). Brookhart (1993) found that teachers believed themselves to be advocates for the student, and that they were influential in the development of students' self-esteem. For this reason, teachers considered the consequences of their grades on the individual student, and also on different groups of students (Sun & Cheng, 2014).

Teachers strongly believed that grading should be fair, according to multiple studies (Brookhart, 1993; Cox, 2011; McMillan, 2003; Simon et al., 2010; Sun & Cheng, 2014; Tierney et al., 2011; Wyatt-Smith, 2010). Nowruzi (2021) found that Iranian teachers' flexibility in grading was an attempt to balance internal and external factors in their grading, in an attempt to be "fair." However, they postulated that this inconsistency in student treatment actually made the grades less valid and less fair, even when applied in the spirit of fairness.

### What Will People Say About Me?

Teachers in the grading zone of contact were bombarded with expert advice from academic scholars, but they were also surrounded by other authoritative discourses like school administrators and district grading policies (McMillan, 2003; McMillan & Nash, 2000). Teachers

felt the external pressure of administrators' negative perception of low grades (Kunnath 2017; Nowruzi, 2021). This balance between rigorous grading and optimal student promotion caused dissonance for teachers. Xu and Liu's (2009) narrative research illustrated the power-laden authoritative discourse from a superior pressuring a teacher to make grade changes in their teaching practice. These administrators sometimes coerced teachers because school districts often had a covert "acceptable" number of failures per class (Zoeckler, 2007).

Teachers also struggled in the zone of contact because of other teachers' negative perceptions of overly high grades or overly low grades (Davison, 2004; Kunnath, 2017). They did not want their colleagues to accuse them of not being demanding enough or trying to unfairly curry favor with administrators by producing artificially high-achieving students. Some teachers were unsure if the grades they gave were fair and consistent with their colleagues, and were concerned about losing respect, especially in cultures like Hong Kong (Davison, 2004).

**Classroom Realities**

The physical context where the teachers grappled with contrasting discourses was the classroom. The students themselves and their parents put pressure on teachers when implementing grading policies (McMillan, 2003; McMillan & Nash, 2000; Nowruzi, 2021). Other classroom realities, like social promotion, absenteeism, disruptive behavior, and heterogeneity, affected teacher decision making (McMillan, 2003). Even pre-service teachers recognized these factors before they reached their own classrooms and prioritized the potential feasibility of grading policies over the theoretical aims of producing valid data (Simon et al., 2010). Teachers took class level into consideration in their grading procedures, whether the course was on-level, honors, advanced placement, or remedial (Kunnath, 2017). They also factored in grade level and class size when deciding on grading policies (Sun & Cheng, 2014).

Internal beliefs, external perceptions, and the logistical complications of the classroom appeared to influence which grading policies teachers implemented. This current study explored how those factors, and perhaps others, impacted how this world language department organized and maintained their gradebooks.

## Conclusion

This literature review looked at what world teachers assessed (the construct), why teachers assessed (the purpose) and how teachers assessed (the method). The wide variety of beliefs and practices across the three themes indicated that language teachers struggled to make decisions among the myriad of choices available. This became complicated by the expert recommendations that often contradicted the beliefs and practices of their fellow classroom language teachers. Authoritative discourses and the internally persuasive discourses of their teaching peers collided in differing cultural contexts in each zone of contact.

This literature review explored the same "what," "why," and "how" questions with regards to grade reporting. However, most of the grading studies focused on the core subjects of math, language arts, history, and science, with world languages either ignored completely or marginally included. Teachers' beliefs and practices for reporting world language grades might have been different than reporting grades for core subjects, but there were few studies to confirm or deny this.

### Mind the Gaps

The contact zones studied in this literature review about assessment were primarily international locations where English was the foreign language being assessed. Little attention has been paid to world language teachers working and living in the United States, where learning a foreign language is not considered an essential skill necessary to succeed in career or life. I

consider this to be a glaring gap in the research on language assessment. This current study attempted to begin to fill this gap in world language teachers' assessment practices and beliefs in the American context. Likewise, grading studies in the U.S. mostly ignored world language teachers, a gap which this study also sought to address.

This review of literature revealed another gap in the body of research about language assessment and grading, namely in research methodology. All of the empirical studies about assessment, except for two, collected data mainly either through surveys, questionnaires, or interviews. Arias and Maturana (2005) and Frodden et al. (2004) required their participants to also submit their actual assessments, complete with rubrics, assessment guidelines, and any self- or peer-evaluation scales. While interviews, surveys, and questionnaires were useful in collecting data, hard copies of classroom documents might have provided more clarity and might have remediated any issues that interviewees may have had with memory or transparency. This current study did not require participants to submit actual assessment documents, but all the titles of the assessments recorded in their gradebooks were included in the Course Content report that they were asked to provide. Those titles were used as an interview artifact to guide the semi-structured interviews, to jog the memories of the participants and to guard against any withholding or embellishment of data.

In the grading research, the same gap in data collection methodology was uncovered. Again, most studies used surveys, questionnaires, or interviews to ask teachers to self-report their grading practices and beliefs. This could mean that they made unintentional mistakes in how they reported their true practice, by using memory recall alone to discuss their grading. Stiggins et al. (1989) and Zoeckler (2007) were the only two studies that also examined the teachers' gradebooks along with their interviews. In this study, I used the gradebook shell (the Course

Content report), along with the Weighted Categories report, to compare how teachers talked about their grading beliefs and how they actually implemented those beliefs in their practice. These documents provided rich data for analysis when examined alongside their interviews.

This current study asked world language teachers to open up their gradebooks to show their category and assessment weighting practices and explain their decision-making process, to uncover their beliefs about what, why, and how they assessed and graded their students. The next chapter will look at how the study unfolded, who was interviewed, how the interviews were conducted, and how the interviews were analyzed and reported.

# Chapter 3 Methodology

## Statement of Purpose

Instruction, assessment, grades, and report cards are tightly bound in an educational system that puts teachers in the uniquely powerful position of evaluating students' achievement and relaying that information to societal stakeholders like the students themselves, their parents, future educators and educational institutions, and to the community workforce at large. However, teachers' assessment and grading practices are very individualized and often private. They are based on teachers' beliefs, training, experiences, and situations. World language teachers' grading and assessment practices are particularly varied due to the rapidly changing constructs of language as a school subject, the contradictory methodologies that have come in and out of fashion over the past century, and the multifaceted nature of language itself.

This study aimed to examine one world language department's grading and assessment practices to answer the following questions:

- How do world language teachers talk about their grading and assessment practices?
- How do they explain their rationale for their grade weighting on their report cards?
- What do world language teachers' actual assessment and grade weighting practices reveal about their beliefs and priorities about language learning and student success?

## Research Approach and Rationale

This qualitative study used interviews to explore world language teachers' rationale and beliefs about their assessment and grading practices. Qualitative research situates the researcher in the world of the research participants as she attempts to make that world visible (Denzin & Lincoln, 2017). These interviews allowed the teachers to speak in their own words about their

practices, as they explored their own issues (Carspecken, 1996). Using the teachers' gradebooks as an opening point of discussion generated explanations about their grading practices and their internal rationale for the way they assessed their students in their classrooms. As this research was underpinned by a hermeneutic or interpretive epistemology, these discussions were part of a double sense-making process, where both the teacher participants and the researcher were attempting to interpret the social world in the process of understanding it (Usher, 1996). The findings in this study were not presented as facts, but rather as the researcher's interpretations of the teachers' lived experiences.

## My Roles as a Teacher and Researcher in the School

I am an insider in this group of teachers; they are my colleagues, whom I greatly respect. But I am also an outsider as a researcher exploring their experiences, since I am not part of their classes. While I do not hold any position of leadership or authority in this group of participants or in the school in general, throughout my research process, I still had to be wary of the power and privilege that conducting research afforded the researcher (Parson, 2019). I tried to use the teachers' voices whenever possible, so that I would not silence or misrepresent them. This strategy also helped me resist any urge to include any assumptions I might have made about their assessment or grading practices because of my previous relationships with them. In my interviews, I attempted to talk less and listen more, avoiding leading questions, so that they were able to express their own thoughts without fulfilling what they felt that I as a researcher wanted to hear (Carspecken, 1996). Throughout the research process, I felt compelled to protect their confidentiality, by anonymizing their identities, never mentioning their specific languages, and avoiding the creation of a readily identifiable profile. As this was a very small study in a very

specific locale, these teachers were more vulnerable to scrutiny, and I did not want to cause them any harm.

## Research Context

This particular high school was a perfect setting for exploring world language teachers' grading practices, because of all the languages represented. The school is a magnet high school for international studies, where Spanish, French, Latin, Portuguese, Japanese, and American Sign Language are all offered to a student population of over 2500. Even though it is a magnet school for international studies, the school is still a typical suburban public high school. For each graduating class of over 600 students, only about 100 of those students are part of the magnet program. These students are bussed in from other parts of the district and go through a rigorous application process to be accepted into the program. The rest of the students live in the school district and follow the regular state academic pathway to graduation. Students in the international studies magnet program are required to take four language credits in high school. For the non-magnet students, the state does not require any language classes for graduation, but the state Board of Regents requires two language credits for admittance to four-year state colleges and universities. Students who do not plan to attend college are not required to take language courses. There is currently no End-of-Course Test that students must pass as part of the course credit, like there is for certain other core academic courses. The population in the language classes therefore consists of a mix of high-achieving magnet students with language course requirements, high-achieving students with aspirations for college, students who are curious about languages and choose them as electives, and students who have been arbitrarily thrown into the courses due to staffing issues in other departments.

## Participants

The world language teachers at this high school– Alicia, Carmen, Daniel, Diana, Julia, Laura, and Martin (all pseudonyms)– are an extremely diverse group of individuals. All of them except for two were born outside of the continental United States and are native speakers of the languages that they teach. All of them have over ten years of teaching experience, five have master's degrees, and one has a doctoral degree. All of them are at least in their forties, with an age range that extends well into the seventies. Two of them are not fully certified and have renewed their work permits for many years. Three of them received their undergraduate degrees outside the United States. Three of them learned English in their native countries in a school setting, and one learned American Sign Language as a child at a school for the deaf. Five of the teachers have also obtained a gifted endorsement for their certification to teach exceptional students, which is awarded after an intense year-long class. Four teachers in this group are "singletons," meaning that they are the only teacher of that language and cannot collaborate with anyone for planning and data sharing.

Because of their ages, these teachers, just like me, learned their second language during the time that the language teaching profession was undergoing massive changes due to the proliferation of approaches and methodologies, along with the introduction of the ACTFL proficiency standards in the 1980's. Their own personal experiences with language assessments as students would likely have been more traditional pen and paper exams, and this might have exerted some influence on their current assessment and grading practices as teachers (Oleson & Hora, 2014). For several of them, their language training experience was also influenced by the necessity to learn English as a dominant language to be successful in their careers. Their

motivation to learn English might have been stronger than American students learning a second, less dominant language.

## Data Sources

My research drew from three primary data sources– the interviews with the teachers, a report showing their grade category weights to show how their grades were calculated, and a Course Content report. Each teacher chose just one class to talk about, from their three possible classes, and they were allowed full freedom to choose what level, from Level One to Honors or Advanced Placement (AP). The teachers provided the two gradebook reports specific to their chosen class before the interviews were conducted. Neither report contained any student information, neither generalized nor individualized. The Course Content report was a shell gradebook that showed the names of course assessments and how grades were calculated for students across the semester, which was the equivalent to a full course at this high school. These reports distinguished this study from several other studies conducted about teachers' assessment and grading practices, by providing a means of triangulating the data across multiple sources, thereby increasing reliability (Creswell & Maxwell, 2000).

In the individual interviews, I asked the participants about their beliefs about the characteristics of high-, medium-, and low- achieving students, and how they distinguished each of those categories. This was an initial attempt to uncover their beliefs about student achievement in a world language classroom. Next I asked them general questions about their gradebooks and their rationale behind how they set up their categories, grading scales, and grading procedures. Finally, I went through their whole gradebook with them, line by line, discussing each assessment, what it measured, how it was administered, how it was graded, what category it fit into, how much it was weighted, and why.

During my research, I started a small research journal in a Word document, where I wrote memos to myself and initial profiles about the teachers. I detailed my process, asked questions, speculated privately, wrote to-do lists, and documented some of the origins of my themes. This journal became a secondary data source as I kept track of my progress and processed my thoughts, while also serving as an audit trail to "provide clear documentation of all research decisions and activities" (Creswell & Maxwell, 2000, p. 128). This journal became an indispensable resource for establishing validity when recalling and describing my data analysis process in the next few sections.

## Data Collection Procedures

In the fall of 2023, I received Institutional Review Board approval to begin collecting data. I received permission from my local site in January, 2024. I began recruiting the world language teachers at the school site with a prepared script which was delivered both in person and by email in early February 2024. Soon after, seven teachers signed informed consent forms to provide the reports and participate in the interviews. I scheduled the interviews during the spring of 2024 at the teachers' convenience. Each teacher submitted a Gradebook Category Weighting report and their Course Content Report from their gradebooks from the previous semester. These reports were stored in my password protected university Google drive.

I conducted and video recorded each individual interview with the teachers over Zoom, taking advantage of the transcription services available. During the interviews, I was in my private locked office, and my interviewees were in the private locations of their choosing. I used a semi-structured interview protocol, as recommended by Carspecken (1996), using lead-off questions as well as follow-up questions. The interviews ranged in length between one and two hours total for each teacher. The video recordings and the transcriptions were stored in my

password-protected university Google drive. The video recordings provided another means of establishing the credibility of my study, since I had an original record of the participants' words.

## Data Analysis Procedures

### Scratching the Surface

For the next stage of my research, I used thematic analysis, using Braun and Clark's (2006) six recursive phases. To familiarize myself with the data, I watched all seven videos all the way through alongside the transcription that Zoom provided, stopping the videos frequently to correct words, phrases, and punctuation, to create a workable document. I also changed my participants' names to pseudonyms to anonymize and protect them. This step helped me look at the big picture of what the participants were telling me and helped me identify sections that required closer examination.

To further familiarize myself with the data, I used the Category Grade Weighting and Course Content reports to create an Excel workbook to chart information about each of the teachers' assessments. I started writing memos in the chart to record my observations, making "notes about the notes" (Dey, 1993, p. 93). At the beginning of this process, I created a separate worksheet for each teacher, and created a chart with the title of each assessment, the grade category it belonged to, the date, the percentage of the overall grade, and a brief summary in my own words of the teachers' description of the assessment. In my summary I noted which standards or traits were assessed, such as listening, speaking, reading, or writing, and whether they were assessing grammar, vocabulary, culture, or more integrated skills. Using the interviews, I also recorded in the worksheet how the teacher graded each assessment, how it was administered, how feedback was offered to the students, and if any remediation, differentiation, or second chances were offered to the students. In the same workbook, I also created a summary

worksheet of all the teachers' categories with their grade weighting, to be able to have a top-down view of all the data (Erickson, 2004). For this summary worksheet, I used the double-entry notes strategy suggested by Sunstein and Chiseri-Strater (2012), writing my interpretations and reflections next to each summary of how each teacher used their categories.

For phase two of my analysis, which Braun and Clark (2006) called "generating initial codes" (p. 88), I converted the interview text from dialogue to monologue form, only preserving my words or questions if they were necessary to understand the participants' meaning. I created a double-columned Word chart, with an initial code on the left, and any relevant paragraphs of interview monologue next to it on the right. Some of the interview monologues related to multiple codes, so the same paragraph was copied and pasted in multiple sections of the initial code document. I decided to leave the paragraphs intact, without isolating smaller chunks of words or sentences, to maintain the unity and full meaning of the participants' ideas. As I worked through the codes, I began to suspect that using fragments of their thoughts might lead to decontextualization and misrepresentation (Maxwell & Miller, 2008).

As I methodically analyzed each teacher's interviews, I added more codes as applicable, revisiting previously analyzed interviews to add the newly identified codes and any pertinent excerpts from the teachers' interviews. By the last interview, I arrived at fifty-nine different codes, such as "Ideas about classwork," "Problem-based learning," "Use of translation," "Characteristics of low-achieving students," "Saving time," etc. Each teacher had their own Word document with these fifty-nine codes along with any relevant accompanying monologue text from their individual interview.

To ensure that I had not missed any data, I went through the interviews one last time, making a list in a new Word document of the teachers' grading and assessment practices, along

with their accompanying reasons for those practices, using their own words. While analyzing these reasons, I completed Braun and Clark's (2006) phase three. I searched for themes that related specifically to the factors in the teachers' contact zone that influenced their grading and assessment practices, through the lens of Bakhtin's authoritative and internally persuasive discourses. For phases four and five, I reviewed and organized these factors by macro, meso, and micro levels of contact zone. I proceeded straight to phase six to write Chapter Four, describing the findings about the teachers' contact zones.

When writing Chapter Four, and subsequently Chapter Five, I used the teachers' words when possible to provide thick, rich descriptions with details and examples to increase the credibility of my analysis (Denzin, 1989). Experienced researchers reported that: "People's spoken words sometimes showed the strength of their views or the depth of feelings or, on the other hand, their passivity and lack of engagement in ways that the researcher's own narrative could not," (Corden & Sainsbury, 2006, p. 13). Using the participants' words attributed value to their voices while simultaneously empowering them and lending credibility to my findings.

However, my participants' quotes presented an ethical dilemma as I prepared to report their experiences regarding assessment and grading. I wanted to maintain the authenticity of the words that my participants had spoken by presenting them verbatim as much as possible (Lingard, 2019). However, I found myself needing to do a "light tidying-up" (Corden & Sainsbury, 2006, p. 18), due to several ethical issues. First, since five of my participants were born outside of the continental U.S., their speech patterns reflected their varying accents that could possibly be detected when reading their quotes, which would compromise their anonymity. Second, experienced researchers reported that their research subjects balked at the way their normal conversational speech patterns were included in reports, because of the inclusion of

hesitations, fillers, repetitions, and stammers, and they felt that they were being disrespected (Corden & Sainsbury, 2006). Third, I did not want foreign or regional accents or disability-based speech abnormalities to distract my readers from the important ideas that my participants were expressing in their interviews. Therefore, where possible, I maintained their original quotations verbatim. When necessary, I went back to the video recordings to confirm their wording and meaning, in order to create a reliable account of their lived experiences. But, when necessary, for the reasons stated, I edited their repetitions, stammers, and fillers, making every effort to avoid changing their intended meanings.

**Digging Deeper**

After writing my findings about the contact zones in Chapter Four, I turned back to my data to investigate the teachers' internally persuasive discourses about assessment and grading. Conducting this thematic analysis proved to be more complicated and required that I utilize my Excel workbook again. To ensure that I did not misinterpret their beliefs or their practices, I copied the teachers' monologues about each assessment into my Excel worksheet about each assessment, pasting the teachers' own words into an additional column next to my own summary description column. I then calculated what percentage of the teachers' gradebook assessments corresponded to each of the skills of writing, speaking, reading, and listening. During this process, I realized that those skills should be broken down further into the corresponding ACTFL standards of interpersonal speaking and writing, presentational speaking and writing, interpretive listening and reading, cultural perspectives and products, connections, comparisons, and communities, for a more thorough examination of the teachers' practices and underlying beliefs. I created a word document to record these percentages for each teacher.

The findings revealed from these calculations spurred me on to calculate the percentages of other teacher practices related to my initial codes, such as "Grading on Accuracy," "Grading on Completion," "Use of Rubrics," "Open-Ended Assessments," "Closed-Note Assessments," etc. As I recorded these percentages by teacher and by category, I wrote memos of my interpretations of the percentages, making a note when I believed that my calculations were more approximate than exact.

At the same time, as many of these phases were overlapping and recursive, I began to see more patterns in the initial codes. To analyze my initial codes more closely, I created a new Word document and grouped the codes together by preliminary categories. In this new document, I transferred all the teachers' monologues from their separately coded Word documents to my initial codes in the summary document, now organized by category, so that I could compare a cross-section of the teachers' statements for each code and category. As themes about the teachers' beliefs about student achievement came more into focus from the cross-sectional analysis of the categories, I was then able to combine my gradebook calculations with the words of the teachers themselves to form a more complete picture of both their practices and the internally persuasive discourses behind those practices. These internally persuasive discourses became the topic of Chapter Five.

From these internally persuasive discourses, I began to see a kaleidoscope of grand learning theories, second language acquisition theories, and new and old methodologies and approaches to teaching languages, that were mixed and mingled to form the internally persuasive discourses of each teacher. I took some time off from analyzing my data to spend time researching these historical approaches, and hypothesized that these teachers' beliefs about student achievement were direct adoptions or modified adaptations of some of these approaches.

I identified these historical approaches and theories as the original authoritative discourses that the teachers had incorporated into their internally persuasive beliefs.

Because I lacked broad knowledge about the plethora of approaches and theories about second language learning, I employed AI to assist me to speed up my research. I typed the characteristics of each teacher's assessment profile into ChatGPT and asked it to identify potential teaching approaches. For each profile, it identified several possible approaches, which I used as a starting point for more in-depth research on each approach and whether it was a true match for the teacher profile. When I was satisfied that I had matched each teacher profile with one or more historical approaches to teaching language, I wrote Chapter Six to discuss the traces of the authoritative discourses that seemed to appear in the teachers' beliefs and practices.

## Establishing Quality and Credibility

In January 2025, I presented my chapters to the participants to review my findings as a means of member-checking for validity (Lincoln & Guba, 1985). As further means of establishing credibility and trustworthiness, as mentioned in the previous sections, I video recorded the interviews to have a permanent record of the participants' words, I used a research journal for an audit trail, I triangulated the teacher interviews with the Gradebook Category and Course Content reports, and I provided thick, rich description of the teachers' beliefs and practices using their own words when possible, using minimal tidying up of their quotations. Qualitative research does not adhere to the same rigorous standards for reliability, replicability, and generalizability as quantitative inquiry but I attempted to make my data collection and analysis as transparent as possible.

The next three chapters will reveal the factors in the teachers' contact zone that affected their assessment and grading practices (Chapter Four), the internally persuasive discourses that

they developed while practicing in that contact zone (Chapter Five), and the original authoritative discourses that may have influenced the development of those internally persuasive discourses (Chapter Six).

## Chapter 4 Contextualizing the Contact Zone

This chapter aims to contextualize the contact zone where the teachers in this study made decisions about grading and assessment. According to Bakhtin (1981), the contact zone is the battleground where authoritative discourses collide with internally persuasive discourses to be assimilated, rejected, or adapted by the individual as they develop their own ideological viewpoints and approaches. In this context, the authoritative discourses were the state, district, school, and department requirements and expectations regarding teachers' assessment and grading practices. The internally persuasive discourses were the multiple, ever-changing, independent evolutions of the teacher's grading and assessment beliefs and practices as they navigated the contact zone of their classroom context. This chapter will examine the authoritative discourses that influenced the teachers, along with the environmental factors in the contact zone that helped shape their internally persuasive discourses.

In their interviews, the teachers in this study—Alicia, Carmen, Daniel, Diana, Julia, Laura, and Martin (all pseudonyms)—discussed their gradebook categories and weighting, and their individual graded assessments over the course of a semester. They also spoke about their environmental context which provided the backdrop for the ways that they chose to assess, grade, and record their students' assessments. This context was examined at three levels that influenced the teachers' practices—the macro level of the state, district and school organization, the meso level of the classroom context, and the micro level of the individual students in the classroom.

## The Contact Zone at the Macro Level

### School Organization

Teachers must abide by, and work within, the expectations and the environment of their school and their district. At the macro level of their context, these world language teachers dealt with the pressures of graduation rates, limited time to instruct and assess due to block scheduling, and the constant worry of maintaining strong enrollment in their language classes to keep their jobs.

#### *Graduation Rates*

In the southeastern state where this study took place, high schools are rated based on many factors, one important factor being the graduation rate. Several years ago, this high school had a goal of "92 in '22" (92% graduation rate in 2022). The expectation after that year was that the graduation rate would continue to improve year over year. To accomplish this, administrators ran gradebook reports throughout the semester and sent the results out to department heads to speak with their teachers about too many failing students in a given class, and to ask how those teachers were working with students to get them to pass their classes. The teachers felt immense pressure to pass their students, which resulted in reflexivity when discussing their assessment categories and grading methods.

In this state, students must complete 23 credits to graduate high school. The world language department involved in the study is in a school that uses a four-by-four block schedule, where students complete four high school credits in their first semester, and four more in their second semester, for a total of eight credits each year. This gives students far more flexibility in choosing electives and pursuing interests like languages, fine arts, and career-tech options while

they are in high school. It also helps boost graduation rates, since students have 32 credit chances to pass 23 required credits during their four years.

However, this flexibility can backfire by producing students who, as Laura stated it, “play too much at our school” because “we’re giving them multiple chances of failing.” In her grading categories, she would have liked to weigh her “Classwork” category less, which was currently at 25%. She explained her potential category weight change by asking, “Why should I have a Classwork category when they should be doing that anyways? Does that make sense? Because we know they're doing it by the way they perform on the quizzes and the test. Period.” Her thinking was in line with experts’ recommendations (e.g., Bonner & Chen, 2019) that grading should reflect achievement rather than effort or academic enabling behaviors. However, she acknowledged “the political side about it,” stating that “what's gonna happen, we're gonna have a lot of students not graduating.” She knew that some of her students struggled in learning language, and she provided “wiggle room” for them in her grading categories, even though she felt conflicted between what she felt were the right ways to measure students’ learning, and the pressures from her school and state to pass students even if they were playing around.

### *Block Schedule*

Another issue with the block schedule at this school is the amount of instructional time that is lost compared to traditional six- or seven-period-day schedules. Spanish and French AP courses are offered for students who complete levels one through four, so teachers in those languages are expected to prepare students in the lower levels to be ready to tackle those challenging courses. However, block schedule only allows approximately 143 hours of instruction per semester course (95 minutes per day X 90 days), compared to 165 instructional

hours (55 minutes per day X 180 days) on a seven-period day or 180 hours on a six-period day (60 minutes per day X 180 days).

These teachers often felt like they ran out of time to assess students' abilities. In looking back at her previous semester's gradebook, Diana noticed a missing assessment, "I guess I didn't have time, like they also have to do the Red Riding Hood where they tell me the story. . . But I don't think I did, because it would say, 'Speaking Something Red Riding Hood.' I didn't get to it." Laura also discovered two missing assessments in her gradebook, "I usually have two more tests on there," saying, "For some reason, this semester, just too many tests, . . . too many EOCs. Everything kind of fell so fast, yeah."

This time crunch forced the teachers to carefully weigh their choices of assessments, particularly in how they designed their final exam, since their final semester grades were due within hours of administering their last final exam of the semester. Daniel mentioned that his final exams used to be more balanced and cumulative, incorporating writing, speaking, and listening/reading. He explained that he "used to do three final exams. But yeah, it's because after the Culture project, it takes time to finish all the projects, sometimes maybe three, four days. And then to have all the speaking, also." So instead he copied and pasted his previous test questions into a one hundred question multiple-choice scantron exam, explaining, "You know we won't have time to do anything. It's like, well, two hours. You have two hours. You finish at eleven. Two o'clock, you have to post grades."

Diana mentioned eliminating a reading comprehension section from her final exam, because in previous semesters the students "were spending too much time, and I was like, no, you know, it's the last day I just wanted that boom, boom, boom." To keep a balanced final, Diana did the speaking and writing sections of her final exam a few days before the official

exam, which was all in scan-tron multiple-choice format to increase speed in grading. But her grading methods for the speaking and writing sections, which she explained sheepishly were borrowed from a prior colleague, were also designed for speed. “If this is an A student, basically, you just give them an A. If this is a B student, you just give them a B, so it's basically based off what they have done in the class. So you don't sit down and listen to all of them and read.” She continued with an embarrassed laugh, “No, you just give whatever they have in the class. So you're like, Oh, yeah, you're a B student. Okay? So out of ten points, you get an eight. Yeah. And that's how I save time.” While on one hand, this practice might have seemed unfair to the students, these extra sections could also have helped cushion a student who did not perform well on the more objective reading and listening sections of the multiple-choice exam.

***Teacher Allotments***

Students are not required to pass any world language credits for graduation in this state. This means that world language teachers in this state feel constant pressure to maintain enrollment in their classes in order to keep their jobs. Some level of world language enrollment is protected by this state’s Board of Regents, which requires two sequential years of world language credits for admittance to any state university. And this school is a magnet school in the district, bussing in some of the top academic prospects who plan to attend in-state and out-of-state universities upon graduation. Because of an overall higher number of potential university candidates, this world language program is strong compared to other non-magnet schools in the district.

Even here, though, the world language faculty has seen the results of not reaching high enough enrollment numbers in particular languages. Over the years, the department has lost German, Chinese, and Arabic due to low numbers, and other languages are struggling to

maintain full-time status. The department loses teacher allotments with each language loss, with the allotments going to other departments where the needs are perceived as more urgent. As Laura explained, “We're in a situation with one of our colleagues. He’s having to do a whole year of ESL.” She recognized the danger of adjusting her Classwork category percentage to focus more on quizzes and tests, saying, “We would definitely lose enrollment, you know what I’m saying? We would lose enrollment because they're not gonna wanna go, you know, take a language.” She realized that taking away the “cushion” that classwork provided for her students’ grades could endanger her own employment.

## Grading and Assessment Expectations

Each school has a grading and assessment culture, as was seen in previous studies (Chen et al., 2004; Lam, 2015; Villa-Larenas & Braunfot, 2022; Vogt et al., 2000; Wang et al., 2020; Yan et al., 2018), and this school was no different. The teachers discussed school and district expectations about final exams, a new district-wide test, and common grading category weights.

### *District-wide World Language Assessment Initiative*

Some grading and assessment policies were determined at the district level. For example, Carmen explained, “The final exam category is a county-based decision.” This category had to be 10% of the students’ final grade. All the teachers in the study adhered to this directive and referenced the district policy in their interviews.

However, Laura was the only teacher who attempted to comply with another district-based initiative. In her gradebook she had a District Assessment category weighted at 0%, to report students’ grades on the pre-tests and post-tests that the district was pushing out “so we all can kind of be in sync.” However, she encountered multiple technological problems with the assessment, stating that after trying multiple times to get it to work, she determined that “it's not

ready yet." This testing program was required for many core subjects in the district, but the complex features such as the audio and video needed to assess the communicative aspects in language learning were not consistently functioning to make this a viable assessment program at the time of her interview. Despite this initial failure, Laura felt pressure from the district to attempt to administer two of the prescribed diagnostic pretests, but she never met with success in either implementation.

***Required Team Collaboration in Grading Categories***

This district expected all the teachers who were teaching the same language and level to collaborate to determine common category weighting. In several meetings, administrators mentioned the "triplets" metaphor, where a hypothetical parent had triplets who were all taking the same class and level, just with three different teachers. The administrators expected that those three students would receive the same opportunities and consequences in the assessment and grading process in all three teachers' classrooms. As Carmen described the category weighting decision process, "We're on a team, and so we discuss what percentage we would like for things to be. And it was a collaborative decision." However, in this world language department with six different languages, several of the members were "singletons," meaning that they were the only teacher of their language, and therefore had no one at the school to collaborate with. Often the department met as one large team, instead of dividing into languages, to share ideas and issues of a more general nature. Martin talked about these group discussions and how over the years his categories have varied a lot, explaining,

> When we've had discussions amongst ourselves, I listen to what you guys do. And naturally, very often, what you guys do makes sense. So I would try and bring it in, and invariably find I'm too lazy to do that. And so I drop it and then go back to the minimum number. So this is a reasonably simple way of doing it for me.

At the same time, Martin recognized that “everybody’s on about consistency, which is very important.” As a singleton, he was expressing the dynamic tension of adapting to team decisions, while also trying to make grading decisions that worked for him and the specific language teaching needs in his classroom.

## The Contact Zone at the Meso Level

While teachers understood the demands from their authorities at the state, district, and school levels, they also felt pressure from the classroom environment. These secondary factors influenced the teachers’ grading and assessment practices and beliefs by shaping the implementation of the requirements from their authorities. These factors included class sizes and makeup, and teacher assessment resources and training.

### Class Sizes

Several of the teachers mentioned the size of their classes as a limiting factor in the kinds of assessments that they could administer, particularly with regards to speaking assessments. As Laura put it, “With the speaking, when you have thirty-eight students, . . . that's a challenge.” For her final exam speaking presentation, she ended up having to divide the class, saying, “Some students have to do face-to-face, some students, I had them record just because in that class I had too many students.” But class size also limited how much feedback teachers could offer students on writing assessments. Diana pondered, “I just wish that I had more time to grade everything individually . . . for accuracy rather than just ‘Oh, yeah, it looks right.’ But . . . when you have ninety students, when am I gonna do this?” At the same time, she wondered if added feedback would actually make a difference to her students. She asked herself, “Would they care if I give them?” She said that even in her advanced level classes that some lackadaisical students would respond disinterestedly to her comments on their papers, “Oh, my feedback.” While others would

respond with more enthusiasm, “Oh, no! I want to do it better! I'm gonna rewrite it with the feedback that you gave me!” While not all the classes were as large as thirty-eight students, many of the teachers felt the strain of balancing effective, appropriate assessments with the limits of their instructional time due to class sizes.

**Combined Classes**

Several of the singleton teachers over the years have had combined level classes, sometimes with levels 3, 4, 5, and AP all in the same class at the same time. Daniel resorted to using Google forms as a way to assess students, saying, “When I have one class, I barely use Google forms. But it's like, when I have combined classes, I don't have a choice. Sometimes I've got to.” It seems that Google forms were not his first choice to assess students’ learning in languages, but they allowed him to work with multiple groups of students at the same time.

**Native Speakers**

Just in the last two years, this school has begun to develop a Native Spanish Speakers program, with classes designed to improve this population’s first language ability. However, in the other singletons’ classes, there usually remained a few native speakers who perhaps did not get placed at the correct level, or who wanted to keep their first language skills fresh, even though the course material was not up to their level. Daniel said that assessing these students was more complicated, because he had “to grade them certain ways based on their performance. It’s that some of them do know the language already. Usually the problem is the writing.” Laura also tried “to make it a little more challenging” for her native students on assessments. On one assessment, she required her native students to take the test without notes, but for her non-native students, she “allowed them to use a folder that they created throughout that semester,” to aid them on a particularly complicated grammatical concept. Assessing native speakers and non-

native speakers in the same class required the teachers to walk a tightrope between making the class too easy or too challenging. Differentiation, along with providing or eliminating accommodations like word banks and grammar notes, become key in assessing a mixed-ability group of students.

**Special Education Students**

Several teachers spoke about the challenges in assessing students with identified learning difficulties. None of the teachers had any support staff to assist with small group testing or to aid during instruction. Alicia explained, "I have a lot of special ed kids," and "I have a few dyslexic students." Her solution was "I treat all the kids the same, extended time, . . . because nobody wants to be pointed out different or unique, or discrimination, or you name it . . . So I do it by individual." She provided the same accommodations for all her students so that none of her students felt singled out, while also watching carefully that her students with learning difficulties received the individual attention that they needed.

Before the semester even began, Carmen researched her students' records to spot students that were "special ed, remedial, 504." She also noted, "I have some gifted in there. I have some magnet students, honor students." In response, Carmen became very highly skilled at differentiation:

> So when I pass these out, I place them on their desk on Monday, and the one they have is the one I want specifically for them to have. So they're differentiated. They're not the same. So, the front is the same for everybody. So it says Monday. It's a grid, and it has the same two questions, or whatever it is. The back is going to be different. So I might have a listening that has true/false, if you're not as strong a student. Or you have to answer in full sentences, if you're a stronger student.

Carmen also provided extra notes for students who needed them: "Again, it's differentiated. So some people's papers, I may underline key sentences, keywords. I might write

in the margins, the conjugation . . . on some people's papers." This individualized level of differentiation took time and added a layer of complication to all her assessments.

Daniel found that the pacing during instruction and assessment was the biggest factor for students with varying abilities. In his words, it "depends on the students, because some students you can go quick. Some of them will have to take time, and this is why I always try to change the tests." He adapted his tests to make them more feasible for all his students' testing speeds.

Laura adapted her assessments, too, by double-checking students' vocabulary knowledge in spoken, rather than written form. She said that some of them, "They just cannot spell . . . for whatever reason." So she would call them up and ask about their misspelled written work, "What is this?" explaining, "I give them credit for even saying them to me, right. So I, again, I try to work with them when it comes to stuff like that, because I don't give them a word bank." This individual attention took time and additional effort when it came to assessing students.

**Student Class Placement**

Students frequently complain about the reasons why they are required to take high school courses, and world language classes are no different. However, since they are electives, the students presumably chose a specific language to learn. But some students float through the system without taking ownership of their choices and get thrown into classes that they did not choose. For example, if they did not fill out their elective selection form, their counselor might enroll them in a world language class in case the student later decides that they would, after all, like to apply to a four-year university with a language prerequisite. Laura saw this problem frequently in her classes, explaining that students sometimes did not know "what they wanted to do, why they're taking the class, that maybe they're just trying to check off the list." Martin viewed the situation slightly differently, in thinking that "As in any language, you would get

students signing up not knowing what they're signing up for." He seemed to be saying that students might have actually chosen the language class that they were in, but they had no idea what it meant to learn a language. If the students did not have motivating reasons to learn the language, the teachers might have struggled to design engaging assessments and students' grades might have been affected.

**World Language Assessment Resources**

The district provided the world language teachers with officially sanctioned testing materials and resources. Some of these were "grab-and-go." In her interview, Diana explained her happiness upon discovering some materials that accompanied her textbook that had been lying around unused in the department, saying, "There's a workbook that comes with it. So Martin found it! We have a bunch!" She thought, "Perfect! So I don't have to make copies! So, there's some good activities; not everything is good. So sometimes I say, okay, 'Do this one, do this one.'" Even with this treasure trove discovery of ready-made materials, Diana realized that not all the assessments were useful for her classes, and she had to spend time curating the materials.

Daniel talked about using the district book materials, but also outsourcing to previously used assessment resources, "Because I use not only the book, but I have other books I use as well. Tests, things like that. Yeah, the books we have sometimes, the texts are really, I would say, difficult." Julia echoed this problem with her commercially published materials: "It was all writing, no multiple choice. So then I adapted from that. Some of the questions I changed. It was very difficult, because it's a college book. So I made it simpler." Sometimes the issue with commercially published materials was not only the difficulty, but also the focus. Julia explained that the textbook she used was "not very listening-oriented, so the listenings are kind of like

checkpoints . . . Normally it's like, they give you a sentence, and they say, 'Is that formal or informal?'. . . There is not much to assess."

Alicia expressed frustration as her language was relatively new to the district and there were few readily available resources: "They need to have somebody to create all these things, not the teachers, because it's very time consuming. Because it took–one test, ten questions– for me, it took me eight hours." She further explained, "We make our own videos, our tests, our assessments. We use each other to incorporate this curriculum, and they're taking a good old time with this new book adoption." In her case, she was spending large amounts of time creating assessment resources from scratch. At the same time, with over twenty years of teaching experience, she also realized that "Within the classroom, the type of kids we have, there's not one shoe size that fits all. Every year, we need to improve, period. No questions asked." She continued, "That's where we always have to revive, revamp. You know, reapply. Those are the three things I do every year and within every day in the classroom, because you can't have anything ever set in stone." What she seemed to be saying was that even after she created these assessments for this group of students, she would still need to "revive" and "revamp" those same materials for a different set of students in the future. The assessments that she created would never be the final version, never "set in stone." The availability of materials for all these teachers affected the way that they assessed their students, since they were constantly revamping, adapting, curating, and searching for assessment resources.

**Lack of Assessment Training**

In world languages, the current push, according to Brown and Abeywickrama (2019) is towards alternative assessments like performance-based assessments, task-based assessments, and integrated skills assessments. Often these kinds of assessments require the use of rubrics to

assess product quality and language proficiency. Several teachers, Alicia, Daniel, Diana, and Laura frequently used rubrics when grading their performance-based assessments, and seemed quite comfortable adapting and adjusting them to meet the needs of their assessments. Carmen mentioned using a rubric once on a skit speaking assessment. Martin stated that he did not use rubrics, although he recognized that they could be useful, since "Some students, at least, are wanting those limits of boundaries. And so every now and again I kind of kick myself for not doing it." Julia laughingly called herself a "rebel" for not using rubrics, "Horrible!" she jokingly described herself. She said she wished she knew how to use rubrics, but she felt that rubrics "don't give them a chance to be better." She said, "I feel like, okay if you don't fit in this box – 'Ah! You're wrong!'" When she listened to her students' speaking assessments, she would remark, "'Oh! I like how you said that!' Then I'm like, 'Okay, I feel it in my heart, you're good. [laughing].'" She felt that rubrics were confining, either "You say it like this, or you don't. So that's why I feel like it's so– ugh!" These statements from Martin and Julia hinted at a possible lack of training in the flexible boundaries that rubrics provide for both language teachers and students. But that lack of training was not addressed in general professional learning sessions at the school. Most of the professional learning provided by the school centered around core subjects like math, language arts, science, and social studies. So these teachers either chose to grade their performance assessments more subjectively, opening themselves up for potential criticism of their grading methods, or they chose to stick to traditional pen-and-paper assessments to avoid having to deal with rubrics at all.

Besides abiding by the requirements of the state, district, and school, these teachers were also influenced in their assessment choices by class sizes, class make-up, classroom resources, and assessment training. Their ideal assessment practices were not always feasible considering

the limitations of their environment and knowledge, and they learned to adapt to those boundaries.

## The Contact Zone at the Micro Level

The focus of education is on the students. But when the students' focus is not on education, the work can be challenging. Today's students come to school with many distractions and issues, and teachers have to meet them where they are to help them be successful on assessments.

### Student Issues

Many American students today work an outside job to help support their families or for extra income. For example, Alicia told a poignant story of a student who basically supported himself because he did not have familial financial support, "I work, you know. Yeah. It's my main income." She also mentioned absenteeism as a problem, "You don't know what's going on behind these kids. You don't know how many dates they're missing." Carmen corroborated, telling a story of a boy that "had a lot of extreme situations in his life." Some of the students missed a lot of school, either because they were working, or they did not have parental figures who prioritized their attendance. Martin explained that he adjusted some of his grading policies for exactly this reason: "We have people, for whatever reason, . . . who are out quite a bit. But then when they are there, they do produce." To accommodate these students, he gave weekly holistic grades instead of daily grades on individual assignments saying, "I'm working on the basis that if they are working when they come in, then I make the assumption that if they were there five days a week they would be working." While school policy tried to discourage truancy, classroom teachers still had to find ways to grade and assess when those policies failed.

When these students did come to school, sometimes they could hardly hold their heads up due to exhaustion, like Daniel described, “There's nothing we can do about these students. And then maybe they have some issues . . . We don't know. And some of them, they come to class sleeping, and there's nothing we can do about it.” Student absenteeism and exhaustion affected the students’ performance in the class and on assessments.

Learning a second language can be a scary experience for anyone, but especially for teenagers. Alicia stated that “These kids have no self-esteem, period,” further explaining, “You know, they’re loners.” She believed, “That's where communication comes in most, they have to learn to communicate.” Language classes generally require students to communicate in social situations, and these teenagers struggled with that. Carmen did fewer speaking assessments for that very reason, saying, “I just felt the two classes I had of Spanish 1, a lot of anxiety, unidentified, I would say, anxiety, autistic in a sense that, producing language, causing anguish.” Julia had to change her presentation format: “Before the pandemic, I used to do this in front of the class. I would put them in there, present and read. But now they can’t. They get super stressed. ‘I have anxiety.’ Okay? So then we go to the hallway.” Laura said, “I usually try to assess when I know they're doing well, like they're not worried about some of our, I think, the issues that they have.” And Diana also did speaking assessments in the hallway, but still described a stressful scene, with “students crying, it's a lot of drama.” She said, “I don’t know if it’s anxiety,” but the students “are trying, you know, like they're shaking and crying.” Even Martin found that “Crosswords freak certain people out. Yeah, because they feel almost claustrophobic by putting letters into a box.” He remarked incredulously, “You never know what you're gonna meet!” All these teachers found it challenging to assess these students with so many personal and emotional issues.

**Student Disengagement**

Several teachers complained about the students' lack of engagement, despite their physical presence in the classroom. Laura explained, "I'm not saying that they're bad . . . They're just quiet and listen. You think they're engaged. But they're not. They got their headphones on, or they're just not there." Daniel agreed, "Some students will spend like 50% of the time on the cell phone not working." Diana pondered how she could change her activities to engage the students to be more productive and motivated on task-based assessments, thinking out loud, "How would I make sure that everybody's working? And you know, it's not just two people working and the other ones on their phone the whole time." Julia confessed that she purposely gave the students kindergarten-like assignments such as researching country flags and coloring, "They had a week to do it. They didn't do it. It was literally coloring." Student disengagement caused these teachers to question whether there was something wrong with their assessments, or if the problem was truly with the students.

**Student Refusal to Work**

Several of the teachers, like Martin, defined a low achieving student as "somebody who simply does not turn in the work." Daniel agreed that "Low-achieving, that is those students who refuse, without trying." Alicia explained that achievement had nothing to do with the students' ability: "It's not the high and low. The lazy ones. It's not the high achievers. Has nothing to do with that. It's the lazy ones. 'We don't wanna do nothing.'"

Some might argue that work output is not an indicator of student learning. However, Diana discussed that this work output was the only way to tell if the student understood and could apply the concepts and language that they were acquiring: "There are people that don't do it. And if they don't do it, cause I see one by one, and I also give them feedback, like 'This is

how you're supposed to do it.'" But she said, "Still, sometimes, they don't do it. I don't get to grade . . . They turn it in the same day of the quiz. 'Oh, yeah, it's too late. We're gonna take the quiz right now.'" Holding the students accountable for their daily work became a way for teachers to formatively assess the students before they were faced with a summative assessment.

Laura explained what happened later in the semester with her classwork activities, "I don't grade all the packets, as you can see, because I can tell how they're doing. But when I start seeing that the kids are not turning in the packets, . . . I need to start collecting them." She was forced to use low grades in the gradebook as negative incentives to ensure that students were learning. She said, "If I don't put a 0 down in the gradebook, they don't, they forget. So I'm trying to be very diligent about going back and putting those, because they need to know what their grade is." Even if she did not grade all her packets, Laura graded all the warm-ups that she gave at the beginning of class as a way to encourage students to engage with the class material: "I'm like, 'Don't you have a warm up?' 'Oh, yeah!' 'I'm like, okay, the bell rang, warm up! Get off your phones!' So, you know, so it's getting them to do that." In her gradebook, Laura had twelve entries for weekly warm-ups in a span of eighteen weeks. In total, she had over forty gradebook entries. Three of the teachers have more than fifty entries, with one teacher having more than sixty total. Two of the teachers had approximately thirty entries, and Martin had the fewest. He used more of a weekly grading system, and explained his reasoning, "Everything is based upon total productivity through a period, for the period of a week." That way he kept his number of gradebook entries down at twenty-one, while the other teachers graded more frequently to encourage students to engage with the coursework.

Regardless of the number of grades that teachers entered into their gradebook, all of them felt forced to use 0's as negative incentives to engage their students in language learning.

Therefore, not only did teachers adapt their assessment habits by modifying the manner in which they probed student learning, they also adjusted the frequency with which they assessed the students to keep students "playing in the game," as Martin expressed it.

**Student Cheating Culture**

All the teachers addressed student cheating in their interviews. Some of them, like Daniel, adjusted their category weighting over the years to account for the lack of oversight when students did homework, saying,

> You don't have any control. I don't have any control. They're cheating. They cheat, and no, by the time you are helping a student, you know, and then . . . you can give them different copies, but they will find a way to do things, using the cell phone to translate and find answers. It's homework . . . There's nothing I can really gauge the students' learning from the homework. And one thing, you can have a student perform well on homework. But when it comes to quizzes, and summative assessment, nope. That's the cheating part. Okay? Because yeah, everything's perfect, you know. And they can use their resources. They can use their cell phone. And all of that. It's a homework. And then you think, "Yeah, they learned the concept," but at the end you realize–nope, they didn't.

Daniel therefore lumped his homework and in-class activities into the same category, and weighted that category at only 10%. But students also cheated on in-class activities. Julia characterized low-achieving students as those who "take pictures of somebody else's work, and just copy down the answers, and then just turn it in because they want a grade." She also found that "Many times they go on Google, and they come up with these crazy sentences. I'm like, 'No.'" Diana added, "They copy from others. Every once in a while they know what they're doing." To combat these bad habits, the teachers got very creative. Alicia changed her tests from year to year, "They have to practice all–'What! two hundred words!' And I only pick ten. You know, they never know what they're gonna get." She even warned her colleague at another high school against students cheating, saying "She was using her old videos. That's like, you can't do that. I mean you can. But you gotta keep revising, revamping for future generations." She also

knew that students cheated on her cultural readings, telling them when they copied from other students and turned the packets in quickly, “Take your time, you know. Read it over.” She purposely included questions about these reading selections on the final exam, because she said, “It tells me if they recall, remember, or did they just copy somebody.” She continued, “I warned them, because I know how people cheat.” She wanted the students to glean knowledge from the readings, and be able to show her their knowledge, not just fulfill the requirements of a reading assessment.

Carmen made multiple versions of her quizzes. For her, “Sometimes it’s a matter of, it’s changing the font. It's changing the order of it. It may not even be that different, but they don't pick up on it.” She also did not give quizzes or tests back to the students after grading them, calling that “a rookie mistake,” because she often had students that needed to take the quiz or test at a later date due to absence. She even had to change her warm-ups from class to class over the years, since she found out that the class at the beginning of the day was sharing the answers with their friends in the class at the end of the day.

For the final exam, the American Sign Language teacher took her students to the media center, where they each took the exam in a Microsoft Form on the computer. She explained, “I had to stand in the library in the front. I stood on a chair.” She is deaf, and she wanted to ensure that they were not cheating and giving each other answers. “I said, ‘Don't think about talking to yourself, moving your lips, whatever, it's gonna be a 0.’ Not one person moved their mouth. Not one person. Cause I said, ‘This is a final exam.’”

To fight this culture of student cheating, teachers went to extraordinary lengths to adapt their assessments to stay three steps ahead of the students. They had to adjust the format, the

expectations, the assessment versions, and the environment to force the students to do the thing they did not want to do the most, which was to LEARN!!

## Conclusion

This chapter examined the zone of contact where world language teachers made decisions, developed practices, and wrestled with their beliefs about assessments and grading. The macro level of the school organization, the meso level of the classes and resources, and the micro level of the students all formed the contact zone where authoritative discourses were adopted or modified into the teachers' internally persuasive discourses about what, how, and why to assess and grade. The next chapter will further explore the teachers' grading and assessment practices, revealing the internally persuasive discourses that developed through the tensions in this zone of contact.

## Chapter 5 Teachers' Internally Persuasive Discourses

This study endeavored to ask questions about teachers' rationale for their assessment and grading practices, to arrive at a full understanding of their beliefs about student achievement. This chapter will examine the teachers' internally persuasive discourses in depth, revealing what they assessed, why they assessed, and how they assessed.

## Gradebook Categories That Reflect Achievement

All the teachers included several categories for measuring student achievement in their gradebooks. Besides the district-mandated final exam category valued at a standardized 10%, the grade category and weighting varied significantly by teacher.

### Tests

Alicia, Carmen, Daniel, Diana, and Laura all had a category with the label "Test" somewhere in the name. Julia had a category labeled "Projects and Presentation," but she clarified that she also meant to include "Tests" in the title as well. Julia said she wanted this category to be higher because she felt "that would measure . . . that would take more effort, not just the day in, day out, in the classroom." Martin called this category "Summative," and Daniel also included this term into the title of his test category called, "Summative Tests." Carmen, Laura, and Diana also used the term "Performance Tasks" or "PBA" (Performance-Based Assessment) in their Test category titles. All the teachers weighted this category at 30%, except Alicia, who weighted it at 10%. In her interview, she said she was planning to change her weighting to add 5% to her test category because, "it's overall." Laura also had a "District Assessment" category weighted at 0%, where she attempted to implement the brand new on-line district-wide summative assessments that were meant to be used as pre-tests and post-tests for each unit. She posted two assessments in this category, but she said that the tests generated no

data because of technical problems like kicking the students out of the system and issues accessing the video and audio files on the tests. The grade postings were empty, but she never deleted the posts or the category. She believed that the assessment could "help teachers" by providing diagnostic data in the form of pre-tests and achievement data through post-tests. She was frustrated by the two failed attempts, because she believed it could provide useful achievement data.

In their Test categories, most of the teachers required students to show their achievement without using class resources, relying solely on what they knew. But the percentage of closed-note assessments varied per teacher. Alicia, Carmen, and Daniel assessed exclusively without allowing students to access resources. For Diana, 18% of her 30% Test category was closed notes. Exactly half of Julia's Project/Presentation (and Test) category was administered without resources, for a total of 15%. Two of Laura's tests were administered without resources (12% of her 30%). Martin never assessed his students using a closed-note format, saying, "I base it on the belief that repetition is more helpful than asking them to remember stuff which they don't know." The decision to allow resources or not correlated strongly with teachers' definitions of achievement, which will be discussed later in this chapter. But all the teachers' gradebook Test categories indicated that measuring student achievement was a priority.

**Quizzes**

All but one of the teachers had a Quiz category, although there were some variations in the titles and weighting. Martin did not include a Quiz category at all, but the title "Review Quiz" appeared once in his Formative category. However, this was not a closed-note quiz, but rather just a different naming for his usual weekly formative grade that included all the class notes, translations, and grammatical, cultural, and historical activities that the students had completed

throughout the week. Again, he stated, “Everything I ask them to do is open notes.” Martin’s beliefs about achievement were not related to students’ ability to perform on closed-note quizzes or tests.

Julia’s Quizzes category came in at the second lowest weighting at 15%. Like Martin, fifteen of the sixteen items in this category were open-note, teacher-assisted. Six of the items in her Quizzes category were evidence of student note-taking during classroom instruction, and only one assessment was closed notes. Four of Julia’s six categories included both open- and closed- note assessments, and she had the second highest number of open-note assessments overall, at 65%. When asked about how she came up with her categories, she replied laughing, “I don't know. I think I copied that from someone- horrible!” She continued, “I adapted with the Quizzes and Tests/Projects/Presentations–they used to be lower,” explaining that “the Writing/Speaking, and the Listening/Reading, and the In-Class—it’s something that you do constantly. So I didn't put that as high.” Two of the assessments in her Quizzes category were identical to two assessments in her Projects/Presentations (and Test) category, but in the Quizzes category, they were weighted .94% each, while in the Projects/Presentations (and Test) category, they were weighted 2.5% each. Chronologically, the higher-weighted assessments came before the lower-weighted assessments, so it does not seem like her intention was to build up in complexity or impact. Julia seemed to struggle to categorize her assessments into like groupings, possibly due to time constraints, or maybe she felt pressure from her collaborative team to use category titles that did not match her grading beliefs.

Carmen, Daniel, and Diana did not allow students to use resources at all on the assessments administered in their Quiz categories, presumably to have a more accurate measure of the students’ acquired knowledge. Carmen explained that her quizzes were “a check-in to see

what you know," saying, "I try to use that information to guide me for the test, and I try to tell them, 'Oh, see, this is the one area you need to work on.'" She used the quizzes as Assessments *for* Learning. Diana described the vocabulary and grammar on her quizzes using phrases like, "they have to know" and "they need to know." Daniel, whose Quiz category was 20%, defined his quiz as "a formative assessment, to see exactly if they study, if they understand the vocab." These teachers believed that a closed-note assessment would be able to show students' knowledge.

Carmen, Diana, and Laura all valued the Quiz category at 25%. Like Carmen, Laura called her quizzes "checkpoints." On seven of her eight quizzes, Laura prohibited students from using notes, and the format of those quizzes increased in difficulty throughout the semester. For example, at the beginning of the course, the students had to translate vocabulary from L2 to English, so only their recognition of vocabulary was assessed. Laura reasoned that she "tried to start them off where they're not intimidated." Later in the semester, students had to label a diagram in L2. Laura specified that the students would have to recall the words from memory, saying, "No word bank. They have to know the words." A later quiz required them to read a definition of the vocabulary in L2, and also write the word in L2 with a word bank as support. And the last quiz of the semester was translating vocabulary from English to L2, again without a word bank. She was pushing them to move beyond just recognition of the vocabulary to achieve full processing and recall of the vocabulary, to show their achievement on the knowledge that they had accumulated.

But Laura did not keep her Quiz category as fully closed resources. She also included an open-note creative writing assignment. Since she was teaching them a new grammar structure, she wanted "to look at" their writings throughout the assessment process, to "give them

feedback" before they turned in the project. She needed "them to know where they're making those mistakes," basically "pre-grading." Even though this assessment in her Quiz category was not closed notes, she was still able to measure their understanding of the use of the new structure, which was part of her definition of student achievement.

Alicia's Performance-Based Assessments/Quiz category was weighted the highest of all the teachers at 30%. This was also her highest category, counting three times more than her Test category. This category was purely performance-based, since it consisted of ten presentations in front of the class. The students practiced and prepared with a partner ahead of time, and they were allowed to use the notes that they had taken during their preparation. Alicia was adamant to look over their notes before they presented, saying, "I go over it with them. So when they get up there, if they don't utilize that, that's their grade." This category was meant to provide group feedback, "That's how we learn. . . the feedback." All these presentations, because of their group format, were Assessments *of* Learning, Assessments *for* Learning and Assessments *as* Learning, all at the same time. The students had to show what they could do in the language (Assessment *of* Learning), while they got feedback on their performance so they could continue to improve in the language ability (Assessment *for* Learning). They also got to watch others perform and receive feedback so that the assessment itself was a learning experience for all students (Assessment *as* Learning).

**Final Exam**

Because of the strict district-mandated 10% grade weighting requirement, combined with the tight time constraints of administering the final exam and posting the final grade for the semester, several teachers found it necessary to vary their usual assessment practices for practical reasons. Some of the teachers were able to conduct business as usual, with similar testing

methods as their normal tests, while others had to adjust. Most teachers were able to adapt the final exam category to reflect their beliefs about student achievement, even if the testing conditions were not ideal.

Since Martin never gave tests throughout the semester, he continued this pattern for the final exam by requiring the students to complete a project ahead of time. He reasoned, “I'd rather have you working for me for eighteen weeks rather than worry about what you do in an hour and a half at the end of eighteen weeks.” Julia also was able to use a testing environment that she had used before, because two of her previous tests had also been fully scantron. Just like her previous two scantron tests, she felt like that format would be able to show her students’ achievement. Diana and Laura were both accustomed to assessing students on their speaking and writing, so they continued that practice by administering the speaking and writing portions of the final exam several days before, so that reading and listening comprehension were the only things assessed using the scantron format for speed’s sake on the day of the last day of class. This way, students were able to show their achievement in other skills besides just reading and listening, to have a balanced reflection of what the students had learned throughout the semester.

Carmen was mostly able to use her normal tests, with the only concession being to use scantron instead of having the students write out their answers or use complete sentences which was her preference. She said that the students were very familiar with the final exam material because they had “done it so many times. It came from a previous quiz. It came from the previous test.” For her, the final exam was a good reflection of what they could do because “It's so familiar. They're actually done very quickly. I want them to feel good about it. ‘Look at me! Look what I can do! I know this!’” However, she did mention that unlike on her tests, which always included a listening section, she did not include listening on her final exam, saying,

"There is no listening. I have found that can be a little stressful." She seemed to want them to finish the semester on a high note without undue stress, while also demonstrating their achievement in the language.

Daniel copied and pasted previous scantron test questions to make his final exam for expediency's sake, but it was not his preferred method of assessing student achievement. "I used to do three final exams. But . . . it takes time to finish." He used to include a spoken and written section on his final, while assessing the listening and reading on his scantron portion on the last day. But since then, he became wary of formally assessing their listening skills, just like Carmen, explaining, "Some students who perform well on writing and speaking, and comprehension, but when it comes to listening—ah no." He felt like the students really struggled in this area, based on their results from previous quizzes. Because the final exam was worth 10% of the grade, he knew that this negative impact could cause big grade fluctuations on the last day of class, "because it's hard at the end of the semester to have your bad grades." He was willing to administer a final exam that did not truly reflect his beliefs about his students' abilities in order to maintain the grades that the students had earned over the course of the semester. He felt that those previous grades better reflected their overall achievement.

Alicia's final exam was the furthest from her normal method of assessing, but it was a good way to measure a different kind of student achievement that she had previously not assessed. Almost 50% of her assessments during the semester were performance-based, across multiple categories. But for her final exam, for speed of grading, she chose to assess the students over the reading packets that they had done in the Classwork/Participation category. She said that asking about those cultural and historical readings on the final exam showed her "if they recall, remember, or did they just copy somebody" when they got a completion grade for

handing in the packets throughout the semester. For her, “History is the main thing, you know, cause if they don't have that, they're not gonna understand.” This final exam was her opportunity to hold the students accountable for their level of understanding of the culture and history of the speakers of her language.

**The Problem with Homework**

Teachers in this group overwhelmingly believed that their grades should reflect student achievement. This belief was further revealed through their discussions about homework and accepting late work. Several of them seemed to struggle with the inclusion of a Homework category in their grading categories. Alicia and Martin did not include a Homework category at all. Julia and Daniel lumped their homework in with “In-Class Activities” and “In-Class Assignments,” with category values of 10% and 15%, respectively. Julia only had one true homework assignment in this combined category, and Daniel had none. Laura, Diana, and Carmen all included a Homework category valued at 10%. However, Carmen and Diana both posted at least one in-class activity in their Homework category, even though they both had a separate dedicated Classwork category in their gradebooks. Several of the teachers expressed concern with homework assignments, since the lack of oversight allowed for cheating, which was not a true reflection of student achievement. Daniel stated, “There's nothing I can really gauge about the students’ learning from the homework.” For this reason, he rarely accepted late classwork completed at home and turned in the next day. Diana concurred, “You know they cheat off each other.” Carmen reasoned that “At home we have other devices and access to other things to help us. And it's not a clear indication of what they can do when I just showed them something.” Carmen also did not accept classwork completed at home, saying, “No, I'm basing it on what you did today with me. And that's what I want to look at.” She mentioned several times

throughout her interview the importance of being able to watch the students throughout the learning and assessment process. Teachers were concerned that homework did not reflect true student achievement.

Carmen only entered two grades in her Homework category, which would have made them each 5% of her overall grade, more than three times the value of an individual quiz grade in her gradebook. However, one of the homework assignments was an in-class warm-up, which was inadvertently valued at 7.5% of the overall grade. This happened because of her frustration with students not completing the homework that she had been consistently assigning throughout the semester. She changed the value of her other homework assignment to a half value, and the gradebook program automatically calculated that assignment to 2.5% of the overall grade, forcing the other homework assignment to be 7.5% to equal the 10% category weight. She explained that she felt that entering all the 0's in the gradebook for the students who did not complete the homework activities that she assigned throughout the semester was not "reflective of what they know." She further explained,

> I don't have anything to show, which makes me reconsider the homework category. But I am on a team, and we do decide the percentages together. So I try to make sure the grade is reflective of what the students have produced, and I have evidence of. I mean, it's a tricky category.

Laura had the same problem with students not turning in homework assignments, and she had to backtrack on her own policy of not accepting late homework at the end of the semester when she realized that her strongest students' overall grades did not match their achievement in her class: "They've done everything else, and they're like 'A.' They have a B my class because they have no homework. Then obviously, I know they've mastered what they have to master." So she allowed them to turn in homework assignments late after all, with a 50% penalty, calling this

homework category her "wiggle room." However, if this 10% Homework were eliminated, all the teachers might feel more comfortable that their students' grades reflected achievement, whether the problem was high-achieving students with low grades, or students faking high-achievement by cheating on the assessments in this category.

It was clear that these teachers had varying, but strong internally persuasive discourses about how their gradebooks should reflect student achievement, particularly in the naming and weighting of their achievement categories. The Homework category caused conflict in the internally persuasive discourses for all the teachers because the reflection of student achievement in this category was questionable. The teachers each responded to this conflict differently, according to how they could resolve the discord between their personal beliefs and the expectations of their collaborative teams in the contact zone of their classroom context.

## Teachers' Varying Beliefs about Student Achievement

These teachers all seemed to believe that student achievement was important when assigning grades, but the definitions of that achievement varied greatly. Their assessments varied according to what the teachers believed demonstrated achievement, whether it was the depth of students' understanding of the underlying concepts of the language, a measure of the quantity and quality of the language acquired, the students' ability to use the language in context, or the students' understanding of, and connection with, the culture and community of the language speakers.

### Belief: Achievement is Understanding Concepts

Two of the teachers leaned heavily into the idea that language is the understanding of concepts. Both Carmen and Martin defined high achieving students as those that "can notice patterns in the language" and "grasp the language." Carmen talked repeatedly about "concepts"

and students who could “recognize the patterns,” “process the language,” “figure out systems,” and “use the formula,” as she discussed how she distinguished strong students from weaker ones. Some of the assessment questions that she used as examples in her interview related to the “formulas” needed to create certain language patterns. In several instances, vocabulary was not assessed as a tool for communication, but rather as a way to “reinforce the idea that some words are masculine, and some words are feminine, and when we learn the words, we learn them with the article.” She also graded most things for accuracy, “I rarely do completion,” because she wanted students to have correct answers to show that they understood the concepts she was teaching. She said she used a lot of online self-graded activities if possible, so that her students could “get feedback right away,” adding, “I could reset it [the on-line activity]. And say, ‘Now, you think you understand it better? Let's try it again.’” She wanted to make sure the students understood the patterns and the formulas of the language.

On her daily warm-ups Carmen often used two similarly worded questions with different verbs that had differing conceptual meanings to see if the students noticed the difference between the two verbs. Students who did not understand the differing concepts told her, “You wrote the same question twice!” She said she repeated these questions multiple times on warm-ups throughout the semester, correcting them with feedback every time, saying, “By the time we get to May, if I use that same question, I shouldn't have the same answer.” Carmen traced the students’ confusion over this particular concept back to a presentational speaking assessment that she administered early in the semester, saying that one of the verbs was “ingrained because of the dialogue,” and so the students struggled to conceptually separate the two similar verbs. This concern about the students’ grammatical errors becoming “fossilized” likely contributed to her lack of logging interpersonal speaking assessments in her gradebook. When she collected student

interpersonal speaking data, she felt that it was “off,” because of “comprehension,” saying, “I don't think about pronunciation at all.” She explained her reasoning behind this practice to her students by using a deep Southern accent to speak her language, illustrating that they could communicate even if they didn’t pronounce the language correctly, as long as their answers were “grammatically correct.” Possibly she was worried that the students would ingrain incorrect phrases into their speech patterns through continual use, even if their answers were comprehensible. She worried, “once they fossilize something incorrectly in the brain, it is hard to undo it.” She prioritized conceptual understanding of the language over communicative use of the language.

Martin used similar terminology to describe “coming to grips with the language,” “recognizing the basics of the language,” and understanding “the framework of the language.” He explained “language is for the most part a language of beginnings and endings,” but that even his third- and fourth-year students struggled “to look at the language in that way.” His stated idea of achievement was “to what degree they think the language, not so much the accurate application of it.” To push students to grasp the structure of the language, Martin “graded more and more on accuracy” as the semester progressed, saying that he was looking for “accuracy of word formation” and for students to use “the formula to come up with a grammatical construction.” By offering the students choices on several of his assessments, which he called “snake sentences,” he was able to map the way that they were thinking about the grammatical concepts that he was teaching them, therefore determining their progress in understanding the language framework.

To a lesser extent, Laura also subscribed to a conceptual understanding of language. Her definition of high-achieving students related more to class engagement and other academic-

enabling behaviors, but in her interview, she also used phrases like, “easy concept for them to grasp,” “the hardest concept,” and “the major concepts.” She described “A” students as those who had “mastered what they have to master.” Laura’s focus on conceptual understanding of the language was reflected in her gradebook in her Test/PBA category. 12% of this 30% measured the students’ understandings of the uses of the past tenses. One assessment was an open-note compilation of all the correct past tense forms and the reasons for using those forms. On the next assessment, the students were allowed to use the notes that they had compiled on the previous assessment to apply the reasons in context. Her focus on these two assessments was less on memorization of the forms, but rather on understanding why and how they were used. On four of her eight assessments in her quiz category, Laura assessed students on “the reasons” for making certain vocabulary and grammar choices, saying, “I tried to get them to let me know that they don't understand, because if they don't understand, I can't help them.” She talked about “key-words” and “buzz-words” for students to better recognize patterns in sentences to make grammatical choices. She said she could tell when students had “mastered” the concepts. Her statement, “Once they get it, they get it,” when talking about one of her grammar quizzes, showed her belief that student achievement was understanding the underlying concepts of the language.

Alicia was the only other teacher that slightly hinted that she defined student achievement as conceptual understanding. On her final exam, worth 10%, she “questioned on the grammar rules” along with history and culture. These were not application types of questions to show students’ ability to demonstrate use of the grammar, but rather conceptual understanding of the language itself. She also ensured that the students understood the structure of the language before they had to present to the class for their Performance-Based Assessment by “walking around

when they're practicing." She said, "So they ask me, 'Is this right?'" And she answered, "Okay, what do you think? . . . Is this a noun? Is this a verb? What should go first?" By answering their questions with more questions, she pushed them to understand the structure of the language, so that they would be able to better communicate during their presentations.

**Belief: Achievement is a Measure of Accumulated Knowledge**

Some teachers did not see language as a set of concepts to understand, but rather as sets of knowledge to accumulate. The assessments that reflected this belief placed a strong emphasis on memorization and discrete-point acquisition of disparate components of the language like vocabulary, verb tenses, and grammar.

***Memorization***

One teacher in particular put a premium on students' accumulation of knowledge in the form of memorizing vocabulary and grammatical forms. Laura explained her strong stance on this accumulation of knowledge: "Half of learning a language, you have to memorize. If you don't memorize, you can't speak." She gave a vocabulary quiz for each unit, most of the time without a word bank, rationalizing that, "They have to know the words." She chose twenty to thirty "essential" words that the students had to memorize. Laura also had a student create an online study tool to share with the rest of the class, so that they could practice the exact format of the quiz. She used mnemonic tricks in the classroom to help students acquire these words, and she said, "If somebody says something good from one class, I'll pass it on to another class," to share this study trick with her other students. She also gave quizzes over verb forms, pushing the same emphasis on memorization. One of her quizzes was a song that was meant as a study tool to memorize difficult past tense verb forms with their infinitive form, which she described as "super easy for them, cause they've known both of them, because they know the song." Knowing

the song would not necessarily mean that they knew how to use those verbs or recognize those verbs in other contexts, but the students could show achievement by demonstrating their memorization of the song. These tools—the one-line study resource, the mnemonic tricks, and the songs—provided her rationale for not using word banks on her quizzes: “I don't really do word banks, because I feel like I give them everything for them to memorize whatever they have to memorize.”

Laura’s presentational speaking assessments, such as an individual country presentation, a group TikTok project, and a group skit, also hinged on memorization. These assessments accounted for 13.39% of the overall grade. For the students’ country presentation, she believed, “They can memorize 6 sentences,” although she added “I can help them out a little bit, you know. I don't want to scare them off at the beginning when it comes to speaking.” But she did want them to get used to the requirement to memorize right away: “I need them to feel comfortable, to go, ‘Okay, I memorized some sentences!’” Her goal in assigning the group TikTok project was again on “memorizing twelve prepositions” because she believed that “They’ll memorize this for their friends.” She told them, “Everybody's gotta participate,” and that she did not want “one person to be the cameraman, not do anything, because the more they practice it,” the better they would memorize the vocabulary in a communicative context. When describing the group skit, she said, “You can tell when the students memorize” and “You know which ones are struggling.” She believed that this struggle came from their memorization issues rather than any difficulties with pronunciation or fluency problems. Even so, she recognized that some kids struggle with language classes: “A lot of our kids’ issues—they can’t memorize,” and “They may have a hard time memorizing these words.” Her solution was, “Okay, pick a song, find a song.” For Laura, “practice and study” were the ways for low-achieving students to

improve their language abilities. In her mind, memorization of the vocabulary and grammar of the language was both the pinnacle of achievement and also the cause of language learning difficulties for some students.

Carmen also indicated her support of an accumulation model of learning languages in her definition of low-achieving students, saying, “They, again, won't be able to process the language, because they don't maybe have ways of studying and keeping the words in their heads.” Her description of how she adapted the final exam to fit her needs served as an example of her strong belief that this stored knowledge in the brain reflected achievement. Instead of using the traditional five multiple choices of A, B, C, D, and E, she also used double letters—AB, AC, AD, AE, etc., to make a larger word bank for the students. “You have to really know which word you need.” She stated, “I don't want process of elimination,” and “You just gotta know which one to use.” She also put all the verb choices in the same conjugation so that students could not guess the correct word to complete the sentence by looking at the subject of the sentence. She wanted to see how much the students retained from their language learning throughout the semester. Unlike Laura’s mnemonic tricks and songs and online study tools, Carmen believed that writing was the key to memorization, explaining, “We have data and information research that tells us when you write something, you encrypt in your brain. When you are physically doing it, it goes in better.” For this reason, she insisted on most things being “handwritten to encrypt the spelling into your brain.” For this reason, she disliked having to design a quickly graded multiple choice final exam at the end of the semester, and her large word banks were a way to ensure that the students’ choices reflected their stored knowledge rather than their guessing skills.

### *Discrete-Point Assessment of Vocabulary and Grammar*

Because of her goal for her students to accumulate sets of knowledge, Laura included a lot of discrete-point vocabulary and grammar assessments in her gradebook. The first half of all her quizzes assessed explicit vocabulary, where she tested lists of "essential" vocabulary through translations, labeling, and definitions. The other half of each quiz assessed a specific grammatical concept. This separate drilling of vocabulary and grammar was also evident in her Homework Category, where the students used a website to conjugate verbs, saying, "It's great for them to spell." Her first test explicitly assessed students' grammar knowledge: "This one has a lot of grammar because I wanna make sure that we're, you know, even if they didn't get it, or they forgot, we're on the same level." She mentioned that her warm-ups and packets in her Reading/Listening/Writing/Speaking category also separated vocabulary and grammar from their communicative uses, saying, "The more they know the vocab, the more they can proceed with the activities that they're doing. . . Every unit has a vocab packet. And some units have a grammar packet . . . Some are embedded together." Altogether, approximately 42% of Laura's overall grade assessed discrete grammar points and about 21% assessed discrete vocabulary sets. Together these show that Laura carried strong beliefs that student language achievement related to their accumulation of sets of knowledge.

Carmen also assessed vocabulary and grammar using discrete-point assessments. In her quiz category, Carmen prioritized the structure of the language over the meaning of the words in four separate instances, explaining, "I chunk it up. So we do one column, and then we do the next column . . . That first one is just conjugating. You don't have to know what they mean at all." She wanted them to understand the grammatical concept before she introduced the vocabulary meanings. She repeated this sequential quiz pairing of structure first, then meaning,

four times in her gradebook, which indicated her strong belief that student language achievement correlated more closely with conceptual understanding than communicating meaning. 20.7% of her gradebook focused on discrete-point grammar assessments, which she graded by prioritizing accuracy in the form of spelling and correct accents, stating, “I like this activity ‘cause it helps me see if they can spell.” Even within her explicit grammar assessments, she divided the material into smaller chunks, saying, “I wouldn't have mixed them all together. I separate them out . . . It doesn't help me if they’re all bundled together, because that's not the concept I'm looking for.” Assessing grammar discretely enabled her to find weaknesses in her students’ conceptual understanding so that she could address those issues point by point, “If they do it wrong, I say, ‘Oh, no, no, we can do it again.’” Carmen wanted to ensure that students grasped the concepts that she was teaching, and assessed those concepts in discrete-point sets.

Daniel’s prioritized communication over conceptual understanding, which will be discussed later in this chapter, but his gradebook still indicated a belief that student achievement was at least partially a measure of accumulated knowledge. He was very mindful of the way he built his gradebook for each of his units, starting with a vocabulary list and a crossword for students to “try to understand the vocab.” Then he gave them some completion or matching exercises with the vocabulary “based on the meaning.” Those gradebook entries were allowed to be completed using resources, but then, “Now comes the quiz . . . They have to study and understand the vocabulary in order to complete the quiz,” in a closed-note format. Then Daniel introduced the verbs for the unit, explaining, “You start with your structure, and they have to use it to see if they understand.” First “they have to conjugate the verb,” then “After they know how to conjugate the verb, now, they can make sentences.” Those two assignments were listed separately in the gradebook as discrete-point grammar assessments. Then, “We're gonna start all

over again," with the second half of the vocabulary for the unit, using the same three discrete-point vocabulary assessments and the same two grammar activities to build towards his summative test and his communicative performance-based assessment. "That's my structure," he said. In contrast with Carmen, who sequentially assessed grammar before vocabulary, Daniel assessed vocabulary before grammar for each of his units. His gradebook weights also demonstrated this priority, with 17.3% total discrete-point vocabulary assessments and only 8.15% total discrete-point grammar assessments. In his Formative Assessment category, he logged seven vocabulary quizzes before he recorded his first grammar quiz, with ten vocabulary quizzes as opposed to three grammar quizzes. This heavy, purposeful vocabulary focus enabled his students to both "master the language" and also "use the language" in later, more integrated assessments of their communication skills. But his gradebook also showed that he believed that his students could not communicate without the discrete-point building blocks of vocabulary and grammar.

Alicia also spoke briefly about how her students had to memorize vocabulary as a first step to being able to use it in context to communicate, saying that every day at the beginning of the semester they "review column one, column two, column three, as a class, as everyday repetition. That's the only way to go." Interestingly, she never assessed the students formatively over that vocabulary, choosing instead for them to use it in context in more integrated, performance-based assessments. Alicia's smallest category, the Test category with only 10% weight, was the only place where she explicitly tested their vocabulary skills, and she did it through presentational speaking, which was only half of the test, or 5% of the overall gradebook. But she said, "The kids really, they like the tests, and they know their stuff . . . But they have to practice all– 'What? 200 words!' And I only pick ten, you know, they never know what they're

gonna get." Alicia realized the importance of straight memorization of vocabulary, but her low gradebook percentage of discrete-point vocabulary matched her belief that achievement was more than just accumulating knowledge sets.

Julia's gradebook revealed the highest percentage of discrete-point vocabulary assessments at 25.64%, but her implementation was totally different from the other teachers. She used translation for these vocabulary assessments, across all her categories, and none of the assessments required memorization or evidence of mastery. At the beginning of each of her units, she gave her students a three-column activity, with the new list of vocabulary words already written in the language in the first column. As a class, Julia gave them the translation in English to write in the middle column. Then students had to copy the new L2 word from the first column into the third column, next to the English translation. These six note-taking activities were recorded in the Julia's Quizzes category and accounted for 6.58% of her overall grade. She included discrete-point vocabulary every day in her warm-ups, too, saying, "It's very small. Five questions to translate, five sentences to complete, ten words to translate." These seventeen warm-up grades added up to 11.56% of the total grade. In her Projects and Presentations (and Tests) category, Julia included three open-note assessments, valued at 2.5% each for a total of 7.5%, that she called "Four Squares." For this type of assessment, the students filled in a grid with four columns with a vocabulary word from the list, along with the English translation, a drawing, and a simple sentence using the vocabulary word. She used conflicting statements when describing these assessments, saying on the one hand, "It is kindergarten work" that must be "nice, neat, and colorful," and on the other hand, she remarked, "That's a test grade," "to see if they can put it all together." At the beginning of the semester, she graded their misspelled words or skipping some of the grid boxes, giving a 100% "if they put some color in there and you see that they put some

effort." Later in the semester, she tightened up her grading, explaining that she wanted "to make sure that they know the order of the things." She said, "The last one is a little deeper. You cannot just say 'I like the park.' So maybe 'I like to walk at the park.' The vocab doesn't bother me as much. But the conjugation, yes." All these assessments addressed discrete-point vocabulary, but none of them required mastery via memorization. But Julia only required 34.69% of her gradebook assessments to be completed without using resources. This high percentage of discrete-point vocabulary assessment without showing mastery or memorization revealed Julia's belief that language learning correlated strongly with the accumulation of vocabulary sets, but that repetitive exposure to the language was more important than memorization.

At first glance, Diana's entire 25% Quizzes category seemed to be comprised of discrete-point vocabulary and grammar assessments, with a heavier emphasis on grammar at 17.5%. She broke three of her five quizzes into a vocabulary half and a grammar half. But Diana's communicative goal of using the language was revealed in these assessments by how she assessed the vocabulary using listening comprehension as the conduit. The students had to listen to her read a definition in the language and then write the vocabulary word that she described. By incorporating previously learned sentence structures and vocabulary in the definitions to assess the current vocabulary, this part of the assessment became more integrated. The students did not just have to recall vocabulary, they had to recall and process the meanings of previously learned vocabulary in context to produce the correct answer related to the new unit. This negotiation of meaning was Diana's priority across all her assessments, and it was not surprising that it showed up even in her seemingly discrete-point vocabulary assessments.

These teachers showed through their assessment and gradebook practices that they recognized to varying degrees that accumulating language skills required separating skills into

components like vocabulary and grammar. How much they separated the parts of the language and their reasoning for doing so depended on the beliefs that they held about student achievement. Their beliefs about memorization also influenced how and to what extent they assessed their students' accumulated knowledge.

**Belief: Achievement is Using the Language**

These teachers did not only believe that student achievement meant understanding concepts and accumulating knowledge. They also believed that students showed achievement by using the language. However, each teacher had differing levels of implementation in their grading, and differing perspectives on what constituted language use.

***Communication***

The ACTFL Communication standards that these teachers adhered to included speaking, writing, listening, and reading (The National Standards Collaborative Board, 2015). Laura, Diana, and Julia all used these terms as category labels, in various combinations. For Laura, her category was "Writing/Reading/Listening/Speaking," while Diana called hers "Reading/Listening/Writing/Speaking." Julia had two different categories—"Reading/Listening" and "Writing/Speaking." According to ACTFL, these communication categories could be further broken down into presentational and interpersonal speaking and writing, and interpretive listening and reading. The teachers in this study demonstrated a patchwork of internally persuasive discourses and classroom practices regarding what, how, why, and to what extent they implemented assessments for these communicative skills.

**Speaking.** Several of the teachers emphasized the overriding importance of communication both in their beliefs and their grading practices. Daniel was the strongest

proponent of this belief, demonstrated by both his statements and his gradebook entries. In his definition of a high-achieving student, he said,

> If the student can use the language anywhere, okay, and that's the goal. The goal is not only to master the language, but to use the language. Because sometimes, you may have somebody who masters well, who learns the language, but has some difficulties to use the language in a regular setting. And that's my goal, out in the world. Because in the classroom, your students can create something, etcetera. But when the student is on the spot, and then he or she has to use the language, this is when you know. And sometimes the student will come back to tell you, "I was there. And then I met somebody, . . . and I was able to communicate with that person!"

Daniel's strong focus on communication was echoed in his grading practices. In his interview, he was able to clearly articulate the reasoning behind his grading practices. He explained that even though the class that he was discussing was an intermediate class, rather than an AP class, he stated, "I always think about the AP structure . . . my summative assessment covers exactly the AP." The communicative requirements of the looming AP exam directed the decisions that he made about his grading categories and assessing, even for his intermediate levels. Therefore, he had an entire category called "Communication," worth 30% of his overall grade. In this category, he had one interpersonal speaking interview, two projects that intertwined presentational speaking and interpersonal speaking, and one project and presentation that could be presented half in English and half in the language. Each individual assessment in this category was worth 7.5%, which was higher than his individual test grade values of 6% each. He talked about one of his interpersonal speaking interviews and how he graded it, saying "It's production. They can produce. They can answer the questions . . . They can make a sentence. Nope, complete sentences. Complete sentences, and this is why usually I add a 'Why?' They have to explain." The students communicated individually with him during their interpersonal speaking interview, both answering and asking questions. But in his other

presentational speaking assessments, they also had to communicate with other students in a group setting, by asking questions to the student presenter, who had to answer those questions at the end of each project presentation. This large Communication category and the assessments that he assigned in it showed his strong belief that language was communication, more than accumulating knowledge or understanding concepts.

Diana's class level was intermediate like Daniel's, and she also kept the AP test in mind when assessing her students, "I show them the AP rubric when it comes to speaking, and I just lower the expectations a little bit. And that's how I grade it, so that they need to expand and elaborate." Even still, she found that students were "shaking and crying," with high "anxiety," and "a lot of drama" during her student-to-student mock speed-date, worth 6% of their overall grade. She found it hard to implement this modified AP rubric, saying to the students that "the most I can do for you is a 70," "if at least I see that they tried, you know, they gave me a basic sentence." While she intended to follow the AP guidelines for interpersonal speaking, implementation in her actual classroom context resulted in having to lower, and then re-lower, expectations for the students because of the reality of the students' skill levels and psychological states. Besides this interpersonal speaking test grade, Diana also included a Flipgrid video, a student created dialogue, and a speaking portion on her final exam, which were all open-note presentational speaking assessments, valued at 9.08% of her overall grade. Together, her total for communicative speaking was 15.08% of her gradebook, which demonstrated a priority for using the language.

Julia also incorporated communication in her grading, with four interpersonal speaking interview quizzes (5% of overall grade) that prepared the students for two larger but similar speaking tests, worth another 5% of the overall grade, for a total of 10% in interpersonal

speaking. In this level one class, she explained that she graded on whether the students could understand and answer the “very simple questions” like “How are you? What is your name? What day of the week is today?” Rather than expecting expanded production like Daniel did, she explained, “And if I ask you what is your middle name? And you answer me with your last name, then that's wrong.” She had four smaller speaking quizzes that prepared the students with the same questions for her two larger speaking tests that would become more complex as the semester progressed, saying, “So that would increase.” This practice of building in complexity through familiarization with the questions reflected her definition of a high-achieving student as someone who, “By the end of the semester, if I ask them a question, they will feel really comfortable.” Julia also required her students to communicate with each other on an assessment where they asked and answered simple questions about their school life. This added another 1.25% to the 10% for a total overall interpersonal speaking percentage of 11.25%. Her interpersonal speaking communication was very basic for her beginners, but it was still language in use.

Laura also included communicative skills as part of her definition of high-achieving students: “They don't mind making mistakes, just so they speak, and they feel more comfortable.” When describing her grading procedures for the speaking section of her final exam, she explained, “Just as long as they can communicate, you know. That's, that's my goal for them to be able to–those little things can be adjusted. The fact that they get over the fear, that they are able to communicate.” This portion of the final exam plus four other presentational speaking assessments in her gradebook added together to account for 16.64% of her overall gradebook.

Interestingly, Laura only included one interpersonal speaking assessment that would require on-the-spot comprehension, negotiation of meaning, and circumlocution, rather than just pure regurgitation of memorized chunks of language. This assessment was worth only 1.39% of the overall grade. She said she found it hard to administer interpersonal speaking assessments "on every single one of them," because of her large class sizes. While her high percentage of presentational speaking correlated strongly with beliefs about achievement being accumulated sets of knowledge, her emphasis on assessing memorization and discrete-point vocabulary and grammar concepts did not relate well to interpersonal speaking assessments. However, she self-reflected during her interview, saying, "I can do more quizzes and tests like speaking," adding, "Maybe I could change some of my assessments to pull them up and be like, 'Yeah, we're gonna do a speaking assessment real quick.'" These ideas to increase the amount of interpersonal speaking in her classroom pointed to a possible evolution in her beliefs about student achievement. She was self-aware about her classroom practices and realized, without having it pointed out to her, that interpersonal speaking was an important aspect of using the language for communication that she was lacking. Her internally persuasive discourse was in the process of changing as it interacted with the contact zone of her classroom.

Martin had a high gradebook percentage for interpersonal speaking at 15%. However, this was for one single assessment, which he did not realize was only one of two assessments in his Summative category. He stated, "From a numerical point of view, no, I didn't think about that deeply . . . There was no plan from that point of view." But he also admitted that he was "doing so many different things for the first time, particularly this spoken thing." He connected this attempt at a speaking assessment to hearing his colleagues discuss their interpersonal speaking assessments and seeing them being conducted in the hallway. His phrase during the interview, "I

will take them out in the same way as other languages do" indicated that he was strongly influenced by the World Language team collaboration meetings to try something new with speaking. But his purposes in trying this new assessment were slightly different than his colleagues, who primarily used it to assess communication skills. Instead, he used interpersonal speaking to determine, "Are they keeping up with their work? And, secondly, how they are thinking [in the language]?" These two questions reflected his beliefs that student achievement was both work ethic and grasping the language as a concept. He recalled doing several other interpersonal speaking assessments, but they must not have made it to the gradebook, and he also confessed that they "tapered off towards the end of the semester." This intermingling of his own internally persuasive discourse about work ethic and conceptual understanding of language with the discourses of his colleagues about communication produced an emerging new assessment that superficially mimicked his colleagues' methods of administering interpersonal speaking assessments, but actually evaluated a totally different construct for totally different purposes.

Carmen never included any interpersonal speaking and allocated only 1% of the total grade for presentational speaking. She mentioned that she had recordings of the students from the language lab, saying "I do Q&A with them. I have evidence of what they can produce. But it wasn't something I wanted to grade and have it reported here. Yes, they would speak. It was . . . it was off." She taught a beginning level, but said the students struggled to answer questions like "What's your name?" She blamed it on "anxiety," saying that "producing language" caused the students "anguish." She recognized her conflicting beliefs and practices regarding interpersonal speaking, explaining, "The four domains are very important. I say it on the syllabus. But it's not reflective in my gradebook." But she also "didn't feel it would be helpful" to record the students' grades on the interpersonal speaking. She acknowledged this hole in her gradebook as a

weakness, saying, “I need to assess them and make it a priority, so that they know this is– to take it seriously,” while also demonstrating her hesitation to include interpersonal speaking as a formal grade in the gradebook, saying, “I might just make it a quiz, that way, it's slightly diluted with the other quizzes, but it is an assessment that they will take seriously.” This idea to dilute interpersonal speaking in the gradebook hinted that this type of communicative assessment conflicted with her belief about student achievement. For Carmen, student achievement meant understanding concepts, rather than communicating in the language, and interpersonal speaking was not a metric that she chose to reflect her students’ conceptual understandings. She felt the pressure from the authoritative discourse about the communication standards for speaking, but her internally persuasive discourse caused her discomfort in implementing communicative assessments.

**Listening.** Assessing listening skills were also important for teachers who believed that language achievement equated to using the language to communicate, even if they expressed tension in their beliefs about assessing students’ comprehension. As mentioned previously, Daniel, Carmen, and Alicia left listening assessments off their final exams, both for expediency’s sake, but also because they had concerns about the students’ possible “bad grades” (Daniel’s words) and how “stressful” the assessments were, according to Carmen. Diana included listening on her final exam, but it only counted 1% of her overall grade, and she assessed the students before the day of the official final exam to save time.

Diana’s gradebook revealed the highest percentage of interpretive listening assessments, allocating 20.8% of the overall grade to the development of this skill. One third of all her tests, and the first half of three of her five quizzes were all listening assessments. Diana believed that listening comprehension was difficult for some students, as evidenced by the way she talked

about a classwork activity, “This one, that is hard. Oh, like for the lower level, they're like, ‘What is going on?’ They're completely lost, because they speak faster.” She wanted the students to hear people from “different countries. It’s different accents. Some people speak slow, or some people speak faster.” But she found that this listening activity unveiled a dividing line between her high achievers and her low achievers. The high achievers tried to guess where the differing accents came from, but the low achievers said, “What? What accent?” She laughed, “They don’t even hear the accent.” She said that all the students found the activity “really engaging,” saying, “they like to do that one.” Nevertheless, some of them struggled.

Diana’s belief that listening comprehension was difficult for the students also appeared when she described her testing procedures. Before each of the tests, she played four audio selections from the textbook assessment package for the students, before choosing two of the four to include on the test. She explained that by doing this, the students were “not as anxious,” because they had “been introduced to it,” saying, “at least they’ve been exposed.” She changed the questions and the question format for the test so that students still had to show their comprehension of the audio selections, but their prior exposure to the audio made the selections accessible to all the students, both high-achieving and low-achieving. The listening portions of the tests and final exam were only worth 5% of the overall grade, and the quizzes were valued at 7.5%. But Diana focused on preparing the students for these summative assessments with the 8.32% of listening that appeared in her classwork category. This large percentage showed a concerted effort to train students and assess them formatively in listening comprehension, so that they could become better communicators.

Daniel’s emphasis on communication was also evidenced in the high gradebook percentage that he included for listening assessments, which was 13.75% of the overall grade. He

believed that for the students to understand their interlocutor in a conversation, they had to practice "listening, listening, listening," because he said, "At the beginning it's gonna be 'Oh, what did they say?' and, 'I could not understand!'" He waited until they had practiced for several months and then "started making it more serious, because at the beginning, it's like, okay, they do it. Yes. But now it's a quiz." He said that his listening practices for the students paid off: "You will see the improvement," and "Now they will listen and understand exactly what they are saying." He wanted them to be able to understand the language in and out of the classroom, as was his stated goal for his high achieving students.

Carmen's overall listening comprehension weight was recorded in her gradebook at approximately 8.5%, but she said that she assessed students frequently in class, saying, "I do feel that I have assessed their listening ability." She used the textbook video series and other online activities as a way for her students to learn to process the language for comprehension. Each of her tests had a listening section, where Carmen played the audio selection for the students three times, adding, "I slow the speed down a little bit." The questions that the students had to answer about the audio were true or false statements, which she said, "That's pretty much what they're able to–capable of doing," implying that more complicated types of questions might have caused her students issues. The repetition, slowing, and simplifying of her listening assessment method showed that Carmen believed that the task was difficult for the students. But her consistent inclusion of listening on unit tests and the weekly formative practices in class showed that she valued listening as an important aspect of language learning. Nevertheless, the fact that the students were rarely required to apply this listening skill to navigate interpersonal communication tasks pointed back to her belief that language achievement was understanding concepts and accumulating knowledge. Her binary assessment method of true and false questions

allowed her to find gaps in her students' understanding that she could address during her weekly Wednesday remediation sessions. She said on those days that she would "take one of these listenings from the test, and I use it for a warm-up the next week." Using the listening text, she would "practice the reading strategies of underlining" with the students. She knew they had to use underlining in their other academic classes to prepare for end-of-course-tests. She said she tried to "encourage that strategy," showing her belief that improved metacognition would aid the students in their understanding of the language. The listening portions of her tests were therefore Assessments *of* Learning, Assessments *for* Learning, and Assessments *as* Learning, because of the way she used them in the classroom.

Julia also assessed the students on their listening skills informally and frequently throughout the semester. Each of her seventeen warm-ups included listening to a song every Friday and completing activities. These songs were a way for Julia to incorporate culture into her classroom, while also connecting them to the outside world. She described a high achieving student as one "that looks for more outside of the classroom." She said that this type of student would come into class after being in the community saying, "'I heard this music! Listen to this! Do you know this song?' or 'I saw this on Netflix. Have you ever heard of it?' They're very curious, and they always bring something from the outside." Exposing her students to songs in her language was one of the ways that Julia was "trying to find connections between what he or she learns in the classroom and outside the classroom." She hoped to encourage the students to be "always interested and not just comfortable with what I offer in the class." She offered so many songs so that her students would begin to hear the language used in the community and make connections to what they were learning.

But for her other class time listening assessments, Julia said, “I don't put it in the grade book, it’s more like an everyday thing, and I see how it is. And then, normally, they're pretty good.” However, her first test of the semester was a listening test for students to write words that she spelled aloud to them, after practicing the letters of the alphabet. Her last big assessment before the final exam was also listening, bringing her gradebook weight for listening to 7.8% total.

Laura’s reticence to assess interpersonal speaking matched the absence of listening comprehension assessments in her gradebook. Laura only assessed listening comprehension once, and it was on a district asynchronous day where teachers had to post assignments online for their students to complete at home. She said she wanted them to “practice and listen and try” to answer the questions that she only graded on completion rather than accuracy. This meant that her only listening assessment, worth 1.39% of her overall grade was not even administered in the classroom. These two gaps in Laura’s gradebook—listening comprehension and interpersonal speaking—revealed a possible conflict between Laura’s internally persuasive discourse about the importance of communication, and her classroom practice as it appeared in her gradebook. Her other competing internally persuasive discourses about student achievement could also have interfered, since for her, accumulating chunks of knowledge and understanding concepts were also high priorities, which could be harder to assess through communicative tasks like listening comprehension and interpersonal speaking.

**Writing.** Daniel’s communication priority extended to the open-ended, closed-note written portions of his unit tests, where students had to answer a personal discussion question related to the unit without access to any resources. These communicative writing portions on his tests added 10% to his 30% spoken Communication category, for an overall communication

percentage of 40%. He did not use a rubric to grade, saying, “I did this on comparison,” because he said, “You don't expect the students to be able to write everything correctly.” He was not looking for perfect accuracy in their writing skills, but whether they could communicate based on their differing “prior knowledge.” Their grade was “based on the class, and . . . on the students’ performance,” saying “This is subjective” and “This is why we have this differentiation.” He measured their written communication by what had been taught in the class compared to their individual abilities and backgrounds, measuring their progress in the learning process, rather than just their written product.

Like Daniel, several other teachers prioritized communicative writing. Diana’s overall percentage for open-ended writing assessments was 21.66%, and for all but 4% of those, the students were not just allowed, but required to use their resources. “The most important thing for me at this level—well, that it makes sense, of course, but that you use the information from the sources,” because she was thinking ahead to prepare her intermediate students for the synthesis-style essays that they might one day write on the AP exam. She said, “There's no excuse for not using the sources,” because they could cite directly from the readings that were on their desks.

Laura also required several open-ended writing assessments in several of her gradebook categories, with an overall percentage of 17.89%. To grade the written portion of her final exam, she used a “small rubric—Did they stay on topic? Was it understandable?” She wanted them to finish the semester with confidence, without “fear” of using the past tense.

Julia’s gradebook included 16.25% open-ended writing assessments, although six of her twelve assignments were completed in partners or groups. She found that one difficult aspect of allowing students to use resources was that “They go on Google, and they come up with these crazy sentences!” This was one of the reasons that Daniel allowed his students to write their

country presentations partially in English, so that they would not create gibberish sentences. Allowing students to write creatively in an open-ended format sometimes made it difficult to assess their actual writing skills because of the resources that the students accessed during the process.

Carmen and Martin, who both prioritized conceptual understandings of their languages, used open-note writing frequently in the classroom, but they assessed open-ended writing at very low percentages. Carmen only included one assessment, with 1% of the overall grade, but the focus was on accuracy, including spelling and accents, rather than on communication. She also included some interpersonal writing questions on her warm-ups, which were 5% of her overall grade, but again, she graded on accuracy. Martin also had his students write to make choices about word endings, but he said, "they would never create" a paragraph from scratch in the language. He used the writing activities to guide them in their grammatical thinking about the framework of the language, rather than as a way to communicate.

But Carmen was the strongest proponent for using writing in her assessments, with a focus on writing in complete sentences, repeating the phrases "I want them to write it out" and "Write the whole sentence out" multiple times across multiple grading categories. She used multiple choice and matching formats on some assessments, but mostly she wanted them answering both closed- and open-ended questions in complete sentences. This was her way of seeing if the students understood the concepts of the language, saying, "You gotta use the formula." But it also ensured that the students memorized the spelling and accents through repetition. She repeated some of the same questions and the same written phrases throughout the semester, "just to reinforce," looking for "ways for them to practice and reinforce it, but they can't make it so they reinforce it incorrectly." She believed in providing quick feedback on their

writing, saying, “Once they fossilize something incorrectly in the brain, it is hard to undo it.” Writing, especially in complete sentences, was a way for her to accomplish her goals of increasing students’ conceptual understandings and helping them store language knowledge in their memory. For her, writing was not a way for students to show their communicative abilities to express their own ideas.

**Reading.** The teachers’ beliefs about student achievement emerged through their use of reading comprehension assessments. Embedded reading comprehension undergirded many of the assessments found across all the gradebooks. Several teachers used a combination of interpersonal reading, which was understanding and answering written personal questions, and interpretive reading, which was reading fiction or nonfiction passages, and being able to demonstrate comprehension of the passages to use the language in differing ways. Interpersonal reading often formed the basis for assessments that involved interpersonal speaking and interpersonal writing, which have already been discussed. Therefore, this section will address assessments containing interpretive reading passages, particularly with regards to the teachers’ purposes in using them.

Teasing out exact gradebook percentages for interpretive reading comprehension was more difficult than for other skills, unless the teachers labeled the assessments as such or discussed them explicitly in their interviews. Alicia and Daniel both explicitly labeled reading assessments in their gradebook at least once. Alicia included five cultural and historical readings totaling 8.3% of the overall grade. The students read from the textbook and answered a packet of questions to demonstrate their understanding of the text. She said students would “boom, hand it in,” just to get the completion grade, or she said they would “just copy somebody,” without processing the meanings of the texts. She had to tell them, “Take your time,” “Read it over,

because some of these questions were on the final exam. And it tells me if they recall, remember." 50% of her final exam consisted of questions about these readings, bringing her overall percentage of reading comprehension about history and culture up to 13.3%. Alicia believed that reading comprehension was a way for students to learn to appreciate the culture of the language that they were acquiring.

Daniel only had one clearly labeled reading comprehension assessment, and it was only worth .4% of the grade, graded on completion. But his final exam was only reading comprehension, and he included reading comprehension passages on all five of his Summative Tests, bringing his overall percentage to approximately 24%. He explained that both the readings and the questions about the readings were in the language, complaining that "The county now is giving everything in English . . . which I don't get." His emphasis on using the language for communication precluded him from using English questions to probe for understanding of the language. He wanted the students using the language all the time, including when they were demonstrating their interpretive reading comprehension skills.

Martin, on the other hand, used translation from his language to English as his chief metric of students' abilities to read passages and understand their meaning. His gradebook percentage for reading was approximately 22%. He said his students would examine a "word to come up with a meaning." But if they did not recognize "the two different parts" of the word, that they were "probably going to come up with their own meaning," which would be an incorrect interpretation of the word. He used reading comprehension via translation to assess their understanding of the "framework of the language," to see if they had learned "what the endings do . . . and being able, therefore, to manipulate it, to manipulate the language and use the

language." Martin believed that students' reading comprehension skills were tightly bound to their understanding of the underlying concepts of the language.

Carmen, like Daniel and unlike Martin, did not use English to assess reading comprehension, saying "I don't do matching translation." Instead, she used reading comprehension passages on her assessments as a means of producing more student writing, insisting, "I need you to answer with full sentences," and "I need complete sentences, I stress that a lot." By having the students write, she was reinforcing through repetition the grammar and vocabulary concepts that she was focused on, to "encrypt" them on their brains. Each of her three tests contained two reading comprehension passages, valued at about 12% total, combined with her 10% final exam which was "lots of reading passages; it's mostly reading passages," to add up to approximately 32% of her overall grade. Carmen later reused these passages on subsequent warm-ups to strengthen her students' metacognitive skills. She explained the follow-up activity saying, "We practice the reading strategies of underlining when you feel you see the answer." She said, "I try to encourage that strategy," mentioning that this metacognitive strategy came in useful for the students on their end-of-course tests in their core subjects, and that it could help them in their second language reading comprehension as well. She would go over the passages with them, saying, "This is what I was asking you all to do. Let's look at it now," trying to get them to "notice patterns in the language," which was how she defined high student achievement. Reading comprehension assessments became a conduit for her to reinforce their accumulation of knowledge through writing and their understanding of concepts through analytical thinking activities.

Diana also used a long reading comprehension passage to produce writing for a test grade (6%). For this assessment, the students had to synthesize the reading passage with a video to

write their own argumentative essay. For this essay they were not just encouraged but required by the grading rubric to cite their sources in an open-note format. This and one other small classwork assignment (1.67%) that had "short readings" were the only two clearly labeled interpretive reading assessments that appeared in her gradebook, bringing her overall percentage to 7.67%. When describing her final exam, she said, "There used to be some reading comprehension, but I took it out because . . . they were spending too much time, and I was like, no, you know, it's the last day. I just wanted that boom, boom, boom." She had seen students struggle on past exams, and she wanted her current students' scores to reflect their actual class achievement. It was also possible that she had done so few formal interpretive reading comprehension assessments during the semester that she felt that they would not succeed in the multiple-choice format of this skill assessment on her final.

Julia's interpretive reading comprehension showed up in her warm-ups, on two of her tests, and on her final exam. All of these categories brought her total reading comprehension assessment percentage to 31.68% of her gradebook. Her students actually read on many more assessments than these, but they were reading aloud, graded on their pronunciation and "comfort" in the language, without any accountability for their level of understanding of the texts that they were reading. Julia's big emphasis on recitation and reading aloud will be discussed in a later section.

Laura's overall reading comprehension percentage was 22%, spread across all her categories, including her final exam where the students had to "figure out some grammar or vocab" from the passages that they had to read. However, she used reading comprehension on one of her quizzes to push her native students to use "academic language," explaining, "They want to put fifty million vocab words that they know," instead of using the new vocabulary that

she wanted them to acquire. She attempted to expand their vocabularies by providing a word bank for them to complete definitions that were all in the language, to force them to use the new vocabulary that was less familiar to them. This assessment reflected her priority for accumulating knowledge, even for her native speakers that already had a baseline of daily useful vocabulary. Furthermore, on one of her tests, she used a reading passage with missing words for students to have to make grammatical decisions about the correct usage of the past tenses based on context. In this case, she used the reading comprehension assessment as a means of checking for their conceptual understanding of the proper uses of the verb tenses. This conceptual understanding was strongly related to her beliefs about student achievement.

While all the teachers believed that students showed achievement by using the language to speak, write, listen, and read, as evidenced by their gradebooks, their assessment purposes in administering them differed significantly. Martin, Carmen, and Laura used these communicative tasks to assess their students' comprehension of the underlying frameworks of the language. Laura and Carmen also used their assessments to show students' accumulated knowledge. But some teachers had other priorities for their communicative tasks.

***Negotiation of Meaning***

Diana had a very specific focus when it came to communication. When grading her closed-note, open-ended essays on her summative tests, she evaluated the students by asking herself, "Were you able to communicate? And did I understand the response? And that's all. That's what it matters to me, if they were able to express, and it makes sense." She said that they needed "to expand and elaborate" and that she graded them by giving them more points the more they elaborated. But beyond just simple communication, she also used the phrase "makes sense" multiple times during her interviews, along with "What do you mean here?" and "Do you

understand the question?" She started to push the students to understand that language is nuanced, and that the way the students used vocabulary and grammar could cause their interlocutor to misinterpret their meaning. For example, on several quizzes and tests, she made the students translate their own sentences back into English, saying, "When they translated it, it was something completely different to what they wrote. So they have to write it and give me the translation, and sometimes it doesn't match. Like in their head, they think it's something else." She used both the students' first and second languages to negotiate meaning when writing in the second language. On another presentational speaking assessment, she noticed that "they have to know their vocabulary in order to make sentences that make sense . . . They just start putting random things together, and then they try to translate." She was concerned that "It's not gonna make any sense." Sometimes a lack of vocabulary caused this lack of sense-making, but other times, she said that this "random stuff" was due to the fact that the students "do it in a rush, and it doesn't make any sense." She found that forcing them to "give the meaning" made them slow down and think through exactly what they wanted to communicate. On another assessment where she was preparing the students for student-to-student interviews, she wanted to make sure they knew what they were asking and answering, so she sat down individually with each student to help them renegotiate the meanings of the answers that they had prepared, saying, "This is how you should say it." Her assessments pushed students to use the language to communicate, but she further helped them negotiate the nuanced meanings of the new vocabulary and grammatical structures that they were using.

***Participation in a Community***

Alicia was left out of the prior discussions about speaking, writing, listening, and reading, because she had very unique beliefs and practices regarding communication. Alicia echoed the

other teachers on the importance of using the language, saying, “Communication comes in most, they have to learn to communicate.” But for her, high achievement meant being “incorporated with the class.” In her class, “Nobody’s left out,” even if she found that high achievers tended to be “isolated,” and that both high- and low-achieving students were “loners.” She felt like it was her job to “redirect them,” to “try to teach them to collaborate as a group,” and to “build that self-esteem.” To make sure that high-achievers maintained a strong bond with the class, she would “pair them up with another. ‘Hey? Can you help so and so?’ ‘Sure, no problem!’” This helped all learners learn to depend on each other.

To increase this group interdependence, Alicia made sure that her students had to use the language to survive in the classroom. For example, if they were absent and needed their make-up work, they had to express their need in the language before she would help them. And in her everyday classroom instruction, Alicia communicated only in L2, and made her students interpret for each other. She would ask, “Okay, what did I say?” And she explained, “They help each other . . . They speak it, because they interpret what I'm saying.” The students started to form a collaborative community of learners because they were forced to assist each other to understand what was happening in the classroom. As further evidence of this hive effort, part of her ungraded formative assessment was calling on individual students in class to respond to her questions:

> If they get stuck, they have to turn – “Help! Can you help me?” And that's the main thing. And mainly I do have some people that don't want to say anything, you know—rebellers. I said, “Okay, I'll wait.” Then slowly, they'll turn around and ask, you know, because their self-esteem–these kids have no self-esteem, period. It’s what we're building up on. And then slowly, they're getting there, and then they warm up, and then at the end of the year, they don't shut up.

Essentially, the students were working together to give her the answer that she was looking for, but they were learning communication skills and building a collaborative community at the same time. By not formally recording these integrated interpretive listening and interpersonal communication opportunities in the gradebook, Alicia created a class community where everyone helped each other. She did not want graded interpretive listening or interpersonal speaking to be separate entities from regular class participation by assigning a numerical rating, but rather sought to build their self-esteem and their confidence by using the language with her and with their classmates. She even talked about applying for a grant for new rolling chairs for her classroom, "so they could turn around, 'See? What did he say?' Instead of just me." She wanted them to be able to swivel quickly to be able to communicate with their classmates for her to be able to ask them, "Johnny, what did Susie say?" to encourage communication among the members of her class community. Her gradebook reflected this emphasis on natural ungraded communication, with only 8.13% for interpretive listening and 6.25% for interpersonal communication. These low percentages of graded participation encouraged student engagement in the group where everyone was working together to be successful in the language.

Alicia's focus on community building was also evidenced in her gradebook, where 42% of the students' overall grade was either practiced or performed as partners or groups. Her "favorite" assessment was a group project that was graded using a group rubric, with no individually graded components. All the students had to do their part, "They had to figure out." She would ask them as they prepared the project, "You guys have a backup plan? It's your decision as a group." She reasoned that it promoted "collaborated brainstorming, helping each other." Other presentations were practiced and informally peer-graded as partners, to help them

"understand the rules and requirements" of the projects. Some of the presentations were presented as dialogues between two students, which Alicia said she liked because students were communicating with each other.

Alicia was the teacher with the highest percentage of presentational speaking with 47.5% of her overall grade. Alicia used these assessments as opportunities to build relationships with her students. Her very first grade entry of the semester was an "About Me" Flipgrid video that she used as a "warm-up," but also as a way for her "to get to know them." She said, "I really like these." They were in the language, and she used a rubric to grade them, but she said she was "flexible," since they were just starting their language learning journey. Her entire Performance-Based Assessment/Quiz category, valued at 30%, consisted of individual and partner sentences and dialogues presented in front of the whole group. During the preparation process, she worked with each student one-on-one, using Socratic questioning to encourage them to look over their work to improve: "Is this a noun or verb? . . . Do you think this is right?" She explained her reasoning for this technique, saying, "So I kind of empower them. They correct themselves." While most of the time, she gave feedback in the large group class setting after these individual presentations, she would also give extra private feedback to her struggling students, "It depends on the student . . . You need to know your students." Even during the group song project, Alicia went "group by group," giving just-in-time instruction as needed, but also keeping all the students on task, "I noticed some people, they'll just sit back. And I'll say, what are we doing?" She said she would also "speak up for" the students who would not complain about doing all the work in the group. She knew her students well, because many of her assessments provided a means for her to build relationships with them.

Another important factor in creating Alicia's community of learners was the public feedback that students received when they had to present in front of the class. "We do it in front of the class, because we all learn from each other, and that is the purpose." Students received public feedback from Alicia on 30% of their grades in the gradebook. When students fought against this, she reminded them that she had to get up in front of them every day, and that if they were going to refuse to present to the class, that she would stop presenting to the class, too. She had to remind them that "We learn from each other, you know, kinda empower them. They're like, 'Oh, okay'. . . But you build that self-esteem. You'd be surprised. The end of the semester. They won't shut up." Alicia felt like she empowered the students, too, because she said they taught her as much as she taught them. The community that she created by forcing survival skills and collaborative learning through group projects with public feedback was her conduit for having the students use the language to communicate.

***Recitation and Reading Aloud***

Julia was another outlier in the way she incorporated language use in her classroom assessments. As mentioned previously, Julia administered several interpersonal speaking assessments, valued at 11.25%, in her Projects and Presentations category and her Writing/Speaking category, with the goal of communication. But she also had a category called Reading/Listening valued at 15% of the students' grade. Four of the five assessments in this category consisted of students reading dialogues from the textbook, out loud in the hallway with a partner. According to Julia, these dialogues were found at the beginning of the textbook units and were meant to introduce the new vocabulary and structures. Julia graded the students on their ability to read the language out loud, saying, "If they feel like they're confident, and I understand what they're doing, I give them a hundred. There is not much of like a 50 or a 60 . . . They feel

very confident." Julia did state that her expectations for the students would grow as the semester progressed, "At the beginning, they do it from the paper. Towards the end, they memorize it." She also expected their confidence and fluency to improve throughout the semester: "The beginning is very like, 'Oh, I'm not so sure.' But at the end I want you reading as if we were talking." Each of these open-note mini dialogues, valued individually at 3% of the overall grade, counted more than the closed-note interpersonal speaking tests in the Project/Presentation category, which were each valued at 2.5% of the overall grade. This meant that 12% of the student's overall grade was reading mini dialogues from a textbook in the hallway, which was more than the value of the final exam that had a weighting of 10%. Julia's gradebook weight showed that she had a strong belief in the usefulness of reading dialogues out loud as a means of becoming comfortable using the language.

The students also did two memorized solo recitations, each worth 1.25% of the overall grade. Julia graded these on "how they pronounce. I'm always on pronunciation, how they pronounce." And they created and read five of their own dialogues with a partner throughout the semester, each worth 1.25%. For the first one, Julia graded on "how comfortable. It's the first one, and it's like, you completed it, used the structures, you're trying to pronounce correctly. Then it's like a zero-or-a-hundred type of thing unless it's a disaster." The other four were graded on pronunciation, but also on grammar and structures. The last student created dialogue of the semester had to be memorized, but the other ones were read from their papers in the hallway to Julia.

These student-created dialogues and recitations added another 8.75% to the other 12% from the textbook dialogue readings, totaling 20.75% of the overall grade for presentational speaking. This percentage was weighted twice as much as the total of her closed-note

interpersonal speaking, which only added up to 10% of the overall grade. This high percentage of grade-weighting on reciting dialogue reflected her stated characteristic of a high-achieving student: "By the end of the semester, if I ask them a question, they will feel really comfortable and answer like something that we saw. They will feel really comfortable." Her belief was that the students would recall some of the dialogues that they saw and read throughout the semester and would be able to comfortably use them to answer questions if they were asked. Therefore, she demonstrated through her grading practices that she valued language in use, but through the conduit of assessing recited dialogues and listening to her students reading aloud.

**Belief : Achievement is Understanding Culture, Connections, Comparisons, and Communities**

Communication, as discussed in the previous section, is the first of the five"C's" on the ACTFL *World Readiness Standards for Learning Languages* (The National Standards Collaborative Board, 2015). But most of the teachers in this group also addressed the other four "C's" of Culture, Connections, Comparisons, and Communities in their assessments, to varying degrees.

***Culture, Connections, Comparisons, Communities***

Martin's gradebook only had three postings that were specifically labeled as cultural or historical topics, but he explained in his interview that culture, history, and mythology comprised one third of all his weekly assessments in both his Formative and Summative categories. His Final Exam project was a cultural interpretation of anything the students wanted to portray, that demonstrated "a connection to the class." These cultural assessments together added up to approximately 40% of his gradebook. For his final exam project and for two of his other projects in his formative category, he explained, "I would make it into an open-ended presentation, and I

would say they could use any medium they wanted." He wanted to "try and give hands-on" ways for the students to access the culture and history, while also acknowledging students' various talents and interests, saying "differentiation comes into this a lot." His goal with these open-ended, hands-on projects was to "empower them," and to "try and take the heat off by giving them a kind of blank sheet to work with." He told the students, "These are the tools . . . Okay? Just go for it. Whatever you do, I'll be happy to see your interpretation." While he would insist that the projects contained certain "elements," he did not grade on a rubric. But he also did not base their grade on "creativity," saying, "I may love words, but I'm certainly not a craftsperson." He said sometimes students would "come back in five minutes and give me the finished product. Then that, to me, doesn't show any particular effort or forethought at all." He said he would give this type of student "a 70" "because they produced something, they did fulfill the terms of the assignment, certainly not spending any time in doing it." Martin defined a high-achieving student as having "a good work ethic," and a low-achieving student as one who did "minimal work." He looked at "the big picture" of the students' achievements, because for him, "One of the big values or the big importances of high school is that they do two things: One, they identify their unique skills and talents and to build on those. And secondly, to develop a work ethic." These cultural projects provided him the means to push the students to discover their talents and to strengthen their work ethic by insisting on maximum effort.

On his weekly review assessments, consisting of grammar, some cultural or historical content, and some translation work, Martin included modern movies "just to show a kind of crossover," to make the students' learning relevant, "to try to cover all of those connections as background." He wanted to "make some connections between the two," between the language and its culture and the world that his students lived in. His crossword puzzle clues sometimes

came from the cultural content, which he graded on "accuracy of information." He explained, "They have to work on it to, quite hopefully, in the process, memorize it to a degree, to find the right answer. It is typically not that easy, that straightforward. Again, you're talking about effort." He wanted them to recognize the cultural connections and comparisons between English and the language. But he was also pushing his students to develop their work ethic through having to be accurate in their daily work, as well as creative and resourceful when producing projects. Through these assessments, Martin demonstrated his priority of making connections and comparisons between his language's culture and the students' modern pop culture.

Like Martin, Alicia had a high gradebook percentage of cultural assessments, with 25%, primarily in her Classwork/Participation category, but also spilling over into her Final Exam and Projects categories. She gave her students five assessments related to cultural videos that they watched in class, a combination of movies, documentaries, and YouTube activities. She also assigned them four textbook readings with a packet of questions to answer, graded on completion. She used some of the same questions from those videos and readings on her final exam. When describing one of the movies, she said, "That was culture and impact, history and community, all of that." When they discussed the movie as a class, she found that making them think with a new perspective caused them to "really enjoy the language more." But Alicia did not just focus on culture, connections, and comparisons, she wanted her students to feel part of the language's community. She considered that all of the four "C's" of the standards tied closely together, saying,

> The community, they get exposed by the connections to the real world. Those are the two main things, when you get to [levels] three and four. [Levels] one and two–it's just the surface. [Level] two, a little more in depth . . . So there's a different level of community at each different level you expose them. Get the taste of, 'Oh, what's community?' . . .We have to base it on history. History is the main thing, you know, 'cause if they don't have that, they're not gonna understand. That's why they watch those movies. They have

> questions. That's why we discuss about that, you know, with today's society, the world today.

These readings and movies tied together to help students understand the culture and history of the communities that the students were becoming a part of by learning the language. But Alicia went further than just reading and watching to expose the students to the culture. She provided ways for them to fully participate. One big project in her gradebook, which had an accompanying essay, was a community simulation, where the students had to communicate as if they were members of the language community for an entire day. This event with the reflection essay added up to 7.66% of the students' grade, which showed Alicia's commitment to introducing the students to the culture and community, by assimilating them into the community itself. Alicia's classroom community that she created by incorporating all members through participation and cooperation in class assessments became integrated into the worldwide community of language users.

Alicia also talked about the ways that she planned opportunities outside the classroom for her students to connect to the local community through events like field trips, participation in graduation assistance, Veterans Day ceremonies, and events for special needs children. She was adamant about providing these opportunities, even when she was blocked by administration or logistics. These were not part of her gradebook, but they were an important external part of her classroom community, which strengthened the internal bond among the members.

Daniel and Diana both had comparable percentages for assessing culture, with 16.15% and 14.32%, respectively. Laura's percentage was lower, at 7.25%, but her purpose was in line with that of Daniel and Diana. Daniel assessed his students on how well they were able to communicate about the culture of his language through interpersonal and presentational speaking

assessments, a closed-note open-ended writing assessment, and a few smaller in-class assignments and quizzes to prepare them for the larger assessments.

Diana used a song as both a grammar lesson and as a jumping-off point for a class discussion about current events on one of her classwork assignments. Her other two assessments that focused on culture became opportunities for students to "expand" their open-ended writing skills. On one, the students watched a cultural video and compared it to their own culture by answering the question, "What would you do in that scenario?" And the other gradebook entry was an AP-style persuasive writing assessment that counted as a test grade, where students watched a video and read some articles to form and express their own opinions about a controversial cultural practice.

Laura's students researched a country for a presentational speaking assessment. She admitted, "I definitely did a lot more PowerPoints with speaking assessments with my native classes, you know. And they had a lot of more culture embedded in theirs, because it was a smaller class, and I was able to do that." Daniel, Diana, and Laura all incorporated culture in their assessments to provide fodder for further speaking and writing, which corresponded to their strong beliefs about the importance of communication as a hallmark of language achievement. For them, connecting with and comparing cultures was an important part of learning the language.

Julia used songs with videos every Friday as part of her warm-up, as a way of exposing students to important artists and countries in the culture of her language. She also had the students research one of the countries that speaks her language to color a flag, which she called, "kindergarten work." She also assigned a test grade for students to research a recipe, make it, and bring it in to share with the class. These three cultural exposures equaled 5.48% of her overall

grade. While this percentage was low, the consistent exposure to songs seventeen Fridays in a row correlated to her belief that high achieving students "always connect with outside" the classroom and are "always interested" in seeing and hearing the language in the real world.

Carmen did not specifically mention any assessments with culture embedded either in her gradebook or in her interview. This did not mean that she did not address culture in her classroom, but rather that it was not as high of a high priority. She did make connections with other academic disciplines, particularly on a "Math Practice" assessment that used numbers in the language to solve simple math problems, and on her warm-ups when she was working with the students to strengthen their reading comprehension strategies.

While all these teachers agreed that grades should reflect student achievement, their beliefs about what that achievement meant was a patchwork of borrowed, modified, adopted, and adapted internally persuasive discourses. No two teachers held identical patterns of beliefs about achievement as understanding, accumulating, or using the language, or in the importance that they placed on connecting to and comparing the culture and communities of the language speakers.

### Gradebook Categories that Reflect Academic Enablers

All the teachers graded students on their academic achievement, but they also included academic-enabling behaviors in their definitions of a high- or low-achieving student. Laura said, "Number one, they're driven. . . they're all over it . . . they're participating." Carmen agreed, "They take their schoolwork seriously." Julia added, "The high achieving student will be one that really works, is interested in learning." Martin concurred, "They do the work, and which means that they have a good work ethic." Diana took her definition a step further, saying that a high-achieving student is "not only engaged in the classroom, but also looks for opportunities outside

of the classroom. . . They're responsible for their own learning. So they go above and beyond what you give them. They're looking for more."

On the contrary, according to Martin, low-achieving students were "not producing daily work." Carmen stated that when things got challenging, these students had "trouble rising to the occasion and pushing themselves, they kind of give up. They don't always ask for help." Alicia did not hold back, "It's not the high and low. The lazy ones. . . It's not the high achievers . . . has nothing to do with that. It's the lazy ones. 'We don't wanna do nothing.'" In fact, Daniel explained to his students in his classes that if they turned in all their assignments, that they may not have a B, but they would probably pass at least with a C. He said that average students "may not be the best students, but the student is trying his best, you know, turning his homework. And as a teacher, our job is to help students like this, you know, because he's willing to learn." Julia was able to distinguish the high, the low, and the average students solely on how they completed their daily warm-up activities that were to be turned in every Friday, explaining, "The C Student, that one you can see a pattern—does it on Monday, doesn't do it Tuesday, does it Wednesday. The straight A does it every day. And then the failing F, just lost it— 'The warm-up? Warm-up? What is that?!'" These teachers believed work production to be a valid grading construct, and this work production showed up to varying degrees in their gradebooks.

**Classwork Completed Individually**

All the teachers had some kind of Classwork category in their gradebooks, and they all allowed students to use "open notes" while completing this work, meaning that they were permitted to access class resources like textbooks, vocabulary sheets, and grammar charts. Daniel had the lowest category weighting, at 10%, followed by Julia at 15%. Alicia, Carmen, Diana, and Laura all had classwork categories at 25%, and Martin had the highest category

weighting at 60%. Alicia mentioned in her interview that she was considering lowering her 25% weight to 20% in order to add 5% to her test category, which she felt was too low, because "it's overall," meaning that she wanted to tilt the grading construct more towards achievement than towards academic enabling behaviors. Carmen, Julia, and Laura all put "warm-ups" in this category, which were daily activities to get the class started and to review previously learned material. Carmen and Laura also spoke about using these warm-ups as behavior management tools. Carmen described the warm-up as "silent, individual. That's how I start the day." Laura would announce at the beginning of class, "Okay, the bell rang—warm-up! Get off your phones!" Classwork was a consistent part of all the teachers' gradebooks, which reflected their beliefs that quality work production was an essential characteristic of a high-achieving student.

**Classwork Completed as a Class**

Several teachers mentioned that classwork was completed as part of large group instruction. Martin in particular used notetaking as a third of his weekly grade. These weekly grades comprised 90% of his overall grade, so approximately 30% of his overall grade was notetaking. He said the students were expected to write down a minimum of notes, but that they could write as many phrases as they wanted. The students would then access these notes to complete grammar and translation activities, and that their success on these activities was "only as good as the notes that they made in the first place." He called this notetaking "regurgitation," but he explained that while the students were copying the notes, "I'm probably gonna be at the board and therefore working my way through examples," "teaching and making sure you got it right." In this way, his teaching process became part of his assessment, Assessment *as* Learning. On one of his summative tests, he described how he related this notetaking to achievement: "I did not need to look at their notes. I would know from the variety of questions I asked if their

notes were any good or not." By valuing notetaking at 30%, he revealed this strong belief that notetaking was a valid grading construct, as it enabled the students to perform better in related language skills.

Julia also included six notetaking grades, worth 5.64% of her overall grade, but she put them in her "Quizzes" category. For these six assessments, she gave the students a paper with three columns: the unit vocabulary in L2 in the first column. She then orally translated the vocabulary, and the students wrote the English in the second column, and then they re-copied the vocabulary back into L2 in the third column. But Julie also included classwork completed as a class in her In-class Activities/Homework category by means of her seventeen warm-ups, worth 11.56% of her overall grade. While the warm-ups were meant to be individual activities, "Five questions to translate, five sentences to complete, ten words to translate, always related to what we saw the day before," Julia admitted, "I always write the answers on the board, always, so literally it's copy work at one point, and then some of them still don't do it." Between the two categories, Julia valued large-group completion of classwork at 17.2% of her overall grade. She believed that constant exposure to the vocabulary in written form was an academic enabling behavior that would increase their learning.

Carmen only included one notetaking assessment in her gradebook. She said, "They had to take my notes. I was writing and drawing pictures with my notes. We found that if they illustrate their notes, you remember things a lot better." This notetaking was part of Carmen's belief that writing encrypted learning on the brain. She also went over parts of her warm-ups every week with the class. She only recorded four warm-ups in her gradebook, so the notetaking and the warm-ups added up to 5% of her overall grade.

Diana mentioned only two of her Reading/Writing/Speaking/Listening category activities that were completed as a class, totaling 4.16%. She said, "'Cause usually I go over the answers for them. So it's basically like . . . if they don't answer, then oh, I take off points." Daniel also went over his students' individual activities in class, but he graded them on completion beforehand, saying, "And this is my class policy. When I go over it, I'm not taking it, 'cause I mean, I cannot give you the answer, all the answers. What is the point? You just fill in!" Diana and Daniel put little value on notetaking or students completing work as a class.

***The Case AGAINST Classwork***

Laura had conflicting beliefs and practices about her classwork, a category which she called, "Reading/Listening/Writing/Speaking," weighted at 25%. In this category, she graded twelve warm-ups, some parts of which she would go over with the class, and some parts which they completed individually. "It depends. If it's vocab, I'm not gonna go over vocab." She further explained, "At the beginning we'll go over them, but I feel like sometimes if I go over them all the time, then they expect, they just wait for me to give the answers." She also entered two assessments into the same category that she called "packets," which she handed out at the beginning of each unit. Laura had four units, but there were only two packets in the gradebook, which she explained by saying, "I don't grade all the packets, as you can see, because I can tell how they're doing. But when I start seeing that the kids are not turning in the packets . . . I had to start collecting them." She was frustrated by this turn of events, because she said, "We go over as a class . . . so I'm like, why should they turn it in?" She did not want to have to record these activities completed as a class, but she felt compelled to do so because the students were not going through the learning process with her. Regardless of her frustration with the lack of work production on the part of the students, approximately 15% of her overall grade came from these

warm-ups and packets that were either done as a class or gone over as a class. This aligned with her belief that "All those assignments they need to do so they can, you know, better prepare for those other categories," like "quizzes or the mastery for the test." Turning in the packets and the warm-ups was a way for her to monitor their progress in understanding and mastering the material, even if they were completing them together as a class. When the students stopped doing the preparation for her quizzes and tests, she believed that their achievement would suffer.

However, in the same interview, she went on a long rant centered around the question, "Why should I have a Classwork category when they should be doing that anyways? Does that make sense? Because we know they're doing it by the way they perform on the quizzes and the tests. Period." She suggested several times in that part of the interview that "we should drop the classwork" or simply that the "Classwork category is too big." She called it a "cushion that's protecting them," saying that without the Classwork category, "We can fix that, to make it [their grade] more accurate," to genuinely reflect the students' language abilities. As much as she would have liked to change her grading system and categories, she lamented "the political side about it," giving three reasons for not dropping the classwork category. First, she said, "We're gonna have a lot of students not graduating" because they needed the Classwork grade cushion to help them pass the class. Second, "There's gonna be a lot of parents that are frustrated. But that puts it right back on them," because parents would have to pay for summer school due to the higher likelihood of class failure. And lastly she explained, "We would lose enrollment because they're not gonna wanna go . . . take a language," since they could take an easier course with a higher passing rate. She believed that classwork was a necessary evil, despite also believing that it did not measure student achievement. She felt that this quandary "puts us in a bind as

teachers," since she could be called upon to justify why a student's language mastery and his grade did not match.

One solution she proposed was to split the classwork category up. Her classwork category was already labeled "Reading/Listening/Writing/Speaking," and contained a hodgepodge of twelve warm-ups that were mostly writing and reading, one presentational speaking assessment, one interpersonal speaking assessment, one interpretive listening asynchronous assignment, three presentational writing assessments, and two unit packets that assessed primarily discrete grammar and vocabulary. Diana had an identically labeled category, which was also a mix of workbook activities, video activities, a skit, and grammar practices that incorporated presentational speaking, interpretive listening, presentational writing, and cultural comparisons. To solve the issue of the mismatched reflection of the students' achievement and their grades, Laura suggested, "If we can split that, we can do writing and reading in one, and then listening and speaking on the other. If we can split that Classwork up, you know, then maybe we can get more accuracy."

But Julia already tried this alternative category labeling, with her "Reading/Listening" and "Writing/Speaking" categories, which at first take would suggest a belief that these two categories would reflect receptive and productive skills, respectively. In Julia's case, however, the Reading/Listening category contained only one interpretive listening, and the four readings that she assessed were actually recited out loud, assessing their speaking, specifically their pronunciation, with no assessment of their interpretive reading skills. In her Writing/Speaking category she recorded grades for four interpersonal speaking assessments, seven student-written dialogues (presentational writing and speaking), and one memorized recitation (presentational speaking). Julia's attempt at renaming the categories into smaller bundles of skills demonstrated

that Laura's idea of splitting the Classwork category might still not clearly reflect the students' achievement on the particular domains that were meant to be assessed in each category.

***The Case FOR Classwork: Effort***

The Classwork category did not cause any conundrum for Martin, however. On the contrary, he maintained a category called, "Formative," valued at 60% of the grade. "Formative is really anything that happens on a daily basis," he said. He had one Formative grade per week on average, but that "grade would not just be related necessarily to that one piece of exercise," but rather would "review of all their work in the week." At the end of the week, he would give them an open-note quiz, explaining "Everything that they do in my class is open book," but they also had to attach all the work for the week to the quiz to turn it in, including the notes that they had taken during instruction. He reasoned that this work was not "necessarily graded on accuracy, but more like you use it as a teaching tool." He believed that this method was a way of "gradually building up proficiency." But he also saw this classwork as a way for the students to demonstrate their effort, which was a priority for him, as reflected in his belief that "Anybody could be a high achieving student, based upon the effort they put into the work which is given them." In his words, "The main thing is putting some thought and effort into it." He saw his job as a teacher as "setting these people up to be successful in the more important parts of their lives," and so he wanted them "to produce and to take ownership for what they want to do." He told an anecdote of two graduating student athletes who jokingly told him that they didn't think that they had learned anything in any of their classes throughout their four years in high school. But Martin regarded them in a different light, saying "They were good people. And so it wasn't so much the fact they could remember the content, but they had progressed." For Martin, this classwork grade was his way to encourage the development of a work ethic, by grading effort,

with the hopes of producing quality people as part of his language teaching. His grading reflected this equal priority of teaching the students to understand the language by grading on accuracy, but also reviewing their classwork, because he believed that "Nothing is ever stand alone. . . Everything is based upon total productivity . . . And so whenever I look at a person's work, a person's grade, it was always in terms of the big picture."

For students who struggled in his class, Martin felt that "purely for attempting, personally, I think it's worth a pass mark, anyway." He based this grading method on his explicitly stated grading construct, "What am I grading on? And that's why I come back to productivity, which is not amount. It is more participation and effort, consistency, a work ethic, all of those things." He tried to "encourage each one where they are," pushing for improvement in their productivity even if they started off very low. Martin said he was influenced at the beginning of his career by a language teacher at another school, who shared his own internally persuasive discourse, saying, "I tend to try and keep people. People feel that they're in the game. I normally don't give anything less than a 50. In other words, it's failing, but it's still reachable." Martin adopted this discourse as his own, mentioning several times that his metric for passing or failing a student was based on his perception of their effort, saying, "They're evidently not showing any interest at all, where I have to say, 'Okay, guys, you aren't even playing in the game, and the reality is now I've got, you've given me no choice but to give you zero.'" But he struggled internally with his interpretation of their effort and an appropriate course of action, asking himself, "I'm faced with, over a period, do I keep this person in the game or not?" And conversely, "Is this person worthy to be failed? And that's a judgment call for me." He acknowledged the dangers of his grading system, saying, "So it is very subjective from that point of view, and I don't know if that leaves me open to any kind of criticism." He mentioned an

incident where he allowed a low-achieving student to remediate her grade in order to keep her "in the game." But his generosity backfired, sharing that after he had been "leaning over backwards to help her" towards the end of the semester, that she and her mother demanded that he give her an "A" in the class, despite her consistently low effort.

Martin clearly stated his two main goals for teaching his students: to "identify their unique skills and talents and to build on those" and to "develop a work ethic." As previously discussed, he also aimed for them "to grasp hold of the language, which typically takes a lot more effort than simply producing assignments." But he realized that in the grand scheme of things, "As content is concerned, they will forget about that in a period of time after they leave high school." He was focused on developing his students as people, thinking long-term, from a "big picture" perspective. His 60% classwork category was the place in his gradebook that gave him space to accomplish these goals. Unlike Laura, he had no conflicting beliefs about the role of Classwork in his grading system. In fact, eliminating or reducing this grade weighting category would completely go against his purpose as a teacher.

Martin had the strongest reasons for valuing academic enabling behaviors as a grading construct, but to varying extents and for differing reasons, the other teachers also believed in the importance of recording students' classwork, as evidenced by this category in their gradebooks.

## Conclusion

This chapter brought to light these seven teachers' beliefs and practices regarding their grade weighting and assessments, showing that they all believed to various degrees that their grades should reflect both achievement and academic enabling behaviors. They had differing views on what achievement meant, whether it was understanding the culture of the language speakers, the framework of the language itself, mastering the language through accumulating

knowledge, or using the language to communicate, negotiate meaning, or participate in a language community. They had differing priorities for assessing speaking, writing, listening, and reading, and they used differing methods of administering, grading, and recording these assessments. They held strong reasons for some of their category weightings, while also haphazardly posting assessments in seemingly incompatible categories. Despite these incongruencies, these teachers were mindful of their underlying purposes for assessing and grading. Martin stated this thoughtfulness best:

> There is one constant thought about grading, it's not so much the structure of the grade book. It's my kind of self-analysis of it, of if I am sticking to my grading principles . . . And this is in my classroom. It's for my grade book. What really are, what really are the key things that, quite frankly, if I was using the word teaching, that I am teaching these children? . . . And so that to me is where the consistency comes in. Am I being consistent? Is that my consistent message to them? . . . And that's what I'm trying to do with this. And I believe for the most part, although I self-question, and I self-doubt, and my interface with people like you and others in the department whom I respect greatly is, "I don't know about this," is that that's what I chose to do when I began teaching. And I am being reasonably faithful to that.

The final chapter will discuss these internally persuasive discourses in the light of the historical authoritative discourses that may have influenced their formation or adaptation. The next chapter will further discuss all the findings in Chapters Four and Five in the context of previous research and the limitations of the scope of this study. Chapter Seven will then explore the possible implications of these findings.

## Chapter 6 Discussion of Authoritative Discourses

This study aimed to explore a group of world language teachers' beliefs and practices regarding their gradebooks and classroom assessments. The goal was to understand what these teachers assessed in their classroom, how they graded the assessments, and why they used these specific assessment instruments and methods of grading. This goal developed in part to answer Leung's (2004) questions, "What do teachers look for when they are assessing? What theory or "standards" do teachers use when they make judgments and decisions?" (p. 880). This group of teachers were asked to share their gradebooks with me and then they were invited to discuss their individual assessments with the hopes of uncovering their rationale behind those assessments and their grade weighting.

I framed my study in relation to Bakhtin's (1981) ideas about the ways that individuals' internally persuasive discourses are formed in contact zones of context where they weigh, consider, accept, reject, adapt, and assimilate the authoritative discourses and other individuals' internally persuasive discourses. In Chapter Four, I explored the contact zone of the teachers' context, including the authoritative discourses coming from their state, district, and administrators, as well as the classroom factors that molded their assessment practices and beliefs such as class size and make-up, available resources and training, and finally the resistance that they faced from their actual students.

In Chapter Five, I reported the assessment practices that these teachers used, as evidenced by their gradebooks, and the beliefs that underpinned those practices. These beliefs, or internally persuasive discourses, at first seemed to be a patchwork of the research that they had incorporated into forming their assessment practices, the collaboration that they participated in, and their experiences in the classroom. But this patchwork organized itself into themes about

how the teachers viewed achievement and the academic enabling behaviors that contributed to student achievement.

In this chapter, I will reexamine the contact zone of these teachers' assessment and grading practices and their internally persuasive discourses as they relate to previous studies. Then I will discuss the teachers' internally persuasive discourses about student achievement considering the historically prevalent authoritative discourses that these teachers likely encountered at some point in their individual contact zones. In the next chapter, I will explore the implications and limitations of this research and how the findings might contribute to the field of world language teaching.

## Contact Zones at Macro, Meso, and Micro Levels

Many of the issues that teachers experience while planning and implementing assessments are well documented in the literature. For these teachers, the issues appeared at three levels—the macro level of the school organization, the meso level of the classroom context, and the micro level of the students that were taking the assessments and being graded.

### *Macro Level of the Context*

The teachers in this study discussed many ways that the organization of their school context affected their practices for grading and assessing. The administrative expectation about graduation rates, the time crunch of their block schedule, their concerns about maintaining their full-time job status, the new implementation of a district-wide world language assessment, and required collaborative gradebook category weighting all exerted pressure on these teachers in their contact zone as they implemented feasible practices and developed internally persuasive discourses regarding assessment and grading.

**Graduation Rates.** When proposing her ideas about eliminating or reducing the weight of her Classwork category in her gradebook, Laura expressed her concern, “We're gonna have a lot of students not graduating.” She was worried that pushing a full achievement model of grading, rather than balancing that achievement with work productivity and effort, would affect the graduation rate at her school negatively. Kunnath (2017) discussed the cognitive dissonance between the desire for rigorous grading and optimal student promotion that teachers felt when making decisions about their grading practices. The shadow of this external pressure about the “acceptable” number of failures on teachers in the contact zone of grading and assessing appeared repeatedly in the literature (McMillan, 2003; Nowruzi, 2021; Xu & Liu, 209; Zoeckler, 2007). Laura’s fears about rocking the boat to make grand changes to her gradebook categories were well founded.

**Block Schedule.** Diana and Daniel mentioned issues with the time crunch and curriculum pacing of the 4X4 block schedule, particularly when administering their final exams at the end of the semester. They had to make choices that went against their own internally persuasive discourses to be able to comply with the district’s grade posting deadline, both in the final assessment that they gave and in the way they graded it. Daniel’s scantron reading comprehension final exam contradicted his high priority for communication. Diana’s artificial assignment of grades on speaking and writing portions of the final conflicted with her desire to provide significant feedback for the students to negotiate meaning. While not specifically addressing the issue of block scheduling, Lam (2019), Yan et al. (2018), and Wang et al. (2020) all found that practical issues like curriculum restrictions, teacher workload, and time shortage due to quantity of syllabus content, all put pressure on teachers in the context of implementing and grading assessments.

**Teacher Allotments.** Laura's fears about losing enrollment in world languages classes was echoed in current events at the university level, where enrollment has dropped by 16.6% from 2016 to 2021 (Quinn, 2023). Many smaller university world language departments have been forced to reduce their offerings or close down. The situation in the K-12 context was more diverse, according to "The National K-12 Foreign Language Enrollment Survey Report" (American Councils for International Education, 2017). This study showed that some language programs had been eliminated or consolidated and some programs had been expanded. These changes were local decisions, based on individual contexts and considerations. Laura's anecdote about her colleague who was pulled from his language to assist with ESL because of low enrollment numbers reflected her fear that changing her grading practices to match her internally persuasive discourse about student achievement could negatively affect her future employment status. According to these recent reports, her fear was valid.

**District-wide Assessment Initiative.** Laura was the only teacher to include the District Assessment category in her gradebook, and she was the only one to mention attempting to implement it. The other teachers' reluctance to administer or include the test in their gradebooks might echo the Villa-Larenas and Braunfot (2022) study where they found that ten of the twenty teachers that they interviewed felt conflicted between their personal beliefs about language assessment and their university's internal assessment culture. Several of the teachers, including Alicia, Carmen, Diana, and Martin, were not accustomed to assessing their students with multiple-choice tests, and might have believed that this online test would not reflect student achievement. They might also have perused the test ahead of time and found that the tests did not assess the material that their students were learning in class. Or, they may have heard about

Laura's cautionary tale of technical difficulties while implementing the test and decided not to waste their time.

**Collaborative Grading Categories.** A surprising finding in this study was the number of inconsistencies in the way that teachers used their collaboratively decided grading categories. These discrepancies harken all the way back to the discovery by Stiggins et al. (1989) that teachers did not know how to properly assign categories and grade weighting.

This team was required to collaborate about their gradebooks within language groupings, and even "singleton" teachers were privy to these discussions about categories, due to the small size of the department. But the agreed-upon categories and weighting caused many issues in the way the teachers classified their assessments.

For example, Julia's Quizzes category contained sixteen assessments, but only one closed-note assessment that would traditionally be called a "quiz." The rest were open-note, teacher-assisted, and six of the assessments were student notetaking. Julia's somewhat embarrassed explanation about the origin of her categories—"I don't know. I think I copied that from someone- horrible!"—showed that team decisions about categories and weighting sometimes backfired with misunderstandings about the purposes of the categories.

Martin inadvertently only included two assessments in his Summative category, which made them each 15% of the students' grade, even though he had intentions of including other similar assessments throughout the semester, saying that he "dropped off a bit at the end." He explained that his choice of categories came from collaborative discussions with the world language team: "I listen to what you guys do, and naturally, very often, what you guys do makes sense." However, his beliefs about open-note assessments and the importance of work

productivity and effort in a holistic sense over the course of a semester did not correlate well with this summative category, and he struggled to find ways to implement it.

Diana and Carmen both made these kinds of random category decisions regarding Homework. Carmen accidentally made one of her homework assignments 7.5% of her overall grade, which was almost as high as her final exam and her individual tests. She had decidedly negative views about homework, and this miscalculation showed her discomfort with using this collaboratively decided category, saying that she wanted to "reconsider the homework category." But she felt obligated to include this category in her gradebook, explaining, "I am on a team, and we do decide the percentages together." Diana decided to assign a Classwork assessment to the Homework category to put pressure on her students who did not seem to be working in class. Ironically, even though she wanted to apply more pressure on this assessment by giving it more weight, she actually caused it to weigh less–at 1.67% of the overall–than the more appropriate category of Reading/Listening/Writing/Speaking, where it would have been weighted 2.08%.

Some teachers seemed to be very mindful of the way they organized their gradebook categories, weights, and assessments. For example, Daniel was very clear in the ways he described his categories and his sequence of assessments, saying, "That's my structure." Other teachers were more capricious in how they categorized their assessments, mixing open-note classwork activities with unit tests, placing two similar assessments in completely different categories with differing weights, and weighing classwork activities equal to unit tests. In other cases, categories seemed to contain homogeneous items, with one or two random items thrown into the mix. It is possible that the limitations of the collaboratively decided categories in this contact zone went in opposition with the internally persuasive discourses of the individual teachers, causing discrepancies and miscalculations.

### *Meso Level of the Context*

At the meso level of the teachers' contact zone, this group struggled with their class sizes and make-up, along with a lack of resources and training. These limitations shaped the assessment practices that they could implement, even if their internally persuasive discourses conflicted with those practices. Simon et al. (2010) found that even pre-service teachers recognized that the feasibility of grading policies eclipsed their own theoretical aims of producing valid assessment data. These teachers were no different.

**Class Size**. Laura and Diana both found class size to be a constraint when designing their speaking and writing assessments. Laura, with thirty-eight students in one of her classes, found speaking assessments to be cumbersome to administer, even though she felt strongly enough about them to include them as part of her final exam, which most of the other teachers did not attempt due to time limits. Diana believed that she should give more feedback to her students on their writing assessments, as part of her internally persuasive discourse about negotiation of meaning, but asked herself, "When am I gonna do this?" when she had ninety students' essays to grade. While Duncan and Noonan (2007) did not find that class size factored into teachers' assessment practices, other studies (e.g., Cheng et al., 2004; Sun & Cheng, 2014; Wang et al., 2020) corroborated Laura's and Diana's struggle to implement quality performance assessments due to the size of their classes.

**Class Composition.** These teachers mentioned many issues with the spectrum of students that made up their classes. Daniel had to use more silent, independent assessments like Google forms in his combined-level classes, which went against his communicative definition of student achievement. To accommodate for his native speakers in another one of his classes, he had to use a totally different grading method on their writing and speaking skills assessments. Laura also

had to differentiate for her native speakers to make several of her quizzes and tests more challenging.

On the lower end of the ability spectrum, several of the teachers addressed the ways that they had to vary their assessments to make the curriculum more accessible, without the help of a special education teacher in the classroom. Alicia provided the same accommodations for all her students so as not to single out her lower-ability students, which strongly corresponded to her belief that language achievement meant participating in a language community. Carmen became a pro at differentiating her individual assessments for specific students so that they were better able to show their understanding of the language concepts. Daniel adjusted his instructional pacing for his slower students, to ensure that everyone was able to communicate. And Laura allowed her students to show their accumulated knowledge in different modes, for example, speaking instead of writing for students who were poor spellers.

Laura and Martin also discussed student class placement as an issue when designing their assessments. They both spoke about the students' confusion about why they were taking the class, either that they erroneously believed that the class was required for graduation and were "trying to check off the list," or that they did not understand what learning a language required. Both McMillan (2003) and Kunnath (2017) addressed this heterogeneity of advanced, on-level, and remedial students all in the same classroom as one of the factors that teachers had to consider in their classroom context as they attempted to assess and grade.

**Resources**. In previous studies, Vogt et al. (2000) and Villa-Larenas and Braunfot (2022) found that teachers used whatever testing materials were available from published textbook programs and from their colleagues' repertoires. In this study, Diana, Daniel, and Julia talked about the textbook assessments that they had to sift through to find appropriate ways to assess

their students. But they also significantly modified those tests and activities in order to conform the published materials to their language achievement priorities. Alicia, however, lamented the amount of time needed to create her performance-based assessments from scratch, since the district had not yet purchased an appropriate assessment program. Both Rogers et al. (2004) and Wang et al. (2020) found that the time required to develop, administer, and evaluate performance assessments in the language classroom caused other teachers issues, just as it did for these teachers.

**Training.** Julia was the only teacher in this group that specifically mentioned a lack of training, saying, "I wish I could know how to use rubrics." But some of the other assessment gaps that were exposed during this study, like the low percentages of speaking, particularly interpersonal speaking, might have been related to a lack of training in using rubrics or other data analysis methods in the assessment grading process. This matches the research results of Vogt and Tsagari (2014) and Vogt et al. (2020), who found that alternative assessment implementation in the classroom could be hampered by lack of training. Several of the teachers in this current study—primarily Laura, Carmen, and Julia, but also, at least to some extent, Daniel, Diana, and Martin—utilized traditional, discrete-point assessments of students' vocabulary and grammar knowledge. Tsagari and Vogt (2017) called these "microlinguistic" assessment methodologies, and posited that teachers needed more training to feel comfortable using alternative assessments. The teachers' discomfort in using alternative assessments due to a lack of evaluation training might have influenced the way the teachers believed about language assessment and grading.

Martin seemed very comfortable using alternative assessments like his task-based student learning projects, but he also acknowledged that his subjective grading method opened him up to criticism without having something solid to base the students' grades on. A well-written rubric

would have allowed him the flexibility to evaluate the students' effort, which was one of his highest achievement priorities, while also providing him some structure to justify his grading. Martin, Julia, and this whole group of world language teachers could have likely benefited from some assessment training, especially using rubrics, whether they realized it or not. Learning how alternative assessments could be effectively administered and evaluated in the contact zone of the classroom could shape teachers' internally persuasive discourses and their instructional practices.

### *Micro Level of the Context*

Along with the school and classroom environment, the students' behaviors and issues also applied pressure to the ways that teachers thought about their assessments and grading. Students' absenteeism, anxiety, disengagement, and cheating habits all caused the teachers to reevaluate their internally persuasive beliefs and subsequently their grading and assessment practices.

**Student Absenteeism.** McMillan (2003) found that classroom realities like social promotion, absenteeism, disruptive behavior, and heterogeneity affected teacher decision making. Daniel, Alicia, Martin, and Carmen specifically mentioned absenteeism as an issue for their students. Martin adjusted his assessment and grading practices to accommodate this absenteeism, by giving one "big picture" grade per week, that would measure the "total productivity" of the student, assuming that "If they were there five days a week they would be working," based on the work that they turned in on the days that they were present. He wanted to keep them "in the game." Both Kunnath (2017) and McMillan and Nash (2000) found this same teacher philosophy of pulling for student success and encouraging motivation and engagement in their previous studies. Martin believed that penalizing the students for their absences would

negatively impact their success, motivation, and engagement, and he felt compelled to create a solution that would be fair for all. However, his students' absenteeism did cause him to struggle in his beliefs, since he realized that there was a "limit" to having to "cover up for people being absent every now and again" through his weekly grading policy. He said this system was "subjective," and that he had to "watch very closely people's habits and ask them why," knowing that sometimes these absences were because of "lack of interest." He explained that in those cases, "the zeros kick in because they are making no effort at all." For Martin, language achievement was tightly related to effort, and accommodating apathetic behavior through his holistic grading system clashed with his internally persuasive discourse. His inclusion of non achievement factors in his gradebook aligned with Norwruzi's (2021) study of three hundred Iranian teachers who considered effort, improvement, ability, and participation, even more than achievement factors in their grading, considering them to be academic enablers to increase learning and motivation.

**Anxiety Issues**. Anxiety in world language classes has been studied since the middle of last century, as evidenced by Scovel's (1978) literature review. However, more recent studies (e.g., Febria & Marsevani, 2024) highlighted the struggles that students had with speaking in the post-pandemic language classroom. Julia specifically referenced changes that she made to her assessment environment to ease students' anxiety upon returning to school after the pandemic. Previously, one of her presentational speaking assessments was read in front of the class with a partner, but she found that the students got "super stressed," so she changed the testing environment to the more private setting of the hallway. She then graded them on how "comfortable" they were in the language. Carmen also dealt with this newer, stronger anxiety, by basically giving up on evaluating her students' speaking abilities, saying that producing language

caused them “anguish.” Diana lowered her expectations for her students’ speaking abilities due to their “shaking and crying.”

Alicia, on the other hand, fought back against the anxiety issue, continuing to have her students present in front of their peers, saying, “We do it in front of the class, because we all learn from each other.” This was part of her high priority to encourage participation in the classroom community, to “empower” them, and to “build that self-esteem.” Brookhart (1993) also found that teachers, like Alicia, believed that they served an important role in developing students’ self-esteem. Regardless of how these teachers responded, clearly the students’ anxiety put pressure on them in their contact zone as their internally persuasive discourses and practices evolved due to a changing world.

**Disengagement and Refusal to Work**. Wang et al. (2020) found that the characteristics of the students at each institution could factor heavily in assessment practices, depending on their language proficiency, motivation, and engagement. These world language teachers encountered issues with student engagement and motivation. For example, Laura, Carmen, and Julia all graded their warm-ups to encourage the students to engage with the language at the beginning of class. Laura clearly struggled with her internally persuasive discourse about student engagement when she discussed her Classwork category and her vocabulary and grammar packets. She believed that she should not have to post these instructional activities as assessments in her gradebook, but she found that without entering the zeros, that the students would not complete the assignments, even when they worked together as a class. These experiences with low levels of student engagement and work ethic became part of the beliefs that all the teachers held about high and low achievement.

**Cheating.** Murdock and Beauchamp (2008) reported that cheating was hard to define because it could mean different things to different stakeholders, whether it related only to the behavior or also to the intent. They found that it was also hard to study empirically with high reliability because the methodology usually relied upon self-reporting. Regardless of how it was defined, all the world language teachers in this study mentioned some form of cheating as an issue that shaped their grading and assessment practices. Alicia constantly updated her exams, scrambled the question order, and remained vigilant as the students were testing, which were some of Murdock and Beauchamp's suggestions for preventing cheating. Another example of the suggestions offered in this previous study was Carmen's policy of making multiple versions of her tests. In another strategy, Daniel significantly reduced his grade weighting category to minimize the effects of the students cheating on classwork and homework, so that their grades would reflect their actual achievement, rather than their ability to copy other students' work. Carmen echoed this belief that students' grades should reflect achievement, as evidenced by the low number of grades in her homework category, where she said, "At home we have other devices and access to other things to help us. And it's not a clear indication of what they can do." All the teachers believed that the students' grades should reflect student achievement and went to great lengths to ensure the reliability of these grades. But the students' cheating behavior influenced the teachers' grading practices and seemed to strengthen their internally persuasive discourses about achievement.

Many of the factors that affected language teachers' grading and assessment practices reported in previous research worldwide also appeared in this study. This contact zone for teachers was similar across time and space, as they had to wrestle with what, how, and why to assess and grade within the limitations of their contexts. In this contact zone they developed their

own unique internally persuasive discourses as they contended with multiple, contradictory authoritative discourses about best practices in grading and assessing.

## Authoritative and Internally Persuasive Discourses in World Language

### Authoritative Discourses in World Language

Authoritative discourses are defined as "the authoritative word" (Bakhtin, 1981, p. 342), "generally acknowledged truths" (p. 344), "organically connected with a past that is felt to be hierarchically higher" (p. 342). Typically, "One cannot divide it up—agree with one part, accept but not completely another part, reject utterly a third part." (p. 343). In the world of teaching and learning languages, authoritative discourses could be considered the pedagogical principles that have become established through tradition, research, and experience by the experts in the field.

In the long tradition of teaching languages before the twentieth century, the authoritative discourse about the approaches to language pedagogy vacillated between analyzing the language and using the language (Celce-Murcia, 2001). In the nineteenth century, the Grammar-Translation ideology related most closely with language analysis while the Direct Method in the United States and the Reform Movement in Europe both aligned with a usage philosophy. Throughout the twentieth century, psychologists began refining theories about knowledge and learning that had been previously explored, and the behaviorist/empiricist, cognitive/rationalist, and situative/pragmatist-sociohistoric perspectives rose to prominence in general education (Greeno et al., 1996). These modern grand learning theories shaped the ways that language teaching approaches evolved over the next century, as methods proliferated, with new methods often rejecting their predecessors because of perceived weaknesses. Celce-Murcia (2001) summarized the newer twentieth century approaches to include Audiolingual, Natural, Reading, Cognitive, Humanistic, Comprehension-Based, and Communicative. Another list of approaches

also included Structural, Task-based, Inquiry-based, Lexical, Immersion, Community Language Learning, Cooperative learning, Content and Language Integrated Learning (Renau Renau, 2016). Larsen-Freeman and Freeman (2008) added Interactional and Critical approaches to the mix. These are by no means comprehensive lists of language teaching approaches, since new empirical research continues to lead to new ideas which in turn evolve into new, competing authoritative discourses about priorities for student achievement in languages.

At least some of these discourses were likely introduced in pre-service teachers' pedagogical and methodological courses before they even reached the context of the classroom. Teachers themselves often previously learned their foreign language(s) in classrooms that utilized one or more of these methods. These exposures to the authoritative discourses during their own language learning process and their teaching training collided in the contact zone of their classroom context to shape teachers' internally persuasive discourse about student language achievement.

**Teachers' Internally Persuasive Discourses**

According to Bakhtin (1981), "Internally persuasive discourse—as opposed to one that is externally authoritative—is, as it is affirmed through assimilation, tightly interwoven with 'one's own word.' In the everyday rounds of our consciousness, the internally persuasive word is half-ours and half-someone else's" (p. 345). It "is gradually and slowly wrought out of others' words that have been acknowledged and assimilated" (p. 345). Some of these words originate from authoritative discourse, and some of them come from others' internally persuasive discourses. "The struggle and dialogic interrelationship of these categories of ideological discourse are what usually determine the history of an individual ideological consciousness" (p. 342).

The teachers in this study seemed to bring the authoritative discourses that they encountered in previous experiences "into new contexts, attach it to new material, put it in a new situation in order to wrest new answers from it, new insights into its meaning, and even wrest from it new words of its own" (p. 346). In their new context, the teachers adopted, rejected, adapted, and modified the historical authoritative discourses in their process of ideological development as teachers. Many of the wildly varying authoritative discourses from the field of teaching and learning languages appeared in pure or mutated forms in the internally persuasive discourses of these teachers. These authoritative discourses, as well as the discourses of others, strongly influenced their decisions about what, how, and why to assess and grade their students' learning.

The findings of this study helped support Leung (2013), who theorized that teachers' views on learning and language were complex, changing, and might align with their assessment practices. Leung and Scarino (2016) called these views "a patchwork of different ideological, institutional, and societal aspirations and requirements that are not necessarily conceptually or educationally coherent" (p. 88). Kunnath (2017) and McMillan and Nash (2000) concurred that grading practices were related to teachers' philosophies, while Bachman (2007) posited that teachers' assessment constructs were based on their beliefs about knowledge, their purposes, and their view of the world. Considering how different these teachers were in their internally persuasive discourses about world language achievement, it is no wonder that their assessment and grading practices also differed greatly. The next section will reexamine the teachers' internally persuasive discourses through a lens of how they adopted or adapted historical authoritative discourses from the field of world language teaching and learning into their own belief systems.

**Grades Reflect Achievement**

All the teachers held strong internally persuasive discourse that students' grades should reflect their achievement, as evidenced by their Test, Quiz, and Final Exam categories. Even so, they gave differing labels, differing weights, differing meanings, and differing purposes to the categories. The Final Exam category was the most regimented, because of administrative expectations and also due to the strong influence of the time limit in the teachers' context. But even here, Martin's strong beliefs about students' effort as demonstrating achievement compelled him to assess the students by means of a project, rather than a traditional test.

All the teachers weighted their Test category at 30%, except Alicia, who weighted it at 10%. Alicia, Carmen, Daniel did not allow students to access class resources on any assessment in this category, while Diana, Julia, Laura allowed resources on at least one of theirs. Martin allowed resources on both of his two Summative (test) category assessments.

The Quiz category was common for all teachers except for Martin, but the grade weighting, the contents, and the purposes varied wildly. Alicia, Carmen, and Daniel weighted their Quiz categories differently, and their content varied from teacher to teacher, but the content for each teacher was coherent and homogenous. Diana, Julia, and Laura, on the other hand, also had differing weights, content, and purposes, but their quiz category itself was not uniform in content. In fact, Brookhart (1991) would likely call this out as a "hodgepodge" (p. 36) grading category for them. These findings aligned with Zoeckler (2007), who discovered that teachers' grades were individually tailored, even if they appeared to be similar, and Cox (2011), who demonstrated that teachers had differing beliefs about the meanings of grades and the contents of the grading categories.

It was in the purposes of the Test and Quiz categories that traces of authoritative discourses about grading could be most clearly found. Martin and Daniel both used the labels "Summative" and "Formative," harkening back to Scriven (1966) and Bloom (1968), who developed the terms to describe the roles of assessment used as terminal evaluations (summative), or as process research (formative) to "discover deficiencies and successes in the intermediate versions" (Scriven, 1966, p. 16). None of the other teachers used the terms *summative* or *formative* either in their gradebooks or their interviews. However, Carmen said, when discussing quizzes, "I try to use that information to guide me for the test, and I try to tell them, 'Oh, see, this is the one area you need to work on,'" indicating her understanding of the dual role of her quizzes, as an assessment *of* learning (summative) and an assessment *for* learning (formative). Alicia went a step further in her utilization of her Performance-Based Assessment/Quiz category. Because students performed in front of the class, their peers were also able to use the assessment *as* learning, as Alicia explained to them, "That's how we learn," from "the feedback."

Many of the other teachers, including Daniel who used the terms *Formative* and *Summative* in his category labeling, seemed to use their Quiz category primarily as assessment *of* learning (summative) at the end of a short-term instructional cycle, even though the content would reappear subsequently in the more comprehensive Test category in a different format. Martin's Summative category contained the same kinds of assessments as his Formative category, just scaled up to be a "larger review than we would normally do." This would mean that although Daniel and Martin adopted the terms from the authoritative discourse, their adaptation of those terms was distorted from the original definitions and intents of those terms.

On the other hand, Alicia and Carmen applied the principles of the authoritative discourse without using the actual terms.

Most of the teachers' actual formative assessments (*for* learning) showed up in their Classwork categories. But because a detailed examination of summative and formative assessments went beyond the scope of this study, those separate percentages were not as distinct in this data set. However, an argument could be made that this study aligned with previous research (e.g., Deluca et al., 2019; Frey & Schmitt, 2010; Vogt et al., 2020b; and Wang et al., 2020) that showed discrepancies in teachers' beliefs, understandings, and practices about formative and summative assessments.

Regardless of how the teachers categorized their students' achievement, their beliefs defining that achievement became evident when observing what they assessed, how they assessed, and why they assessed the items in their gradebooks. Those differing internally persuasive discourses will be discussed in the next section, along with the authoritative discourses that seem to have influenced those beliefs.

### *Achievement as Conceptual Understanding*

Carmen and Martin, and to a lesser extent Laura, prioritized student understanding of language concepts as they discussed student achievement. They mostly graded their students' assessments based on accuracy, as a way to see if the students could "grasp the language," the "formulas," "frameworks," "patterns," and "systems." Several of their assessments that seemed on the surface to be communicative in nature were actually used to test students' understanding of grammatical concepts.

One of Martin's internally persuasive discourses seemed to align with an analytical Grammar-Translation approach to teaching languages. This approach originated from Latin and

Greek studies in the sixteenth century, and later became the predominant method of teaching languages through the influence of Karl Plötz in the nineteenth century (Celce-Murcia, 2001). The Grammar-Translation approach focuses on "grammatical parsing, i.e., the form and inflection of words" (Celce-Murcia, 2001, p. 6), using L1 to teach L2, frequent use of translation from L2 to L1, and low priority for speaking or using the language. Martin only assessed for understanding of the language and its grammars, never on the students' memorized recall of specific vocabulary or grammar, in the fact that all of his assessments were open-note. He never made them translate from L1 to L2. Instead, he believed that "There is grammar behind it all." His priority was for the students to "think the language" and "come to grips with it," rather than to use the language for communicative purposes, even when he implemented interpersonal speaking assessments meant for communication. Martin fully adopted the historical authoritative discourse of the Grammar-Translation approach, although he combined it with several other approaches that will be discussed later.

Carmen, however, who also viewed language learning through an analytical lens, rejected the use of translation and the use of L1 in her assessments. Her assessments aligned better with a Cognitive approach, as defined by Celce-Murcia (2001) to prioritize grammatical rule acquisition, whether acquired through deductive or inductive means. This teaching approach is rooted in Cognitive learning theory, which James (2006) described: "A particular focus is on how people construct meaning and make sense of the world through organising structures, concepts and principles in schema (mental models)" (p. 8). The Cognitive approach to teaching languages, as opposed to Cognitive learning theory, emphasizes reading, writing, and vocabulary, while downplaying the importance of pronunciation while speaking. The focus is on rule acquisition and conceptual understandings, rather than rote habit formation. The students

bear the responsibility for their individualized learning, and error correction is used as part of the learning process.

Most of Carmen's assessments were based in reading and writing, with only 1% speaking and 8.5% listening. She did not concern herself with the students' pronunciation. She placed a heavier emphasis on acquiring grammar concepts than on utilizing vocabulary for communication, as evidenced by her sequential quizzing of conjugations before meanings when assessing the students on their verbs. Even her method of getting to know the students seemed to be based more on their cognitive abilities than on their social identities, saying, "I do my research prior to. And I look and see how many students I have that are special Ed, remedial, 504. I go back and look at what previous teachers say in their IEP reports." Based on these reports and their performance in class, Carmen was able to successfully differentiate their instruction on an individual basis, adding extra notes and underlining for students who she knew would need extra support. While she did her best to accommodate for their differing cognitive abilities, she also held them accountable for their responsibility to learn by including a section for students to give her feedback at the bottom of each of their tests. The students responded to prompts like, "I came in for help when I didn't understand" or "I asked for help when I didn't understand." In one anecdote, she said that she was excused from a parent-teacher conference because she could show the parent the students' feedback responses on a test that said, "I didn't try, not my best work, didn't really study, didn't ask for help repeatedly." Carmen was almost textbook in her adherence to the Cognitive approach, indicating that she had fully accepted this authoritative discourse.

### *Achievement as Accumulated Knowledge*

Laura also subscribed to a belief that language achievement meant understanding concepts. For her, these understandings showed up in the way that her assessments sometimes measured the students' mastery of the grammar rules themselves, rather than strictly observing their ability to apply the rules. Several of her assessments demanded that the students explain their reasoning for using the grammar rules, which aligned with Cognitive learning theory, but not necessarily with a Cognitive language teaching approach.

However, overall, Laura's teaching beliefs seemed to match the Acquisition metaphor of learning languages (Sfard, 1998), where "Concepts are to be understood as basic units of knowledge that can be accumulated" (p. 5). According to this metaphor, knowledge is seen as a commodity that can be collected and stored. This acquisition model stems from Behaviorist learning theory, as James (2006) explained: "Achievement in learning is often equated with the accumulation of skills and the memorisation of information (facts) in a given domain, demonstrated in the formation of habits that allow speedy performance" (p. 7). According to James, teaching approaches based on behaviorist learning theory break the whole into deconstructed parts that can be "practiced and reinforced and subsequently built upon" (p. 7). Simple skills progress to more complex skills.

Laura adopted this authoritative discourse into her own internally persuasive discourse, with a strong emphasis on assessing discrete-point vocabulary and grammar memorization by means of repetition, songs, and mnemonic tricks. Her vocabulary quizzes progressed from simple recognition to full recall throughout the semester. Her assessment routines of the daily warm-up, repetitive homework grades, and her vocabulary and grammar packets provided the

students with the perfect environment for acquiring language units, according to Greeno et al. (1996).

Several other teachers also incorporated this behaviorist ideology into their accumulation-based teaching approaches. Carmen emphasized discrete grammar and vocabulary on her quizzes, but her beliefs about the importance of understanding concepts compelled her to assess their grammar first, and then their vocabulary skills. She also focused on repetitive writing as a means of encrypting the sentence structures, vocabulary, and spelling onto her students' brains. She valued quick feedback, which was a chief aspect of the stimulus-response nature of behaviorist theory. She tried to find ways for her students to practice "errorless learning," (Terrance, 1966, as cited in Greeno et al., 1996, p. 28), which was possibly why she did not use her homework category to its full potential or assess the students on their interpersonal speaking.

Daniel, while his focus was on communication, incorporated the behaviorist principle of breaking down the whole into its parts in his assessment chronology. He was very mindful of his sequential pattern, explaining that he assessed vocabulary first, which would be essential for his students to communicate in the language. Next, he assessed their discrete grammar skills. Then he integrated the parts back into the whole on their summative test to assess their reading, writing, and listening skills using the vocabulary and grammar that he had previously quizzed. And finally, he assigned a performance-based speaking and listening task for the students to demonstrate their newly acquired skills. Greeno et al. (1996) called this kind of curricular design "sequences of component-to-composite skills" (p. 27), when discussing behaviorist learning theory. Daniel's gradebook structure showed his partial adaptation of the authoritative discourse of behaviorist accumulation principles, even though communication was the end goal of his assessments, rather than just sheer acquisition.

Julia, Diana, and Alicia also used varying levels of discrete-point vocabulary and grammar assessment, showing an inclination to view language as the sum of its parts, but they also had achievement goals for their students other than just accumulation of knowledge. In general, these findings regarding discrete-point assessments corresponded to previous studies that found that many teachers assessed microlinguistic units of language like grammar and vocabulary in isolation (Frodden et al., 2004; Mede & Atay, 2017).

### *Achievement as Using the Language*

The pendulum of the authoritative discourse about language teaching approaches has always swung between an analysis focus and a usage focus (Celce-Murcia, 2001). The next section will discuss the discourse about language usage and how it appeared in the internally persuasive discourses of these world language teachers.

**Communication.** The ACTFL standards that these teachers followed, often called "the five C's," start with the first "C" of "Communication" (The National Standards Collaborative Board, 2015). This standard includes interpersonal speaking and writing, presentational speaking and writing, and interpretive reading and listening. A deep dive into the teachers' assessment practices showed that several teachers prioritized communication as the highest indicator of student achievement.

Daniel was the most explicit in his communicative emphasis, both in his assessment practices and his verbalized beliefs. His approach was most closely aligned with a Communicative Approach (Celce-Murcia, 2001) or Communicative Language Teaching, as it has also been called (Savignon, 1987). Communicative Language Teaching is more of a set of principles based on communicative competencies, rather than an organized approach. Celce-Murcia (2001) explained, "The basic principle involved is an orientation towards collective

participation in a process of use and discovery achieved by cooperation between individual learners as well as between learners and teachers" (p. 25).

Daniel's gradebook revealed equal, purposeful attention to all the subsidiary standards of Communication, such as interpersonal speaking and writing, presentational speaking and writing, interpretive listening and reading. He employed integrated projects, presentations, interviews, reading and listening excerpts, and essays as conduits for his students to show their competence in using the language. His strong beliefs about avoiding the students' L1 in the language learning process and his open-ended writing and speaking that he graded using rubrics, or through differentiated subjectivity based on students' previous knowledge, pushed students to use the language to communicate on almost all his assessments.

Other teachers like Alicia, Diana, Julia, and Laura also held high commitments to communication in their practices and beliefs, but their implementation was either not as comprehensive as Daniel's, or they had more specialized purposes or approaches. However, this small study of these teachers' gradebooks confirmed previous studies that showed that regardless of their regular practice of using traits-based assessments, many teachers believed that they should be measuring their students' communicative abilities (Díaz et al., 2012; Tsagari & Vogt, 2017).

The next few sections will examine some of the more specialized approaches for communicative language use that this group of teachers employed in their assessment practices.

**Negotiation of Meaning.** Diana's percentages for assessment of the communication standards were similar to Daniel's, showing a strong belief in the importance of using the language in the classroom. But Diana was very specific in her priority for the students to "expand and elaborate" in their interpersonal speaking and presentational writing in order to

produce language output. Her repetition of the phrases, “It doesn’t make any sense” and “What do you mean here?” when talking about her students’ language output pointed to a belief that language could be acquired through negotiating meaning between interlocutors. She used individual spoken feedback on their interpersonal writing to help them renegotiate their answers to better communicate their ideas in preparation for an interpersonal speaking assessment. She even forced them to translate their own L2 output back into their L1 so that they could see where there was a disconnect between their interlanguage and the correct target language forms. She also wished that she had more time to provide her students with better, more complete feedback, while also realizing that some of her students would not appreciate it or use it. Wang et al. (2020) found that students often ignored written comments about their language production and instead focused on the gradebook score. Diana’s classroom experience fell right in line with this previous research.

This focus on negotiation of meaning indicated Diana’s alignment with the principles of the Interaction Hypothesis, which states that “Output serves as oral practice, aids fluency and provides learners with the opportunity to test hypotheses about the rules they have constructed for the target language" (Mackey, 2002, p. 380). According to this hypothesis, “Tasks in which there is a need for the participants to exchange information with each other promote more interactional restructuring" (Ellis, 1991, p. 182). Interaction Hypothesis encourages a task-based approach to learning, which correlated with Diana’s open-note, open-ended assessments that she graded using a rubric. These tasks, like Diana’s interpersonal speaking and presentational writing assessments, forced the students to notice and compare language forms and functions, thus seeing how they needed to modify their language output as they negotiated their meaning to be more comprehensible by their interlocutor (Ellis, 1991). Interestingly, this approach typically

does not make use of translation to negotiate meaning, but instead uses recasts, which are negative conversational feedback episodes that force the speaker to modify his output (Long, 2014). So Diana did not adopt this authoritative discourse in its pure form into her internally persuasive discourse, but she adapted it to fit her classroom context.

**Participation in a Community.** Alicia's emphasis on using the language was more in line with Sfard's (1998) participation metaphor, unlike the other teachers that subscribed to the acquisition metaphor. According to the participation metaphor, "Learning a subject is now conceived of as a process of becoming a member of a certain community" (p. 6). Alicia was teaching a language by encouraging the students to participate in the community environment that she set up for them. She accomplished this through forcing classroom survival skills in L2, encouraging hive answers to her cold call questions to individuals, where the spotlighted student was allowed to ask for the community's help as long as he used the language. She used group projects, peer work, and presentations in front of the class with public feedback so that all of her students, including herself, could "learn together."

Alicia seemed to be using a Situated Learning approach, described by Lave and Wenger (1991) as a "community of practice" (p. 29). A Situated Learning approach means that "Learners inevitably participate in communities of practitioners and that the mastery of knowledge and skill requires newcomers to move toward full participation in the sociocultural practices of a community" (p. 29). Learners, initially considered to be peripheral members of the community, are included in new activities, tasks, and functions, that allow them to be gradually incorporated as central members of the community. As they learn the discourse and "the culture of practice" (p. 95) of the community, they develop their own identities as full practitioners. Lave and Wenger cautioned "that legitimate peripheral participation is not itself an educational form,

much less a pedagogical strategy or a teaching technique. It is an analytical viewpoint on learning, a way of understanding learning" (p. 40). In a community of practice, learning the knowledge that is common to the group is subsumed by the social process of group involvement.

In Alicia's classroom, the discourse of the group was the language that she was teaching. For peripheral members to become full community participants, they had to speak the language, where collaboration and communication were essential to being successful in the social environment of the group. This focus on community also translated to her gradebook, in her partner and group assignments, her language simulation project, and her insistence that the students read and watch movies about the language community's culture. Alicia appeared to fully embrace the authoritative discourse of Situated Learning as her own internally persuasive discourse, implementing it without modifications into her assessment and grading practices.

**Recitation and Reading Aloud.** Julia adopted an older authoritative discourse into her classroom assessment and grading practices. Her recited and memorized dialogues, her focus on pronunciation, and sequential introduction of skills (listening, speaking, reading, writing) all mimicked the Audiolingual approach from the 1940's. This approach aimed for native-like oral fluency by means of drill and repetition and defined "language as mental habit" (Larsen-Freeman & Freeman, 2008). The Audiolingual approach stemmed from Behaviorist learning theory, particularly B.F. Skinner's ideas about operant conditioning based on positive and negative reinforcement (East, 2021). The Audiolingual approach was not Julia's only focus for language assessment, since she also employed a Communicative approach and even some Grammar-Translation-like assignments. However, her gradebook showed that 20.75% of her overall grade calculation was based on AudioLingual-type assessments like recitations and reading aloud.

**Achievement as Understanding Culture, Comparisons, Connections, Community.**

For two of these teachers, Martin and Alicia, achievement was centered not just on the language itself, but also on the culture of the speakers of the language. Martin and Alicia both had high percentages of postings in their gradebooks that assessed the students' ability to make comparisons and connections to the perspectives and artifacts of the language community's culture.

Martin employed a Constructivist experiential learning approach to assess his students' understandings of culture and history. "Constructivism is realised in approaches that emphasise students' active construction of their own knowledge, in contrast to sitting passively and receiving information from teachers or textbooks" (East, 2021, p. 170). In this approach, students work independently or collaboratively to construct meaning from the complexity of the world around them. Celce-Murcia (2021) explained that in this type of experiential approach, the teacher serves as a facilitator and language learning is a process of self-actualization for the student. This Constructivist project approach has been found to be particularly effective in developing students' insider perspective of culture (Troyan, 2012), which is how Martin used it in his classroom. Martin's stated goals when assigning his final exam project and his other cultural projects throughout the semester were to help his students discover and develop their talents and potential, to connect to the language's culture, and to develop a strong work ethic by showing high productivity and effort throughout the process. Martin accepted the authoritative discourse of constructivist, experiential learning in the form of task-based instruction as part of his own belief system about how his students showed their achievement for the standards of Culture, Comparisons, Connections, and Community.

Understanding culture was essential in Alicia's classroom community of practice. She incorporated historical readings and movies that highlighted the community of her language into her Classwork category, providing fodder for discussions comparing the students' native culture to the new culture that she was exposing them to in her classroom. These readings prepared them for a simulated immersion experience, where they could become an L2 language speaker for a day. This experiential learning episode, with a high grade-weighting to encourage full participation, helped bond their classroom community even more. Alicia explained, "Everybody participates," moving them towards the center of full group membership from their previous peripheral involvement. Because of the readings and movies, this immersion experience connected them not just with their classmates in their local community, but also with the worldwide community of language speakers across space and time. The reflection essay that she required as a follow-up assessment of their experience helped them express what it was like to be part of this global community, and Alicia's individual feedback on each of their essays helped them feel fully accepted into their local classroom language community. While Alicia was not able to provide a true immersion experience for her students, she used her Situated Learning approach to expose her students to the culture of the language community. Her instructional and assessment practices pointed to her strong internally persuasive beliefs about the efficacy of a community of practice, which seemed to be a full adoption of the historical authoritative discourse of Situated Learning.

Troyan (2012) found that the ACTFL Communication standard, that included listening, reading, writing, and reading, was the only standard that had developed fully operationalized assessments of student achievement, via the ACTFL proficiency guidelines and integrated performance assessments. The other 4 C's of Culture, Comparisons, Connections, and

Community, and how they were meant to be integrated with language learning, have never been fully actualized. Several of the teachers incorporated Culture into their Communication standard assessments. But some of the other teachers in this study struggled to include Culture, Comparisons, Connections, and Community in their assessments and had low gradebook percentages for these standards, similar to the teachers in the studies cited in Troyan's literature review.

**Grades Reflect Academic-Enabling Behaviors**

All the teachers mentioned student effort, engagement, or work production in their definitions of high-achieving or low-achieving students. These academic-enabling behaviors showed up in their gradebooks, typically in some form of Classwork category. Back in 1989, Stiggins et al. showed that grades often measured effort, motivation, personality, attitude, and interest, against the directives of grading experts at the time. This current study indicated that a quarter of a century later, effort has still not been eliminated from the list of non-achievement characteristics that teachers considered in their grading practices. The recommendation to exclude these factors has been repeated in the more recent authoritative discourse on grading (e.g., Guskey & Brookhart, 2019), but these teachers rejected this imperative in their practice and in the way they revealed their internally persuasive discourse throughout their interviews.

Carmen, Julia, and Laura used daily warm-up grades as a way to control student behavior by getting the students to begin working at the beginning of class, similar to how Simon et al. (2010) found that pre-service teachers believed that assessments could be used for classroom management. All the teachers in this study seemed to use their Classwork category as a way to hold their students accountable for the instruction that they were receiving, by way of class notes, vocabulary and grammar practices, or cultural readings. These teachers found that when

they did not grade these instructional activities, that the students stopped engaging in the learning process, which in turn affected their achievement grades negatively. This finding was in line with Nowruzi's (2021) study that effort, improvement, ability, participation, and behavior could contribute to future learning and motivation, and with Brumen and Cagran (2011) who found that teachers believed that students would not learn anything if they were not given assessments. However, none of the teachers in this current study referred at all to an economic system of "paying" students for their work using grades, as was the case in several previous studies (Brookhart, 1993; Simon et al., 2010; Sun & Cheng, 2014).

Laura, despite her high Classwork weighting percentage, had strong reservations about grading her students' effort. Her long discussion about eliminating or reducing the classwork category reflected her struggle to reconcile the authoritative discourse about grading solely on achievement with the daily realities of her contact zone. In her context, she feared being judged for having a low passing rate, which could result in lower enrollment in her classes, which could eventually cause her to lose her job. Her internally persuasive belief that the classwork cushion masked her students' low language abilities aligned with a study by McMillan and Nash (2000), who speculated that grading for effort might dilute student achievement. They suggested that students might associate hard work with achievement, instead of pure language ability. Laura worried that her grades were not accurate depictions of her students' true language achievement because of her 25% Classwork category.

Martin was an outlier in the group who prioritized effort, work ethic, and productivity in 100% of his gradebook assessments. His assessment and grading beliefs and practices aligned with a Humanistic approach to teaching languages. Stevick (1990) explained that humanism in the language classroom emphasizes students' feelings, relationships and cooperation,

responsibility, the free use of the intellect, and self-actualization (as cited in Shirkhani & Ardeshir, 2013). A Humanistic approach values student self-expression and freedom. The teacher's role is that of a facilitator, as learning becomes more important than teaching. Assessments are meant to highlight students' abilities, with errors viewed positively and used for self-reflection as a guide for future learning. Martin's stated life goals for his students were for them to discover and develop their strengths and talents, and to improve their work ethic, while his intellectual goal for his students was for them to understand the concepts and framework of the language. Martin graded all his assessment holistically, based on these goals, with the endgame of setting his students up "to be successful in the more important parts of their lives." While he recognized the importance of teaching his language, he also concerned himself with developing his students as humans.

Several previous studies showed that some teachers and cultures valued effort, improvement, ability, study habits, and participation even more than achievement (Nowruzi, 2012; Sun & Cheng, 2014). And another study demonstrated that teachers believed that these non-cognitive elements could be included in grading as part of students' moral character development (Mulyani, 2023). Martin considered the students' effort and achievement equally, in the way he holistically graded their weekly work, which showed that he valued both their work ethic and their language abilities. He also graded down their final exam projects if he watched their assessment process and saw that they put in no effort, in order to push his students to develop the kind of character that they would need in their adult lives. He might not have valued these non-cognitive elements more than their language achievement, as in the other studies, but he certainly valued them equally.

Martin was the only teacher that treated missing or incomplete work in accordance with the grading experts, who opposed the use of zeros as punishment (O'Connor, 2008). He stated his policy, "I normally don't give anything less than a fifty. In other words, it's failing, but it's still reachable." This policy was based on his Humanistic approach, in that he wanted to keep his students "in the game." Even when their assessments contained many errors, he said, "Purely for attempting, personally, I think it's worth a pass mark." Only when his students made no effort did he feel compelled to give a zero, saying "You aren't even playing in the game, and the reality is, now I've got, you've given me no choice." None of Martin's peers used this policy. They seemed to fall more in line with the teachers in the McMillan and Nash (2000) study that used zeros to motivate students, while allowing late work and second attempts.

Martin traced the origin of his Humanistic internally persuasive discourse back to a conversation that he had at the beginning of his career with "a guy over in the next county who's got a very big program regarded as being one of *THE* people [Martin's emphasis]" in his language. This Humanistic approach to teaching was the authoritative discourse that his colleague espoused, and Martin adopted it as his own. He explained, "Although I self-question, and I self-doubt . . . that's what I chose to do when I began teaching. And I am being reasonably faithful to that."

Martin's conversation with his colleague was one of the only instances throughout the interviews where a teacher traced the roots of their internally persuasive beliefs directly to a specific influence. Daniel also identified the source of his strong opposition to the use of translation in his assessments when he referred to his teacher preparation, saying, "When I was in school, they were always asking us, 'Don't do translation,' and that sort of thing. Okay, I keep that. And yeah, I don't translate." Maybe because of this early experience, Daniel was drawn to a

Communicative approach to assessment which avoids the use of translation. Two teachers, Diana and Julia, mentioned the adoption of other teachers' grading categories as the origin of their grading practices. Diana also borrowed a former colleague's time-saving strategy for grading the writing and speaking sections of her final exam. But these were borrowed practices, not beliefs. Martin and Daniel were the only two teachers in this study that could identify the sources of their internally persuasive discourses.

## Conclusion

These seven world language teachers worked in the same context, with the same constraints coming from the macro level of their zone of contact, such as the block schedule, the fear of losing enrollment and teacher allotments, and the requirement to have collaborative grading categories. Many teachers mentioned the same issues with class size and composition, along with a lack of available assessment resources at the meso level of their contact zone. And all the teachers identified the same problems that they encountered with their students at the micro level of their context, namely absenteeism, anxiety, disengagement, and cheating.

When it came to assessing the students in this context, all the teachers believed that their gradebooks should reflect student achievement. Even so, just as Leung (2004) predicted, there was "a great deal of variability within individual education systems and among teachers within a locality" (p. 880). In fact, no two teachers shared the exact same combination of internally persuasive beliefs about what this achievement meant, nor how or why it was to be assessed. Even though the teachers were working from the same set of ACTFL standards, it was quickly evident that "The publication and standardization of curriculum and assessment statements does not by itself lead to shared practice or perception" (Leung, 2004, p. 876). All these teachers adopted or modified more than one authoritative discourse regarding language achievement that

they used as the construct of their graded assessments, although the composite of these beliefs for each teacher differed from anyone else in the group. Some of these beliefs were clearly stated, while others were inferred from the teachers' gradebook and assessment practices.

Just as James (2006) postulated, these teachers used a blend of theories and approaches about language learning in their attempts to align their beliefs with their assessments. Their internally persuasive discourses were individually synthesized combinations of the historically accepted authoritative discourses about language teaching and learning. Below is a brief summary of the aggregate beliefs of each teacher, as evidenced by the assessments found in their gradebooks.

- Alicia came the closest to a "pure," non-synthesized approach in her implementation of assessments using her Situated Learning approach in her community of practice. Still, she mentioned using Behaviorist "repetition" in her instructional practices, which also showed up in small percentages in her Test category.
- Carmen indicated a close alignment with a Cognitive approach, believing that language achievement meant understanding concepts, while also utilizing some of the assessment practices of an accumulated knowledge model of learning, pointing to a partial adaptation of Behaviorist ideology.
- Daniel assessed his students using a Communicative approach, believing strongly that language achievement meant using the language rather than analyzing the language. However, he built his assessment sequence and his gradebook according to component-to-composite principles, which were most closely associated with Behaviorism.
- Diana also assessed based on a Communicative approach, as evidenced by her high percentage of listening. But she also emphasized the negotiation of meaning in her

interpersonal speaking and writing assessments, showing a modified form of Interaction Hypothesis in her assessment, grading, and feedback practices.

- Julia used a Communicative approach by way of her interpersonal speaking assessments, but she leaned heavily into an Audiolingual approach with her recitation and reading aloud assessments. In opposition to these two approaches that insisted on using L2 exclusively, she also incorporated translations of word lists into her assessments, which mimicked a Grammar-Translation approach.
- Laura's blend of approaches was eclectic like Julia's. She showed strong beliefs that language achievement related to a Behaviorist approach of accumulating knowledge, placing a premium on memorization and discrete-point assessments. However, she also assessed her students' Cognitive understanding of grammatical concepts, insisting that they show their mental reasoning. But in her assessment practice and in her beliefs, she also showed her alignment with a Communicative approach.
- Martin assessed his students holistically, both on their cognitive abilities to analyze the language via a Grammar-Translation approach, and on their effort, work ethic, and productivity in the class. He clearly articulated his Humanistic stance about the purpose of language learning, to develop his students' talents and character to set them up for life. He also used some Constructivist task-based learning as part of his assessment profile.

These findings supported Leung and Scarino (2016) who stated that language learning beliefs are sometimes incoherent patchworks of ideas and aspirations. These hodgepodge formulations could be clearly seen in these teachers' internally persuasive discourses about language learning.

The final chapter will discuss why these findings are important, and how they can possibly be used to enhance world language teachers' understandings of assessment and grading.

# Chapter 7 Implications

Leung (2004) stated that "teachers exercise considerable amount of asymmetric power over students in regulative and pedagogic decision making in terms of both day-to-day classroom construction of the curriculum and longer term trajectories of student progress" (p. 874). Because of this power, teachers' decision-making process in their assessment and grading practices need to be "theoretically and empirically" (p. 874) examined. This study has been an attempt to do just that. Prompting these world language teachers to explain the assessments found in their own gradebooks brought about two processes—self-reflection and collaboration.

## Self-Reflection

The interviews gave these teachers the opportunity to articulate and consolidate their internally persuasive discourses about their practice, and to express pride in the assessments that they used with their students. Throughout the interviews, the teachers were confronted with their own practices, and as their interviewer, I was privy to their real-time, out-loud self-reflection as we discussed the what, the how, and the why of their assessments. These reflections centered around making changes in their practice and recognizing gaps or weaknesses in their gradebook categories, assessments, and methods.

### Change

Alicia, Laura, and Martin reflected aloud about their gradebook categories, with Alicia looking toward the future, saying "I'm thinking about changing it next year." Meanwhile Martin recalled his past attempts to add more categories, but how recently he reverted back to his previous ways explaining, "I invariably find I'm too lazy to do that, and so I drop it and then go back to the minimum number." Laura wondered aloud for several minutes about the possibility

of changing the structure and weighting of her categories by dropping or reducing the Classwork category.

Julia and Martin thought about the changes that they had made to their past assessments. Julia noticed that the assessments in the gradebook that we were examining together looked different than the gradebook assessments that she was currently working on, saying, “I change the assignments constantly. If you go to my level one right now, these assignments will not be there.” Martin admitted, “I come up with ideas which I don't think through very often” and “I'm kind of making things up as I go along.” His attempts at several new types of assessments in his gradebook showed that he was willing to take risks while he also self-assessed the effectiveness of the changes he made to his practice. Alicia mused out loud about a new rubric element that she wanted to add to one of her future assessments, “I'm gonna input creative, how creative they are,” as she planned for the next year.

**Gaps**

Even though the interviews were never meant to judge the teachers, several of the teachers self-identified gaps in their own assessment and grading practices. Carmen felt that she could have used more descriptive gradebook labeling. Regarding her speaking assessments, she said, “I need to assess them and make it a priority,” adding, “I need to have it in there, so that the four domains are very important. I say it on the syllabus. But it's not reflective in my gradebook.” Diana wondered to herself about how to add more projects into her assessment portfolio to increase student engagement. But she self-questioned, “How would I make sure that everybody's working? And you know, it's not just two people working and the other ones on their phone the whole time . . . How would I do that to make it more engaging with projects?” Julia wished out loud that she could learn more about implementing rubrics, because they made her

feel very uncomfortable. Laura mused about the possibilities of incorporating more informal interpersonal speaking assessments in her gradebook. And Martin struggled to understand why his other summative assessments did not show up in his gradebook, saying, “There was no plan from that point of view. If you're using the past tense, the answer is there's no particular thought to it.”

**So what?**

Leung’s (2013) discussion about teachers’ language learning perspectives and orientations showed that it was “important that the relationship between these language-related views and assessment be made explicit.” (p. 7). Teachers should be encouraged to look for gaps or areas of weakness in their assessment practices as evidenced by their gradebooks to question whether they were caused by oversight, a lack of time, or a true disconnect between their internally persuasive discourse about language learning and the usefulness of the skill or standard that should have been assessed. For example, assessing interpersonal speaking might not be a high priority for a teacher with an analytical Cognitive approach to language teaching. Nor would interpretive listening be the most effective method to assess students’ accumulated knowledge for a teacher who sees language learning as amassing vocabulary. But teachers could use the standards to self-reflect on the changes that they need to make in their practice in order to close those gaps, regardless of what they internally believe about the missing standard.

Furthermore, understanding their own personal language learning theories and approaches could help teachers “analyze and understand the often implicit language models and assumptions underlying externally produced tests and assessment frameworks” (Leung, 2013, p. 7). As James (2006) explained, “By using models of assessment borrowed from elsewhere, teachers may find themselves subscribing, uncritically or unwittingly, to the theories of learning

on which they are based" (p. 3). Most of the teachers in this study used at least some commercially produced testing material. Several of them mentioned having to curate or modify the materials to make them usable for their context. Understanding their own beliefs could enhance those modifications to make the commercially published assessments align better with the teachers' own internally persuasive discourses.

Participating in one-on-one interviews at the end of every semester with a researcher to uncover their perspectives and orientations is not a practical way to stimulate teachers' self-reflection. But teachers could be encouraged to think about their beliefs about language learning and how they apply those beliefs to their assessment and grading practices. They could also take time at the end of a semester to scour their gradebooks, looking for hidden gaps and weaknesses.

One possible solution could be to create a personal assessment checklist, related to the standards, for teachers to track how many times, or at what percentage of their overall gradebook, that they assessed the different standards throughout a given course. This checklist could provide fodder for private consideration, or could be shared with a collaborative group, to contribute ideas and receive support. Future assessment goals could be decided privately or collectively, to raise gradebook percentages for certain standards, and to push teachers out of their comfort zone to try new ways of assessing to broaden their internally persuasive beliefs about language learning approaches.

## Collaboration

Another discovery during these interviews was the evidence of collaboration and team spirit. As Alicia put it, "We're supposed to be together, teaching together." Carmen added, "We're on a team." Laura talked about working with Diana and making a joint decision about their students' use of a resource, saying, "We talk about it. Okay, what do you think? Are they

ready?  Should we let them use it?" Martin talked about his "interface" with the department, saying, "When we've had discussions amongst ourselves, I listen to what you guys do. And naturally, very often, what you guys do makes sense." These statements pointed to a strong sense of professional camaraderie among this group of teachers.

**Copying, sharing, and borrowing**

These teachers showed evidence of copying other teachers' grading and assessment practices. Diana talked about inheriting a class from a prior colleague and copying her gradebook categories, "I just continued doing what she was doing . . . because I was like, okay, yeah, it works." Julia also said about her categories, "I think I copied that from someone—horrible!" Teachers also copied from others' grading and assessment methods. Martin talked about interviewing his students in the hallway, "in the same way as other languages do," and Diana used a previous colleague's grading methods to save time when grading the speaking and writing sections of the final exam. The teachers mentioned sharing and borrowing resources and ideas for assessments, like a video worksheet, a "little man" image that the students had to label with body parts, and a board game project idea. These examples of collaboration pointed to an untapped resource that could aid teachers in honing their assessment and grading practices.

**So What?**

Leung (2004) stated that "Group moderation can clearly play an important part in promoting some degree of consistency in the outcomes of teacher assessment" (p. 880). This already-established collaborative community could tap into their own resources of knowledge and specialties to copy, share, and borrow ideas to improve their assessment and grading practices as they self-reflect about category changes or assessment gaps. Brumen and Cagran (2011) found that sharing assessments sparked collegial discussions that caused teachers to think

about and change their teaching practices. Group or peer collaboration about assessment could provide an outside perspective when self-reflection is not enough. At the end of her interview, Alicia turned the tables to become the interviewer, saying, “May I ask you a question based on what you hear? As a teacher, peer, what's your insight? What's your input of me as a teacher? What could I do better?” Despite her confidence in her abilities, Alicia knew that she could still have blind spots in her grading and assessment practices. Martin also fully recognized the power of the interviews that we were conducting, “What you're doing right now, I think, is a good idea.” He continued, saying that we should “discuss, at the department level, if there should be more planning of our gradebooks. Because we can all do the same thing, but still be independent, individual.” Even after the interview, he texted me with more thoughts about a specific assessment, showing that the collaborative nature of our interview continued to spark self-reflection well past the initial conversation. In his words, “Whoa! This does have a beneficial impact on the whole department!”

I came to this department as their peer and colleague when I asked to interview them, without claims of being an expert. I made sure they knew this before we started, saying things like, “There's no right or wrong” and “It's not a judgment.” I was in a unique position as no longer just their colleague, but also a researcher. Leung (20024) talked about the advantages of this type of research, saying,

> The researcher is positioned as an interested outsider whose claim to empirical knowledge crucially depends on access to actual and teachers’ perceptions of classroom practice. This kind of researcher–teacher collaboration clearly has a professional development dimension with teachers’ views and practices serving as starting points of discussion, not as “problems” to be fixed through in-service development activities (p. 884).

During their interviews, these teachers' views and practices gave me much fodder for my own self-reflection and inspiration to try new assessment methods in the classroom throughout the course of the research process. I was surprised and enlightened by the variety of beliefs that these teachers tapped into and the assessments that they employed, even though I have worked on the same hallway with them for years. Seeing some of their strengths highlighted some of my weaknesses as a classroom assessor, and I began thinking deeply about my own internally persuasive discourses and what authoritative discourses that I have assimilated, rejected, or modified throughout my career.

As their researcher, as Leung (20024) posited, I am in a position to "provide opportunities of critical engagement with information and theories emanating from relevant fields of research and education" (p. 884). I believe that these opportunities can stem from understanding our differing internally persuasive discourses and the authoritative discourses that helped shape them. From there, the collaborative group could take a deeper look into our gradebook categories and use informal, reciprocal training to enhance all our assessment practices as we share and collaborate together.

### *Recalibrating Gradebook Categories*

Several teachers mentioned making changes to their gradebook categories, either in the past or in the future, with Alicia making changes right in the middle of the interview. Laura was very clear that she felt that the Classwork category did not reflect the students' achievement. Several teachers posted assessments in categories that were not homogenous with the rest of the assessments in that category, or they underused a category, like Homework.

In this group of teachers, the singletons, who are the only teachers of their language, can make individual decisions about their categories and weighting. For those teachers, I propose a

hard, self-reflective look at the categories that are not working, ones that contain haphazardly placed assessments, ones that have too high or too low of a weight. Perhaps the category needs a new name and a new focus. These would be personal redeterminations, based on individual internally persuasive discourse. Asking questions like, "What does student achievement look like for me?" and "What are my assessment priorities?" could help guide these teachers to revamp or rename their grading categories so that the purposes of the categories are clear to all stakeholders-- administration, students, parents, and to the teachers themselves, who will have to post assessments to those categories. This transparency about grades is important (Quinn, 2013), as grading systems are a shared understanding between teacher and students (Zoeckler, 2007).

For the teachers who work on small teams, these category changes would have to be joint ventures, negotiated with an understanding of each of the members' assessment approaches and views on achievement in mind. Perhaps collaborative discussion about members' internally persuasive discourses, as well as a mini crash course on how those discourses were adopted or adapted from historical authoritative discourses about assessment approaches could spark new ideas about category uses and weights. In this case, a researcher-moderator could be effective in guiding some of these discussions, as Leung (2004) suggested.

Often teachers get into ruts in their profession and need fresh ideas for revitalizing floundering practices. As teachers find gaps in their assessment practices through self-reflection and collaboration, they could apply new, targeted grading category names to help fill those gaps. Ken O'Connor (2008) recommended organizing grades by learning goals, as one of his fixes for broken grades. Laura, Julia, and Diana already partially adopted this recommendation with their "Reading/Listening/Writing/Speaking" category. However, this category could be broken apart to give each skill more specialized attention. Instead of viewing this category as just an

alternatively named Classwork category, these teachers could break off Speaking into its own category, to isolate and emphasize interpersonal and presentational speaking, with either classwork to learn the skill, homework to practice independently, quizzes to formatively assess, or tests to summatively assess students' abilities. To give varying weights for instructional activities as compared to summative assessments, these teachers could use points instead of percentages. For example, five points could be given for a speaking homework assignment, ten points for classwork speaking, twenty points for a speaking quiz, and thirty points for a speaking test.

On the other hand, Martin, who had three categories—Formative, Summative, and Final Exam—could further reduce his number of categories due to his holistic grading approach. To match this holistic approach, he could have just one category for his weekly reviews, along with his Final Exam category. Within the weekly review category, he could use a point system rather than a percentage system where he could give more emphasis to one of his Constructivist task-based projects than he would normally give to his weekly translation, note-taking, and grammar reviews. For example, his projects could be valued at twenty points, while his weekly reviews could be valued at ten. The point numbers don't matter as much as the ratios, and the title of the holistic gradebook category could be anything that aligned with Martin's achievement goals for his students, rather than an attempt to apply a traditional, overused, meaningless category title.

These are just two ideas of many that teachers could implement to creatively refresh their gradebook practices to continue to allow for flexibility in their assessment approaches while maintaining unity and consistency in their gradebooks in their collaborative groups. Prescribing a definitive solution for gradebook inconsistencies is beyond the scope of this research,

particularly when these decisions should be made locally within the collaborative group to best align with the contact zone of the department and the school context requirements.

***Mutual Training***

Another advantage of an established collaborative community like this group could be mutual training. Each teacher came with a strength that they could share with others, and each teacher came with a weakness that could be aided by another teacher's strength. When these teachers shared resources and ideas and methods with each other, they were not coming from a hierarchically different plane, but rather as equals. In the same way, each teacher, even though they may not consider themselves to be experts, could informally share their assessment and grading strengths with the group without necessarily having professional-level expertise. In this particular case, their strengths were identified for them through research, but these strengths might also be discovered through guided collaborative discussions or self-reflection.

- Alicia's strength lay in building community through her group and partner projects, as well as establishing an atmosphere of interdependence among her learners in their Community of Practice.
- Carmen showed strong differentiation strategy skills, which are highly sought after in this school system, considering that teachers are evaluated on this specific ability during their formal and informal observations.
- Daniel's structured, consistent gradebook was a wonder to behold, showing that he had mindfully considered chronology while making a concerted effort to include and adhere to all the standards.

- Diana has received training and has extensive experience with the AP rubrics, which she has also learned to modify for other levels as she has developed engaging cultural speaking and writing prompts for her intermediate level classes.
- Julia's frequent use of recitation and interpersonal speaking activities lowered her students' affective filter while making them feel comfortable with the language's structure and proficient in their pronunciation.
- Laura was the queen of memorization techniques, always on the lookout for songs, drills, mnemonic tricks, and games to shore up her students' vocabulary and grammar basics.
- Martin excelled at creating engaging, project-based learning activities for his students to construct their knowledge while simultaneously developing their individual talents.

These teachers could take turns conducting mini lessons focused on their area of strength on departmental professional learning days or in collaborative meetings. Or they could just be identified within the department as the "go-to" person for questions and advice about their area of expertise.

***Issues with Collaboration***

Collaboration among teachers with differing approaches to teaching languages and beliefs about student achievement could be a catalyst for positive change and improvement, but it might also cause issues when agreements must be made about what, how, and why to assess. This issue would arise in departments where administration requires common assessments. For example, a teacher with a strong Communicative approach, whose assessments center around the student using the language, might struggle to find common ground with a teacher with a Cognitive approach. The teacher with a Cognitive approach would likely grade the students on whether they understood the concepts of the language through analysis, with right-or-wrong,

closed-ended questions, rather than the open-ended, rubric-graded kinds of assessments that a teacher with a Communicative approach would prioritize. A teacher who sees student achievement as the accumulation of knowledge and assesses based on how many language elements have been stored in the students' memory through closed-note quizzes and tests would balk at a Constructivist's open-note, task-based approach of allowing students to explore the language, building knowledge through experimentation and inquiry. A teacher coming from a Situated Learning approach that encourages use of the language to participate in a community of practice would struggle to accommodate a Behaviorist emphasis on differentiated individual performance on hierarchically assembled skills testing. These conflicting ideas about achievement would have to be negotiated among a group of teachers in a professional learning community to design common assessments that appease all the teachers' approaches to language learning. While this collaboration could aid in expanding the teachers' views about learning by adding new tools to their assessment arsenals, the process of compromise and conciliation could be painful, difficult, and possibly in the end, unsuccessful.

This section discussed the local implications of this research on the individuals and the group involved in the interviews. These were (a) a focus on self-reflection about gaps in assessment and grading practices, which could result in positive change, (b) collaboration to gain a better understanding of teachers' beliefs, to realign their gradebook practices, and to share assessment and grading strengths in the form of mutual training, and (c) issues in collaboration efforts between teachers who do not share similar internally persuasive beliefs about assessment. The next section will show how this study contributes to the global body of literature about world language assessment and grading.

## Contribution to the Field

This study was designed to fill the gap found in the American context of world language assessment research, since most recent language assessment studies were conducted abroad in contexts where English was a highly sought language (e.g., Bahtiar & Purnawarman, 2020; Berry et al., 2019; Brumen & Cagran, 2011; Lam, 2015; Lam, 2019; Mede & Atay, 2017; Nowruzi, 2021; Sun & Cheng, 2014; Tsagari & Vogt, 2017). It was also intended to explore teacher beliefs that influenced their assessment and grading practices, which both Bachman (2007) and James (2006) believed could be tightly related to their beliefs about learning and knowledge. McMillan and Nash (2000) found that teachers' philosophies about learning factored heavily in their grading processes, but this study focused on language and math teachers. This current study focused on beliefs specific to learning languages, which have evolved vastly over the past one hundred years. Based on the myriad of newly developed authoritative discourses about language learning, it was not surprising to find a variety of those learning beliefs deeply embedded into this group of teachers' assessment practices.

One global implication of this tiny study related to the ways that teachers' beliefs affected their implementation of similar assessments. Their internally persuasive discourses influenced them to change the "what," and the "why" of the assessment, even if they used the same "how." For example, while Martin administered an interpersonal speaking assessment that he saw implemented by his colleagues Daniel and Julia, he assessed the students on their effort and their cognitive grasp of the language, which aligned with his Humanistic and Cognitive approaches to language learning, whereas Daniel and Julia were looking for communicative competence. And one of Laura's open-ended presentational writing assignments was not used to assess her students' communication skills, but rather their conceptual understanding of the language. This

could mean that two teachers could administer the same assessment but evaluate their students' performances from different approaches to reach completely different conclusions about the students' abilities, therefore complicating the meaning of the grade that appears in the gradebook.

Another potential implication derived from looking at gradebook percentages for the assessments was that the teachers seemed to avoid assessing the skills that did not closely align with their beliefs about learning. Carmen had a strong analytical Cognitive approach to language learning and had a very low percentage of speaking assessments in her gradebook. Speaking assessments would evaluate the students' communicative abilities, but that was not Carmen's focus. Laura had an equally low percentage for interpretive listening and interpersonal speaking. Her approach aligned most closely with Behaviorist ideology, which valued the accumulation of knowledge rather than using that knowledge for receptive or productive communication. Without deep analysis of the teachers' gradebooks through self-evaluation or collaboration, these gaps could go unnoticed and unaddressed. The grades in the gradebook might then be an incomplete picture of the students' language achievement if all the standards are not assessed throughout a course.

## Recommendations for Future Research

The next step for this particular research project could be to apply my implications to this same set of teachers, by observing and collecting data on their self-reflective and collaborative efforts to recalibrate their gradebooks, and to track possible changes in their assessment practices that might occur from their mutual training in a longitudinal study. The results could be used to provide insights about how teachers' internally persuasive beliefs change over time with direct exposure to other authoritative discourses held by their colleagues and peers.

More research could also be done to trace the roots of teachers' internally persuasive discourses and how the teachers came to believe the way that they do. Martin and Daniel are the only teachers in the group that hinted at the origins of some of their beliefs. A narrative study of the teachers' careers could help to uncover the sources of their ideologies, whether they developed from their experiences learning their second language as a child or a student, or if they identified closely with the authoritative discourses taught in their pre-service teacher training, or if their beliefs came from on-the-job experiences, other colleagues, professional learning, or other unknown sources. This research could reveal important information about the impact of pre-service teacher programs or professional learning.

This study collected data not just from teacher interviews, but also from the teachers' gradebook reports, similar to Stiggins et al. (1989) and Zoeckler (2007), which made it more methodologically robust than some other studies that solely relied on interviews, surveys, and questionnaires. However, a more in-depth study could be conducted, similar to studies by Arias and Maturana (2005) and Frodden et al. (2004), to include teachers' actual assessments, complete with rubrics, assessment guidelines, and any self- or peer-evaluation scales. This more expanded data set could provide even more insights into teachers' beliefs and practices about assessment and grading.

## Limitations

Clearly, this was a very small study, in a highly contextualized setting. While many of the factors in this particular department were similar to the factors in previous studies regarding language assessment, there were still differences between this teaching contact zone and the ones previously studied. The participants themselves made this study unique. The teachers were all over the age of forty, each with at least fourteen years of teaching experience. There were no

young teachers, fresh out of pre-service training, to bring new, vibrant assessment and grading ideas to the study. At least one of these teachers started their career before the first publishing of the *ACTFL Standards for Foreign Language Learning: Preparing for the 21st Century* (The National Standards Collaborative Board, 2015) in 1996, meaning that they were acclimating to the developing standards right along with their students. The participants were also from a diverse group of nationalities, and their exposure to language learning in international contexts could have affected the way they believed about language learning. The findings from this study would likely not be replicated in other contact zones with differing participant backgrounds.

## Conclusion

While teachers' grading practices are normally private, (McMillan and Nash, 2000) these world language teachers opened up their gradebooks to talk about their assessment practices, revealing a wide variety of internally persuasive beliefs about language learning and student achievement. Hopefully this research can inspire deeper self-reflection and more candid collaboration about what, how, and why teachers assess, and how that translates to their gradebook.

## References


ACTFL. (2025). *ACTFL Proficiency Guidelines 2024.* ACTFL Language Connects. https://www.actfl.org/educator-resources/actfl-proficiency-guidelines?resources=view

ACTFL. (2025). *Save a language program.* ACTFL Language Connects. https://www.actfl.org/advocacy/save-a-language-program

Adair-Hauck, B., Glisan, E. W., Koda, K., Swender, E. B., & Sandrock, P. (2006), The integrated performance assessment (IPA): Connecting assessment to instruction and learning. *Foreign Language Annals, 39,* p. 359-382. https://doi.org/10.1111/j.1944-9720.2006.tb02894.x

Alderson, J. C., Brunfaut, T., & Harding, L. (2017). Bridging assessment and learning: A view from second and foreign language assessment. *Assessment in Education: Principles, Policy & Practice, 24*(3), p. 379-387. https://doi.org/10.1080/0969594X.2017.1331201

American Councils for International Education. (2017, June). *The National K-12 foreign language enrollment survey report.* https://www.americancouncils.org/language-research-fle-state-language-us

Arias Toro, C. I., & Maturana Patarroyo, L. M. (2005). Evaluación en lenguas extranjeras: Discursos y prácticas. *Ikala, Revista de Lenguaje y Cultura, 10*(16), p. 63-91.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41– 71). Ottawa: University of Ottawa Press.

Bahtiar, I., & Purnawarman, P. (2020). Investigating English teachers' comprehension in language assessment literacy (LAL). *Proceedings of the 1st International Conference on*

*Information Technology and Education (ICITE 2020).* Atlantis Press. https://doi.org/10.2991/assehr.k.201214.253

Bakhtin, M. (1981). *The dialogic imagination: Four essays.* University of Texas Press.

Berry, V., Sheehan, S., & Monro, S. (2019) What does language assessment literacy mean to teachers? *ELT Journal, 73*(2), p. 113-123. https://doi.org/10.1093/elt/ccy055

Bloom, B. S. (1968). Learning for Mastery. *Instruction and Curriculum.* Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1.

Bonner, S. M., & Chen, P. P. (2019). The composition of grades: Cognitive and noncognitive factors. In T. R. Guskey & S. M. Brookhart, (Eds.), *What we know about grading : What works, what doesn't, and what's next* (pp. 57-83). ASCD.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), p. 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M. (1991), Grading practices and validity. *Educational Measurement: Issues and Practice, 10,* p. 35-36. https://doi.org/10.1111/j.1745-3992.1991.tb00182.x

Brookhart, S. M. (1993). Teachers' grading practices: Meanings and values. *Journal of Educational Measurement, 30*(2), 123-142. https://doi.org/10.1111/j.1745-3984.1993.tb01070.x

Brown H. D., & Abeywickrama P. (2019). *Language assessment : Principles and classroom practices* (Third). Pearson Education.

Brumen, M., & Cagran, B. (2011). Teachers' perspectives and practices in assessing young foreign language learners in three Eastern European countries. *Education 3-13, 39*(5), p. 541-559. https://doi.org/10.1080/03004279.2010.488243

Campbell, C., & Duncan, G. (2007). From theory to practice: General trends in foreign language teaching methodology and their influence on language assessment. *Language and Linguistics Compass 1*(6), p. 592–611. https://doi.org/10.1111/j.1749-818X.2009.00183.x

Carreira, M. M. (2012). Formative assessment in HL teaching: Purposes, procedures, and practices. *Heritage Language Journal, 9,* p. 100–120. https://doi.org/10.46538/hlj.9.1.6

Carspecken, P. H. (1996). *Critical ethnography in educational research: A theoretical and practical guide*. Routledge.

Celce-Murcia, M. (2001). *Teaching English as a second or foreign language* (3rd Edition), Heinle & Heinle Publisher, Boston.

Cheng, L., & Wang, X. (2007). Grading, feedback and reporting in ESL/EFL classrooms. *Language Assessment Quarterly, 4*, p. 85-107. https://doi.org/10.1080/15434300701348409

Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing, 21,* p.360-389. https://doi.org/10.1191/0265532204lt288oa

Chiekem. E. (2023). Grading practice as valid measures of academic achievement of secondary schools students for national development. *Journal of Education and Practice, 6*(26), p. 24-28.

Corden, A., & Sainsbury, R. (2006). *Using verbatim quotations in reporting qualitative social research: Researchers' views* (pp. 11-14). York: University of York.

Cox, K. B. (2011). Putting classroom grading on the table: A reform in progress. *American Secondary Education, 40*(1). p. 67-87. https://www.jstor.org/stable/23100415

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory Into Practice, 39*(3), p. 124–130. http://www.jstor.org/stable/1477543

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28,* p. 43-56. https://doi.org/10.1016/j.asw.2016.03.001

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), p. 305-334. https://doi.org/10.1191/0265532204lt286oa

DeLuca, C., Coombs, A., & LaPointe-McEwan, D. (2019). Assessment mindset: Exploring the relationship between teacher mindset and approaches to classroom assessment. *Studies In Educational Evaluation, 61*, p. 159-169. https://doi.org/10.1016/j.stueduc.2019.03.012

Denzin, N. K. (1989) *The research act: A theoretical introduction to sociological methods* (3rd Edition). New Jersey: Prentice Hall.

Denzin, & Lincoln. (2017). The discipline and practice of qualitative research. In N. K. Denzin, Y. S. Lincoln, M. D. Giardina, & G. S. Cannella (Eds.), *The SAGE handbook of qualitative research* (pp. 1-26). SAGE Publications, Inc.

Devlin, K. (2015). Learning a foreign language a 'must' in Europe, not so in America. Pew Research Center. https://www.pewresearch.org/short-reads/2015/07/13/learning-a-foreign-language-a-must-in-europe-not-so-in-america/

Dey, Ian. (1993). *Qualitative data analysis: A user friendly guide for social scientists* (1st edition). Routledge. https://doi.org/10.4324/9780203412497

Diaz Larenas, C., Alarcón, P., & Ortiz, M. (2012). El profesor de inglés: sus creencias sobre la evaluación de la lengua inglesa en los niveles primario, secundario y terciario. *Íkala, Revista de Lenguaje y Cultura, 17*(1), p. 11-22.

Downer, J. T., López, M. L., Grimm, K. J., Hamagami, A., Pianta, R. C., & Howes, C. (2012). Observations of teacher–child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly, 27*(1), p. 21-32. https://doi.org/10.1016/j.ecresq.2011.07.005

Duncan, C., & Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *Alberta Journal of Educational Research, 53*(1), p. 1-21.

East, M. (2021). Foundational principles of task-based language teaching. Routledge. https://doi.org/10.4324/9781003039709

Elder, C., McNamara, T., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us. *Language and Communication, 57*, p. 14-21. https://doi.org/10.1016/j.langcom.2016.12.005

Ellis, R. (1991). The interaction hypothesis: A critical evaluation. In E. Sadtono (Ed.), *Language acquisition in the second/foreign language classroom* (pp. 179–211). Anthology Series 28, Singapore: SEMEO Regional Language Centre.

Erickson, F. (2004). Demystifying data construction and analysis. *Anthropology & Education Quarterly, 35*(4), p. 486–493. http://www.jstor.org/stable/3651351

Febria, D., & Marsevani, M. (2024). The post-pandemic reflection on students' speaking anxiety: A case in higher education. *Linguists: Journal of Linguistics and Language Teaching, (10)*1, p. 1-14. DOI : http://dx.doi.org/10.29300/ling.v10i1.4492

Frary, R., Cross, L., & Weber, L. (2005). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12,* p. 23-30. https://doi.org/10.1111/j.1745-3992.1993.tb00539.x

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*(2), p. 275-301. https://doi.org/10.1017/S0272263104262064

Frey, B. B., & Schmitt, V. L. (2010) Teachers' classroom assessment practices. *Middle Grades Research Journal, 5,* p. 107-117.

Frodden Armstrong, M. C., Restrepo Marín, M. I., & Maturana Patarroyo, L. (2004). Analysis of assessment instruments used in foreign language teaching. *Íkala, Revista de Lenguaje y Cultura, 9*(15), p. 171-201.

Giraldo, F. (2019). Language assessment practices and beliefs: Implications for language assessment literacy. *HOW, 26*(1), p. 35-61. https://doi.org/10.19183/how.26.1.481

Greeno, J. G., Collins, A, & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46), New York: Macmillan.

Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *The Journal of Educational Research, 79*(2), p. 96-100. https://doi.org/10.1080/00220671.1985.10885657

Guskey, T. R., & Brookhart, S. M. (2019). Conclusion: Where do we go from here? . In T. R. Guskey & S. M. Brookhart, (Eds.), *What we know about grading : What works, what doesn't, and what's next* (pp. 57-83). ASCD.

Guskey, T. R., & Link, L. K. (2019). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice, 26*(3), p. 303–320. https://doi.org/10.1080/0969594X.2018.1555515

Guzman-Orth, D. A., Lopez, A. A., & Tolentino, F. (2019). Exploring the use of a dual language assessment task to assess young English learners. *Language Assessment Quarterly, 16*(4-5), p. 447-463. https://doi.org/10.1080/15434303.2019.1674314

Hill, K., & Mcnamara, T. (2011). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing, 29*, p. 395-420. https://doi.org/10.1177/0265532211428317

James, M. (2006). Assessment, teaching and theories of learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 47–60). London, England: Sage.

Jamieson, J. (2014). Defining constructs and assessment design. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1-17). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla062

Jia, Y., Eslami, Z. R., & Burlbaw, L. M. (2006). ESL teachers' perceptions and factors influencing their use of classroom-based reading assessment. *Bilingual Research Journal, 30*(2), p. 407-430, https://doi.org/10.1080/15235882.2006.10162883

Kunnath, J. P. (2017). Teacher grading decisions: Influences, rationale, and practices. *American Secondary Education, 45*(3), p. 68-88. https://www.jstor.org/stable/45147907

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32*(2), p. 169-197. https://doi.org/10.1177/0265532214554321

Lam, R. (2019). Teacher assessment literacy: Surveying knowledge, conceptions and practices of classroom-based writing assessment in Hong Kong. *System, 81*, p. 78-89. https://doi.org/10.1016/j.system.2019.01.006

Larsen-Freeman, D. & Freeman, D. (2008). Language moves: The place of "foreign" languages in classroom teaching and learning. *Review of Research in Education, 32,* p. 147–186. https://doi.org/10.3102/0091732X07309426

Leung, C., & Scarino, A. (2016). Reconceptualizing the nature of goals and outcomes in language/s education. *The Modern Language Journal, 100,* p. 81-95. https://doi.org/10.1111/modl.12300

Leung, C. (2004). Classroom teacher-based assessment of second language development: Construct as practice. In E. Hinkel (Ed.), *Handbook of research in second language learning and teaching.* Mahwah, NJ: Erlbaum.

Leung, C. (2013). Classroom-based assessment issues for language teacher education. In A. J. Kunnan (Ed.), *The companion to language assessment.* John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla064

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic inquiry.* Newbury Park, CA: Sage Publications.

Lingard, L. (2019). Beyond the default colon: Effective use of quotes in qualitative research. *Perspectives on Medical Education, 8*(6), p. 360-364. https://doi.org/10.1007/S40037-019-00550-7

Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing, 28*(3), p. 367-382. https://doi.org/10.1177/0265532211404188

Long, M. (2014). *Second language acquisition and task-based language teaching.* John Wiley & Sons, Incorporated.

MacGregor-Mendoza, P. (2012). Spanish as a heritage language assessment: Successes, failures, lessons learned. *Heritage Language Journal, 9*(1), p. 1-26. https://doi.org/10.46538/hlj.9.1.2

Mackey, A. (2002). Beyond production: learners' perceptions about interactional processes. *International Journal of Educational Research, 37*(3/4), 379-394. https://doi-org.proxyiub.uits.iu.edu/10.1016/S0883-0355(03)00011-9

Maxwell, J. A., & Miller, B. A. (2008). Categorizing and connecting strategies in qualitative data analysis. In S. N. Hesse-Biber & P. Leavy (Eds.), *Handbook of emergent methods* (pp. 461–477). The Guilford Press.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*, p. 20-32. https://doi.org/10.1111/j.1745-3992.2001.tb00055.x

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22*, p. 34-43. https://doi.org/10.1111/j.1745-3992.2003.tb00142.x

McMillan, J. H., & Nash, S. (2000). Teachers' classroom assessment and grading decision making. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.

McNamara, T. (2007). Language assessment in foreign language education: The struggle over constructs. *The Modern Language Journal, 91*(2), *p.* 280-282. https://doi.org/10.1111/j.1540-4781.2007.00543_14.x

Mede, E., & Atay, D. (2017). English language teachers' assessment literacy: The Turkish context. *Dil Dergisi,* p. 43-60. DOI: 10.1501/Dilder_0000000237.

Mulyani, S., Rahman, A. A., Akhmad, F., & Ardin, H. (2023). EFL lectures' grading practices in Islamic and general higher education institutions. *Indonesian TESOL Journal, 5*(1), p. 14-27.

Muñoz, A. P., Palacio, M. B., & Escobar, L. (2012). Teachers' beliefs about assessment in an EFL context in Colombia. *Profile Issues in Teachers' Professional Development, 14,* 143-158.

Murdock, T. B., & Beauchamp, A. S. (2008). Cheating in academic contexts. In T. L. Goode (Ed.), *21st century education: A reference handbook.* Sage Publications, Inc. https://doi.org/10.4135/9781412964012.n95

National Academies of Sciences, Engineering, and Medicine. (2020). *A principled approach to language assessment: Considerations for the U.S. foreign service institute.* Washington, DC: The National Academies Press. https://doi.org/10.17226/25748

The National Standards Collaborative Board. (2015). *World-readiness standards for learning languages* (4th edition). Alexandria, VA: Author.

Nowruzi, M. (2021). A study of EFL teachers' classroom grading practices in secondary schools and private institutes: A mixed methods approach. *Language Testing in Asia 11*(29), p. 1-22. https://doi.org/10.1186/s40468-021-00145-2

O'Connor, K. (2008). *A repair kit for grading: 15 fixes for broken grades, a study guide.* Educational Testing Service.

Oleson, A., Hora, M. T. (2014). Teaching the way they were taught? Revisiting the sources of teaching knowledge and the role of prior experience in shaping faculty teaching practices. *Higher Education, 68,* p. 29–45. https://doi.org/10.1007/s10734-013-9678-9

Parson, Laura. (2019). Considering positionality: The ethics of conducting research with marginalized groups. In K. K Strunk & L. A. Locke (Eds.), *Research methods for social justice and equity in education* (pp. 15-32). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-05900-2_2

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal, 100*, p. 190-208. https://doi.org/10.1111/modl.12308

Quinn, R. (2023). *Foreign language enrollment sees steepest decline on record.* Inside Higher Ed. https://www.insidehighered.com/news/faculty-issues/curriculum/2023/11/16/foreign-language-enrollment-sees-steepest-decline-record

Quinn, T. (2013). *On grades and grading: Supporting student learning through a more transparent and purposeful use of grades.* EBSCO Publishing.

Randall, J., & Engelhard, G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement, 46*(1), p. 1–18.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*(7), p. 1372–1380. https://doi.org/10.1016/j.tate.2010.03.008

Renau Renau, M. L. (2016). A review of the traditional and current language teaching methods. *International Journal of Innovation and Research in Educational Sciences, 3*(2), p. 82-88.

Rogers, W., Cheng, L., & Hu, H. (2007). ESL/EFL instructors' beliefs about assessment and evaluation. *Canadian and International Education, 36*(1), p. 39-61. https://doi.org/10.5206/cie-eci.v36i1.9088

Savignon, S. J. (1987). Communicative language teaching. *Theory Into Practice, 26*(4), p. 235–242. https://doi.org/10.1080/00405848709543281

Schellekens, L. H., Bok, H. G. J., de Jong, L. H., van der Schaaf, M. F., Kremer, W. D. J., & van der Vleuten, C. P. M. (2021). A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL). *Studies in Educational Evaluation, 71,* Article 101094. https://doi.org/10.1016/j.stueduc.2021.101094

Scovel, T. (1978). The effect of affect on foreign language learning: A review of the anxiety research. *Language Learning, 28,* p. 129-142. https://doi.org/10.1111/j.1467-1770.1978.tb00309.x

Scriven, M. (1966). *Methodology of evaluation.* Social Science Education Consortium, Publication 110. https://eric.ed.gov/?id=ED014001

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27*(2), 4-13.

Shirkhani, S., & Ardeshir, D. (2013). Humanism in the foreign language classroom. *Journal on English Language Teaching, 3,* p. 1-5.

Simon, M., Chitpin, S., & Yahya, R. (2010). Pre-service teachers' thinking about student assessment issues. *International Journal of Education, 2.* https://doi.org/10.5296/ije.v2i2.490

Stein-Smith, K. (2019). *Foreign language classes becoming more scarce.* The Conversation. https://theconversation.com/foreign-language-classes-becoming-more-scarce-102235

Stiggins, R. (1991). Assessment literacy. *Phi Delta Kappan, 72,* p. 534–539.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989), Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8,* p. 5-14. https://doi.org/10.1111/j.1745-3992.1989.tb00315.x

Stoynoff, S. (2012). Looking backward and forward at classroom-based language assessment. *ELT Journal, 66*(4), 523-532. https://doi.org/10.1093/elt/ccs041

Sun, Y., & Cheng, L. (2014). Teachers' grading practices: Meaning and values assigned. *Assessment in Education: Principles, Policy & Practice, 21*(3), p. 326-343. https://doi.org/10.1080/0969594X.2013.768207

Sunstein, B. S., & Chiseri-Strater, E. (2012). *Fieldworking: Reading and writing research* (4th ed.). Boston, MA: Bedford/St. Martin's.

Tian, W., Lê, H. D., & Nguyễn, N. T. (2023). Vietnamese university EFL teachers' practices and literacy in classroom assessment: A sociocultural perspective. *System, 116,* p. 1-15. https://doi.org/10.1016/j.system.2023.103066

Tierney, R. D., Simon, M., & Charland, J. (2011). Being fair: Teachers' interpretations of principles for standards-based grading. *The Educational Forum, 75*(3), p. 210-227. https://doi.org/10.1080/00131725.2011.577669

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment, 6*, p. 41-63. http://dx.doi.org/10.58379/UHIX9883

Troyan F. J. (2012). Standards for foreign language learning: defining the constructs and researching learner outcomes. *Foreign Language Annals: American Council on the Teaching of Foreign Languages, 45*(S1), pp. S118–S140. https://doi.org/10.1111/j.1944-9720.2012.01182.x

Usher, R. (1996). A critique of the neglected epistemological assumptions of educational research. In D. Scott & R. Usher (Eds.), *Understanding educational research* (pp. 9-32). Routledge. https://doi.org/10.4324/9780203131923

Villa Larenas, S., & Brunfaut, T. (2022). But who trains the language teacher educator who trains the language teacher? An empirical investigation of Chilean EFL teacher educators' language assessment literacy. *Language Testing, 40*(3), p. 463–492. http://dx.doi.org/10.1177/02655322221134218

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), p. 374-402. https://doi.org/10.1080/15434303.2014.960046

Vogt, K., Tsagari, D., & Spanoudis, G. (2020). What do teachers think they want? A comparative study of in-service language teachers' beliefs on LAL training needs. *Language Assessment Quarterly, 17*(4), p. 386-409. https://doi.org/10.1080/15434303.2020.1781128

Vogt, K., Tsagari, D., Csépes, L., Green, A., & Sifakis, N. (2020). Linking learners' perspectives on language assessment practices to teachers' assessment literacy enhancement (TALE):

Insights from four European countries. *Language Assessment Quarterly, 17*(4), p. 410-433. https://doi.org/10.1080/15434303.2020.1776714

Wang, L., Lee, I., & Park, M. (2020). Chinese university EFL teachers' beliefs and practices of classroom writing assessment. *Studies in Educational Evaluation, 66,* p. 1-11. http://dx.doi.org/10.1016/j.stueduc.2020.100890

Welsh, M. (2019). Standards based grading. In T. R. Guskey & S. M. Brookhart, (Eds.), *What we know about grading : What works, what doesn't, and what's next* (pp. 113-144). ASCD.

Widiastuti, S., Mukminatien, N., Prayogo, J., & Irawati, E. (2020). Dissonances between teachers' beliefs and practices of formative assessment in EFL classes. *International Journal of Instruction, 13,* p. 71-84. http://dx.doi.org/10.29333/iji.2020.1315a

Xu, Y. and Liu, Y. (2009). Teacher assessment knowledge and practice: A narrative inquiry of a Chinese college EFL teacher's experience. *TESOL Quarterly, 43,* p. 492-513. https://doi.org/10.1002/j.1545-7249.2009.tb00246.x

Yan, X., Zhang, C., & Fan, J. (2018). "Assessment knowledge is important, but …": How contextual and experiential factors mediate assessment practice and training needs of language teachers. *System, 74,* p. 158-168. http://dx.doi.org/10.1016/j.system.2018.03.003

Young, J. R. (2023, November 30). *As schools move to change how kids are graded, some families push back.* EdSurge. https://www.edsurge.com/news/2023-11-30-as-schools-move-to-change-how-kids-are-graded-some-families-push-back

Zoeckler, L. (2007). Moral aspects of grading: A study of high school English teachers' perceptions. *American Secondary Education, 35,* p. 83-102.

# Melinda Cotton

melcott@iu.edu

## ACADEMIC DEGREES

Ed.D., Indiana University Bloomington, Curriculum and Instruction, March 2025

Major: Literacy, Culture, and Language Education

Dissertation Title: *World Language Teachers' Assessment Beliefs and Grading Practices*

Committee: Karen Wohlwend (chair), Sharon Daley, James Damico

M.A., Spanish, Georgia State University, 2005

Thesis Title: *La Tradición Deformada: Surrealismo en el Romance de "Thamár y Amnón" de Federico García Lorca*

B.A., Spanish, Kennesaw State University, 1996

## PROFESSIONAL TEACHING EXPERIENCE

**Teacher (Spanish, French) at High School,** 2014- Present

- Taught levels I, II, and III of both Spanish and French
- Taught in a hybrid format, both in-person and online in 2020-2021
- Mentored teachers new to the school

**Teacher (Spanish, French) at Kell High School** in Marietta, Georgia, 2005-2014

- Taught up to level V in French and level II in Spanish
- Sponsored a local chapter of La Société Honoraire de Français
- Sponsored French Club
- Assisted with annual International Festival

**Adjunct Faculty at Kennesaw State University** in Kennesaw, Georgia, 2003-2005

- Taught Spanish 1001, 1002, 2001, 2002

**Spanish Teacher at The King's Academy** in Woodstock, Georgia, 1999-2005

- Developed and implemented Spanish curriculum for a new K-12 hybrid school
- Taught Kindergarten through tenth grade
- Served as department lead

**Contract Spanish Teacher** in Metro Atlanta, 1999-2003

- Designed, marketed, and taught Spanish classes to private groups
- Instructed all levels of students, children, and adults, tailoring curriculum to each group's needs

**Spanish Teacher at Burruss Elementary School** in Marietta, Georgia, 1998-1999

**Spanish Teacher in Bartow County Schools** in Cartersville, Georgia, 1997-1998

## OTHER PROFESSIONAL EXPERIENCE

**Teachers Pay Teachers Seller**, 2018 - present

- Designed, marketed, and sold over 300 French and Spanish teaching resources

**French Exchange Coordinator**, 2011- 2018

- Teamed with another local high school to place and chaperone 30 high school students from Rennes, France, in American high school students' homes for a 12-day cultural exchange in 2011, 2013, 2015, 2017
- Organized daily local field trips for the French high school students and their American hosts in metro Atlanta
- Organized and chaperoned reciprocal trips to the high school in Rennes, France in 2012, 2014, 2017, 2018 for 25 American students and 2-3 chaperones for 12 days, keeping costs down and safety a priority
- Organized and chaperoned group sight-seeing excursions for three days in Paris at the end of each trip
- Managed a budget of approximately $50,000 per trip

**Local Coordinator**, Loisirs Culturels à l'Etranger, (LEC), (2000 - 2002)

- Recruited metro-Atlanta families to host French teenagers for summer cultural exchanges
- Teamed with French chaperones to monitor families and students

## CERTIFICATION

T-6 Georgia Educator Certificate, French and Spanish, with Gifted Endorsement