AN EXPLORATION OF THE LAL DEVELOPMENT OF PRE-SERVICE ESL TEACHERS THROUGH THE PROCESSES OF ITEM WRITING

BY

ERIC CHEN PEI HO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Masters of Arts in Teaching English as a Second Language
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Master's Committee:

Assistant Professor Xun Yan, Chair
Professor Paul Prior

ABSTRACT

Language assessment literacy (LAL) has been recognized as an important component of language teacher development. Traditionally, TESOL programs have addressed the need of LAL for language teachers through language assessment courses and formal instruction. Such methods have proven valuable in helping teachers develop the theoretical knowledge of assessments. However, it is unclear how such instruction addresses some practical aspects of test development competencies, which include item writing. The current study investigates the features and practices of item writing for an institutional English Placement Test in a university context, and explores the potential of the activities of item writing in developing LAL for pre-service teachers. Employing a thematic analysis and adopting a community of practice perspective, this study examined data from recorded group discussions of participants, individual interviews, and draft comments of test items. The results suggest the importance of collaboration as an important competency of LAL that was developed through item writers' engagement through test development. Additionally, the study proposes the inclusion of content knowledge as a competency within the LAL framework. Finally, the potential of employing perspectives of genre studies in further research on item writing practices is discussed.

ACKNOWLEDGEMENTS

I would like to thank my advisor and chair, Dr. Xun Yan, for his time, patience, ideas, and generous support provided for me during the process of writing this thesis. I am also grateful to Professor Paul Prior for his guidance as my committee member and the time that we spent discussing the various concepts and ideas from writing studies, which have been a great source of inspiration for this project. I would also like to thank the participants of this study, including the coordinators of the EPT and members of the item writing team for their support and cooperation. Finally, I would to thank my family and my friends at UIUC who have helped and supported me during my time here.

TABLE OF CONTENTS

Chapter 1: Introduction

The knowledge and use of assessments, ranging from large-scale standardized tests to small-scale classroom assessments, have been recognized as key components of teaching and learning, both in the fields of general education (Popham, 2011) and language education (Shohamy, 2001). These competencies related to assessment are referred to as *assessment literacy* (AL) (Stiggins, 1991) and *language assessment literacy* (LAL) (Inbar-Lourie, 2008) respectively. Definitions of AL/LAL have been varied thus far in the literature; scholars in the field continue to identify new assessment competencies and reconceptualize the role of assessment in response to the varying contexts and needs of the stakeholders involved (Davies, 2008; Fulcher, 2012; Scarino, 2013; Stiggins, 1999; Popham, 2009; Taylor, 2009). Although there is not a standard definition of LAL, Brown and Bailey (2008) surveyed a sample of language assessment instructors and identified essential competencies such as theoretical knowledge of measurement, as well as the practical skills involved in test development; the goal of this current study is to investigate the role of LAL development in the context of the latter. Specifically, the study focuses on the development of pre-service ESL teachers, for which test development is a necessary assessment competency. Based on Taylor's (2013) visual representation of the various components of LAL, Baker and Riches (2018) offer a representation of LAL for language teachers, recognizing that language teachers have different LAL needs than test developers.

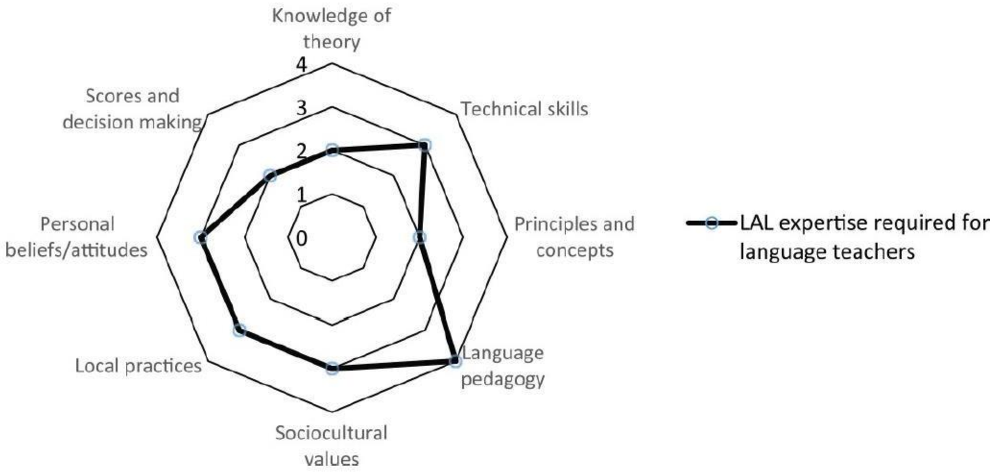


Figure 1.1. *Components of LAL for language teachers* (from Taylor, 2013)

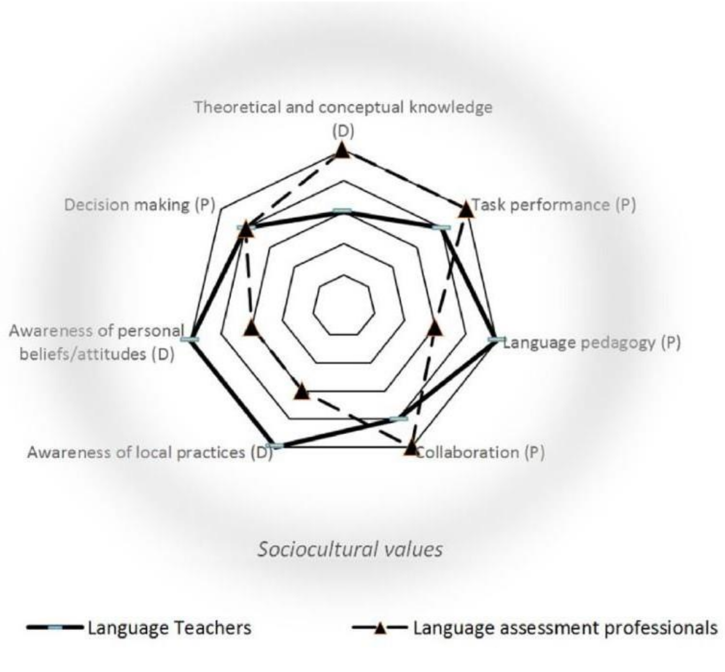


Figure 1.2. *Modified components of LAL for language teachers and language assessment professionals.* (from Baker & Riches, 2018)

Previous studies have acknowledged the importance of developing competencies in the theoretical knowledge of assessment as a component of LAL (Fulcher, 2012). Since the notion of LAL has appeared in the literature, there have been a considerable number of studies investigating how it can be developed in various stakeholders, especially language teachers (Brown & Bailey, 2008; Coombe, Troudi, & Al-Hamly, 2012; Scarino, 2013). Recent studies (e.g. Lam, 2015) evaluating the effectiveness of language assessment training in developing LAL in teachers continue this important work across different contexts. Although *theoretical and conceptual* knowledge has been identified as an important component of LAL, there is a notable lack of research in developing what Baker and Riches (2018) have identified as the procedural components of LAL, or the practical skills associated with language assessment. This study will focus on the latter in the context of item writing as a process during test development. In this context, it is assumed that LAL development will occur mostly through the *practice* of item writing rather than *instruction* from language assessment professionals (Kleinsasser, 2005). This current study attempts to address the LAL development of pre-service ESL teachers in a university graduate program, working together to develop an integrated writing assessment for placement purposes. One way to draw a connection between item writing and LAL is to examine the features of item writing that correspond to components of LAL identified in the previous literature (Taylor, 2013). To accomplish this, the study draws on concepts from genre studies to conceptualize item writing as the creation of *texts*, rather than simply test items. Therefore, item writing is not only the development of the products of assessment, but can be investigated as a process of individual and group LAL development. Writers create these test items through the social process inherent in collaborative writing, which include the practices of feedback, discussion, and revision. These practices eventually form recognizable genre practices of item writing within this particular context, through interactions between the individual and group practices of the writers. Thus, this study examines these practices to determine the values of the writers and features of item writing and how they align with LAL. Following this, the study examines the development of LAL through item writing and interactions of the writers by taking a *discourse socialization* approach, a theoretical perspective derived from linguistic anthropology (Ochs & Schieffelin, 1986).

Chapter 2: Literature Review

2.1 Defining LAL

Before LAL was established as a concept in language testing, scholars in the field of general education recognized the lack of assessment knowledge in teachers and the urgent need for assessment training to be one of the essential components of pre-service teacher education. The term “assessment literacy” was first coined by Stiggins (1991) in an article in which he described assessment illiterates, most notably teachers, as “easily intimidated by apparently technical information”, and lacking “the tools to be critical consumers of assessment data” (p. 535). Such a definition of AL referred mainly to the statistical and technical knowledge associated with the field of measurement and educational psychology. In an earlier study, Stiggins (1988) suggested that “classroom assessment specialists” be installed in schools to assist and train in-service teachers in responsibilities related to assessment. Similarly, Hills (1991) recommended a training program for school administrators and teachers and a system to monitor the assessment practices of in-service teachers. Schafer (1993) also recognized the importance of developing AL and called for the implementation of pre-service training programs and professional development opportunities. These earlier studies recognized the importance of assessment literacy for teachers, not only assessment professionals, and called for more rigorous training programs and courses for pre-service and in-service teachers.

The term *language assessment literacy* as distinct from AL appeared in the literature only in the past decade (Inbar-Lourie, 2008; Malone, 2008). Davies (2008) recognized the development of language assessment as a field, developing a focus on general assessment principles, specialized knowledge about language, and more recently, ethics of language assessment. He categorized the three elements of language assessment literacy respectively as "skills (which includes item-writing), knowledge, and principles". However, as Taylor (2008) notes, Davies was focused on the competencies required for the professional language testing community. She argued for the need to improve LAL for other stakeholders who may not be as familiar with the specialized knowledge of assessment, such as language teachers.

An article by Brindley (2001) represents one of the earliest treatments of AL in the context of language testing, and recognition of the role of language teachers in assessment. In response to the demand for stricter standards of reporting on the development of students and accountability for educational institutions, he expressed the need for language teachers to develop the competencies needed to use and participate in conversations of decisions related to assessment. Rather than focusing on the (lack of) measurement knowledge of teachers, Brindley acknowledged that teachers were aware of assessment issues and capable of evaluating the quality of assessments due to their experiences and *practices* in the classroom. Echoing Mertler (2003), Brindley argued for assessment training that addresses the needs of the classroom, moving away from an emphasis on statistical techniques for use in large-scale testing and more recognition of the classroom practices of teachers. To that end, he created an outline of a knowledge base that presented core and optional competencies for language assessment training, establishing a foundation for future work in defining LAL and recognizing the importance of classroom assessment practices in the knowledge base of LAL.

Inbar-Lourie (2008) expanded on Brindley's (2001) acknowledgment of the social context of classroom assessment and highlighted the social consequences of language assessment practices. Referencing the "social turn" (McNamara & Roever, 2006) in language testing, she drew attention to the critical issues of assessment and effects they have on educational and social contexts within and outside the classroom (Lynch, 2001; Shohamy, 1998; Shohamy, 2017). Her critical perspective drew attention to reflexive approaches in consideration of the power relations among the stakeholders of language testing, bringing light to issues such as culture, identity¸ ideology (Pennycook, 1999). Thus, there is a need to understand how and *why* item writers write in the ways they do, as their writing practices have social consequences for test-takers. For example, the ways in which item writers write may be related to how they construct the identities of test takers, the audience of the test. These practices are further complicated in collaborative settings where it is necessary for item writers to negotiate the appropriate writing practices.

Additionally, Inbar-Lourie (2008) noted “the need to democratize assessment” (p. 390) and a critical view of the responsibility of stakeholders in language testing. Previous views of AL development (Stiggins, 1991) described a top-down approach, in which the competencies of measurement would trickle down from “assessment literate” professionals to the teachers. In contrast, Inbar-Lourie’s view implies that LAL should be developed from the bottom-up; the practice and knowledge of assessment would lie more equally among teachers, administrators, and other stakeholders. Thus, the theoretical knowledge of assessment alone is insufficient to develop LAL, and the role of the language assessment *practices* of language teachers and item writers need to be considered in research on LAL development.

After the importance of LAL for different groups of stakeholders had been established, it was recognized that the components of LAL should be different for those groups. Taylor (2013) conceptualizes the components of LAL for different stakeholders as stages in development.

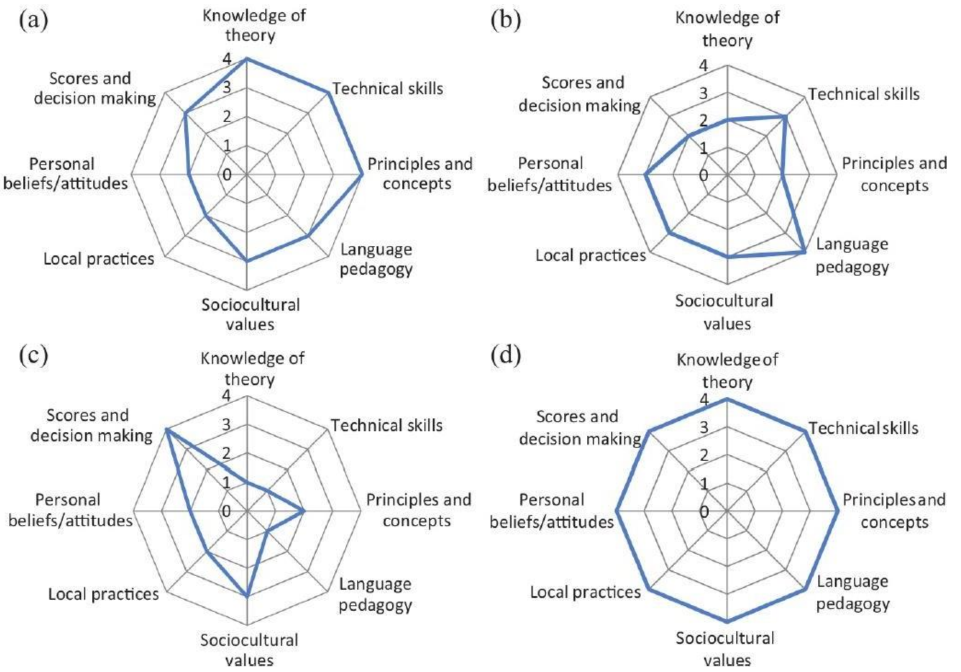


Figure 2.1. *LAL profile for different stakeholders. a) test writers. b) classroom teachers. c) university administrators. d) professional language testers*

Building upon this model of LAL, Baker (2016), as well as Baker and Riches (2018) in a later paper, make a distinction between components of declarative knowledge ("what") and procedural knowledge ("how to"). Another way of conceptualizing this is the difference between the theoretical components of assessment and practical components (e.g. decision-making). To reflect these distinctions, Baker and Riches present a modified depiction of LAL. They also add a new component of *collaboration* to emphasize the professional competencies involved in language assessment activities. For example, in the context of item writing, it can be expected that participants will need to develop the professional competencies of providing critique and implementing received feedback to write their items. Table 1.1 presents a comparison between the components of LAL identified by Taylor (2013) and Baker and Riches (2018)

Table 2.1. *Comparison of two models of LAL with definitions*

Taylor (2013)	Baker and Riches (2018)	Description of Baker and Riches (2018) components
Knowledge of Theory	Theoretical and Conceptual Knowledge (D)	Combined *knowledge of theory* and *principles and concepts* from Taylor (2013) Includes theoretical knowledge and principles of language assessment, including language pedagogy
Principles and Concepts	(Combined with *knowledge of theory)*	
Technical Skills	Task Performance (P)	"Broader term referring to all procedural knowledge related to the design, administration, and validation of language assessments" (p. 574)
Language Pedagogy	Language Pedagogy (P)	Enactment of language pedagogy, as opposed to knowledge about pedagogical principles
Local Practices	Awareness of Local Practices (D)	Declarative knowledge of local context/practices of assessment. (e.g. suitability of assessment materials)
Personal Beliefs/Attitudes	Awareness of Personal Beliefs/Attitudes (D)	Beliefs/attitudes regarding teaching and assessment (e.g. beliefs about the purpose of assessment)
Scores and Decision Making	Decision Making (P)	Procedural processes involved with scoring and decision-making, such as how test scores should be used.
	Collaboration (P)	Collaborative processes between stakeholders and participants of test development
Sociocultural Values	(Removed)	

Baker and Riches conceptualize sociocultural values as encompassing the other components of LAL; therefore, it is not considered as an individual part of LAL.

2.2 LAL Development

2.2.1 LAL Development Through Instruction

Following the expressed need for assessment courses in pre-service teacher education (Popham, 2004), many studies have investigated the gaps and needs of teachers in designing courses in assessment, primarily relying on surveys and interviews with teachers (Crusan, Plakans, & Gebril, 2016; DeLuca & Klinger, 2010; Fulcher, 2012; Volante & Fazio, 2007). Regarding teacher development of LAL, studies have discussed the effectiveness of instruction, which are typically in the form of language assessment courses that have been established in recent years as a requirement in graduate programs in TESOL. Lukin, Bandalos, Eckhout, and Mickelson (2004) reported the early successes of district-wide assessment training programs, emphasizing the need to design flexible programs that can be applied to various educational contexts. However, early language assessment courses, and their instructors, (Baily & Brown, 1996) were concerned mainly with traditional concepts of measurements such as validity and reliability, rather than the test writing processes of teachers.

Although a later survey (Brown & Bailey, 2006) revealed the continued concern of language assessment instructors in teaching theoretical concepts, some language assessment instructors began to develop a social perspective of assessment (e.g. Kleinsasser, 2005) and the recognize the importance of practice. Kleinsasser, in his own language assessment course for pre-service language teachers, sought to empower teachers and encourage the dialogue and negotiation of assessment practices, rather than simply transfer theoretical and statistical knowledge. In line with this approach to language assessment education, Scarino (2013) underscored the importance of developing the interpretive framework of teachers in relation to LAL. Drawing on language teachers' discussions and written reflections of their assessment of students’ intercultural understandings, Scarino highlighted the importance of teacher self-reflection in revealing their preconceptions of assessment practices. For example, in one case, a teacher considered the conflict between being objective in her assessment of a student and the inherent subjectivity in evaluating intercultural competence. Studies such as Scarino’s and Kleinsasser’s present possible approaches to address the social dimensions of language testing, which involve close considerations of the practices of assessment. As recent studies from various educational contexts (Lam, 2015; Malone, 2013; Vogt & Tsagari, 2014) continue the important work in bridging the gap between measurement theory and practice in LAL, others also acknowledge and develop further research in the social concerns of assessment (e.g. Jin, 2010) and bring to light the need to consider the social practices of test development.

2.2.2 LAL Development Through Practice

In addition to formal instruction, early studies (e.g. Mertler, 1999) found that teachers acquire assessment skills through their classroom assessment practices, effectively “learning on the job”. In some cases, practical experience with assessment may be more crucial to the development of LAL than formal instruction. As Kleinsasser (2005) noted, instruction may be a hindrance to LAL development in terms of preparing teachers for practice in the classroom, as teachers may become stuck in the theoretical underpinnings of assessment and lose sight of practical concerns of their own assessment practices. In his assessment course for language education students, Kleinsasser drew upon Wenger’s (1998) communities of practice theory to argue that participation in developing assessment materials was a key component of his students’ professional and LAL development. Although research on the implementation of language assessment courses continues to present useful suggestions for teacher training, further investigations on the role of the assessment practices of teachers, which include test writing, may illuminate another important source of LAL development. Coniam (2009), in a study of language teachers in Hong Kong involved in a test development project, noted the poor quality of tests they wrote because they lacked theoretical measurement knowledge. However, he acknowledged that the teachers seemed to display a good awareness of test principles such as validity and reliability during the test development process, even if they lacked the technical knowledge to ensure such principles. Given that his participants were experienced teachers, this may suggest that teachers may develop implicit knowledge of appropriate testing practices from their experiences with classroom assessments. Even if they lack the explicit theoretical knowledge of measurement, they may accumulate a repertoire of good items and test-writing practices (Popham, 2001) that would allow them to succeed in their daily instructional tasks. Given our limited understanding of the relationship between implicit and explicit knowledge in test development, it is crucial to examine test writing practices and what test writing entails.

2.3 Item writing as Distinct Genre Practices

Although the language testing literature focusing on test writing is sparse, three case studies (Green & Hawkey, 2011; Kim, Chim, Huensch, Jun, Li, & Roullion, 2010; Ryan & Brunfaut, 2016) explored the social processes of test writing across various institutional contexts. They presented findings that lend support to conceptualizing writing as genre practices, and examining test items as texts. This allows for new approaches for examining what test writing entails from a genre studies perspective, which has been a crucial lens in writing studies research of many domains of professional writing (e.g. Bhatia, 2008).

The following three studies have revealed item writing as a unique set of practices that share a commonality with other genres of writing, in that they are characterized by rules and conventions unique to their social context and purpose. Kim et al. (2010) conducted a case study to examine how individual test writers responded to the constraints and writing rules outlined by test specifications (Davidson and Lynch, 2002; Fulcher and Davidson, 2007) in a university test development setting. Their study argued the limitations of the use of test specifications as a static document and suggested the need to incorporate the test writing practices of individuals in revisions of the test specifications. However, notwithstanding their critique of test specifications, Kim et al. reaffirmed the necessity of rules and guidelines for test writing, suggesting that there *are* genre conventions in test writing that need to be followed to successfully write test prompts. Thus, test writing can be conceptualized as specialized literacy and genre practices, similar to academic writing that has been recognized as a set of crucial literacy practices in university settings. Ryan and Brunfaut (2016) examined a case of collaboration between language assessment literate test writers with no proficiency in the tested languages and “language informants” who speak those languages. The fact that such collaboration between participants of different expertise is required to create a test highlights the complex and specialized nature of test writing. Language tests are created from a combination of the theoretical knowledge of assessment and content knowledge of the language being tested. The theoretical knowledge of assessment serves as a guide to writing test items that appropriate for the context and purpose of assessment. It follows that there are particular genres of item writing that are required for the validity of tests. Green and Hawkey’s (2011) study further elucidate the distinct nature of language test writing in a professional test development setting. The researchers tracked the processes of trained writers as they adapted texts for the academic reading portion of the International English Language Testing System (IELTS), a widely used standardized English proficiency test. From their findings, they discussed the issue of test authenticity, noting that some test writers felt that the prompts that they had adapted and the writing that they asked test-takers to perform did not accurately reflect the kinds of academic writing that students would encounter in universities. Despite this disparity, the test writers defended the appropriateness of the prompts because they served their purposes of assessing their students, even with the lack of authentic "academic" content. The writers' recognition of test purposes led them to adopt particular genre practices in the way they presented the information in the test. Item writing is a complex process in which writers must take into account various factors from both knowledge of the assessed content and the theoretical knowledge of assessment to guide their specific writing processes. Furthermore, the previously mentioned studies of item writing occurred in close collaborative contexts, which illustrate the importance of examining the writing as a social activity.

2.4 Genre Theory: Acquiring the Practices of Item Writing

When item writers write tests, they not only produce the products of their inscriptional practices but are simultaneously engaged in producing genre practices and negotiating the purpose and conventions of test writing. One way to understand LAL development through item writing is to understand how its genre practices are negotiated and acquired by the writers of a test development team. Although genre is defined differently in a number of disciplines, this study will adopt an understanding of genre that draws from the school of Rhetorical Genre Studies (RGS). A key tenet of RGS comes from Carolyn Miller's seminal piece, *Genre as a Social Action* (1984). Miller argued that genres are constructed from social and rhetorical action, rather than the form, or "text types" (e.g. writing style). Within this view, a collection of discourses that are similar in content and form may still fail to constitute a genre unless they share a pragmatic function and purpose; although, similar styles of writing may be a result of writing for that shared purpose. Miller's concept of genre drew upon the concept of *typification* that was developed in Alfred Schutz's (Schutz & Luckmann, 1973) work in sociology, a notion that has also been adopted in later works in genre studies (e.g. Bazerman, 1988). Schutz argued that our knowledge and understanding of the world is based on types, or categorizations that allow us to generalize and recognize similarities between new experiences and previous ones. Therefore, genres become *typified t*hrough recurrences of social action and interaction between participants. For example, as individuals come to understand the meaning and consequences of writing for a particular purpose and social context, they gradually learn and *define* the practices of genre that will allow them to perform those social actions. Situating this in a item writing context, it can be expected that item writers gradually come to a consensus on what the appropriate practices of item writing are as a group by recognizing the common practices of the particular item writing context.

Individual item writers have the potential to shape social test writing processes through their practices as other writers come to recognize them as test writing genre practices and adopt them as their own. A useful notion to explore how genre practices are recognized and adopted by others is *intertextuality*, or the interactions and relationships between texts. For Bazerman (2003), intertextuality refers to "the explicit and implicit relations that a text or utterance has to prior, contemporary, and potential future texts" (pg. 86). In a study of tax accountants' writing practices, Devitt (1991) introduced the notion of the *genre set*, which refers to the routine set of genre practices enacted by tax accountants. Moreover, she underscored the stabilizing feature of genre sets; existing genre practices become more recognizable through their recurring use and by intertextual association with other practices in that genre. Texts that are situated in the same social context and purpose may be recognizable as such because they draw on and make connections to aspects of other previous texts. For example, the writing conventions of a test writing group may become stabilized as genre practices as more texts adopt those conventions. In this manner, the formerly idiosyncratic practices of an individual writer have the potential to become accepted as routine practices in a particular social situation. Bazerman (1994) expanded on the notion of *genre sets* by introducing the concept of a *genre system*, referring to a system of interrelated genres, rather than simply interrelated texts. Social actions become subsumed within a shared activity; multiple genres, rather than a single set of genre practices, may interact with each other in particular settings. In the case of test writing, writers may draw from various genres practice such as academic discourses to participate in the activity of test development. Using the notions of interrelated genres and intertextuality allows for considering the ways that writers interact with texts (test items) and other members of a test development team. These types of interactions eventually develop into the valued practices of a group of item writers, and thus it is important to examine which practices and values of writing are acquired.

Although the coordinators of a test development team may provide feedback and instruction to item writers, it is likely that the practices of item writing are established more from practice than explicit instruction. Studies of the experiences of vocational writers within the workplace (Beaufort, 1999; Uhrig, 2012; Winsor, 1996) demonstrated instances where genres were acquired through practice. In these cases, writers seemed to acquire new genres without explicit instruction from more experienced members of a professional group, and “learned by doing” instead. In another case, Parkinson, Demecheleer, and Mackay (2017) described trainee carpenters and their process of learning to write diary logs for the workplace while receiving instruction from more experienced carpenters. Although a characteristic of carpenter log writing is the use of passive voice, the instructors taught a narrative style of writing logs, which they perceived as a more accessible writing style to beginning carpenters. However, the trainee carpenters developed a style of using the passive voice through their practices of writing in the workplace and their professional and intertextual interactions on the job. Interestingly, although the experienced carpenters had the authority in terms of their knowledge of the field, they played a minimal role in shaping the practices of the trainee carpenters. Similarly, item writers may develop more from bottom-up interactions with a team than top-down prescriptive instruction from test development coordinators.

Test specifications and exemplar models of test prompts based on those specifications, also play an important role in defining and producing genre practices. Learners of new genre practices seem to draw on models and examples of writing in both instruction-based acquisition (Kelly-Laubscher, Muna & van der Merwe, 2017) and practice-based acquisition contexts (Tardy, 2005). Tardy (2006) also noted that learners of new genres tend to seek examples of exemplary writing in that genre when a model is not provided for them. Purcell-Gates, Duke, and Martineau (2007) explored the literacy education of elementary school students by comparing the explicitness of teacher instruction with authentic reading and writing activities. The findings showed no significant effect in the degree of explicitness of teacher instruction on student reading and writing growth, but did suggest that the authentic reading and writing activities had a positive effect on learning. This study further supports the importance of exemplary writing models and examples of genre-specific writing for novice writers. It seems that models can be valuable resources by allowing writers to draw on texts that present already established genre practices. However, it is unclear which aspects of models learners draw from, such as organization of the writing or sentence-level features, and to what extent they follow the model.

Exemplar models and peer writing samples are crucial to the role of genre acquisition due to the nature of intertextuality. Through oral and textual interactions, individuals come to recognize and adopt the genre practices of test writing and may also transmit their practices to others. In other words, they are *socialized* into these genre practices through their interactional discourses.

2.5 Discourse Socialization

Discourse socialization is an approach to theorize how genre practices can be acquired through social practices. Understanding both how and why test writers adopt certain writing practices will provide a new perspective to complement the work that has been done to develop LAL through instruction. The concept of discourse socialization was derived from the theoretical perspective of *language socialization*, which refers to the processes in which individuals are socialized to *use* language, *through* language (Schiefflin & Ochs, 1986). Typically, the novices of a community learn to participate in the discourses of that community through their social interactions with more experienced members who have competency in those discourses. Through these interactions, individuals come to develop social competence within a particular social context. In the case of language/discourse socialization, the language/discourse is both the target of and medium through which competencies are developed.

This concept has been adopted as a theoretical perspective in fields such as L2 education (Morita, 2000; Zuengler & Cole, 2005) and applied linguistics (Duff, 2010). Discourse socialization has been used in studies of academic discourse (Kobayashi, Zappa-Hollman, & Duff, 2017; Zappa-Hollman, 2007) with an interest in the interactional processes by which newcomers (such as L2 English speakers) acquire the discourses and competencies required to engage in the cultures of academic institutions. Similarly, one could frame the development test writing as a discursive socialization process. Discourses of assessment can be formalized (Davies et. al., 1999), such as with the theoretical knowledge and technical language associated with assessment. They can also be the informal discourses that item writers engage in when discussing how to create items with their peers or other stakeholders.

With socialization processes, a common assumption is that newcomers will acquire competencies by interacting with more competent individuals in the social context, the more experienced “experts”. However, socialization processes may also result in the learning of the more competent members of the social group. As Talmy (2008) and other scholars have emphasized, socialization is multidirectional. Novice writers are not simply passive recipients of knowledge and practices, they also play active roles in socializing other novices and experts of a community. The approach of discourse socialization is compatible with Lave and Wenger’s (1991) notion of the community-of-practice, which frames participation in these social activities as a means of learning. Thus, for newcomers to a social activity, learning how to *participate* discursively in that activity should facilitate their development of the competencies and practices related to that activity. Discursive socialization processes include both oral (e.g., Seloni, 2012), and textual (Okuda & Anderson, 2018) interactions. The concepts of discourse socialization from linguistic anthropology and intertextuality from genre studies are useful in examining test items as *texts* that are composed of recognizable genre practices. These practices could potentially be developed in the community-of-practice of item writing, as the writers learn how to participate in collaborative writing within an idiosyncratic context.

2.6 Research Questions

The context of an institutional test development project at a U.S. university provides an opportunity to investigate a site where individuals participate in a collaborative item writing process, from which different discourses and practices emerge and contribute to the development of item writers and pre-service language teachers. The nature of item writing is still underexplored, and this work seeks to address this by reconceptualizing item writing as genre practices that are developed through a community of practice (CoP) of novice item writers.

Research Questions:

a. What are the key features of item-writing for an integrated writing test, and how are these features related to LAL?

b. How do novice item writers/pre-service language teachers develop (LAL) through the process of collaborative writing for an institutional test?

Chapter 3: Methodology

3.1 Context and Process of Test Development

Each year, the English Placement Test (EPT) developers at UIUC recruit several graduate students from the department of Linguistics/TESL for paid RA positions to adapt source texts and write prompts for the EPT. International students from countries where English is not a first language are required to take this exam if they are admitted to the university and have TOEFL scores that do not meet institutional standards. This exam is administered before the beginning of the Fall and Spring semesters and places the students into academic writing classes according to the scores they have received. The process of writing and revision takes place over the course of a semester, and the item writers are overseen by two EPT coordinators who manage the development and administration of the exam. Each test writer is required to adapt and write two test prompts based on texts from online sources, and each prompt consists of articles, powerpoint slides, and a lecture that is presented to test takers.

The recruitment of the item writers was completed before the end of the Fall semester, and the test development team had their initial meeting in January 2018 at the beginning of the Spring semester. This meeting allowed item writers and coordinators to become familiar with the other members of the team. The two coordinators also explained the general test development process and responsibilities to the item writers. During the meeting, the coordinators provided the item writers access to a website from a previous year that the test takers used to take the test in its entirety. The writers were asked to return home and take the test in the same manner as the test-takers, beginning with a video that explains the procedures for the EPT. The writers then listened to a video lecture accompanied by powerpoint slides on the topic of the prompt and proceeded to read the article on the same topic before writing their essay response.

After the initial introduction and briefing of responsibilities, the test development team, consisting of two EPT coordinators and five writers, met once a week for one hour. The first few meetings consisted of brainstorming sessions, in which each writer would discuss potential topics they wished to create prompts for, as well as provide source texts of information about their topic. The team would discuss and give feedback on the viability of each individuals' proposed topics, but the bulk of the feedback at this stage came primarily from the test coordinators. The test coordinators had previous experience with similar prompts that were written in previous years and could provide insights into what would be appropriate and effective topics for testing purposes. After the first prompt for each test writer had been decided, the coordinators provided test specifications, a document that provided general guidelines for writing the prompts. They also provided a model of a completed prompt from a previous year as a reference for the test writers. The model included an article with six sections. Three sections consisted of evidence in favor of one side, while the other three consisted of evidence against that side of the issue. The test writers were required to label each section either as "pro" or "con". Within the articles, there were also visual aids such as charts and infographics. In addition to the articles, the writers were also required to write a script for a lecture that would be presented in audio format to the test takers. Finally, the model included a powerpoint presentation that consisted of concise text and visuals. This powerpoint would complement the lecture in the form of a video, and transitions between slides would be synced with the audio during the test.

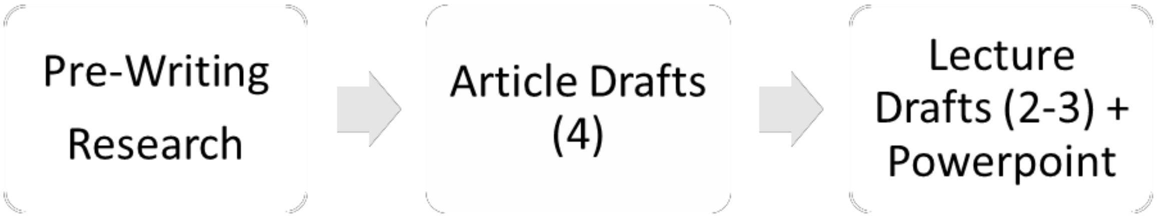


Figure 3.1. *Writing process and number of drafts for each section of the prompts. Peer feedback occurred at each stage of the process.*

The test development team followed a pre-determined schedule and assigned the item writers a task to complete each week before the team meeting. Each week, the writers were given a deadline to complete their draft and submit to a shared document a few days before the meeting to allow time for other members to read the other drafts. Before the meetings, each member was required to give comments on the drafts of the other team members. A typical meeting would begin with announcements by the coordinators and a review of the progress of the team as a whole. The team would then discuss the draft of each writer in turn, either elaborating on the written comments they had given the day before or giving some general comments and advice to the item writer. After the first prompt was completed, the process of writing the second prompt proceeded in a similar manner. For both the first and second prompts, the writers created the articles first, and the lecture and accompanying powerpoint second. Each of the prompts took a little more than a month to finish, with approximately four to five group meetings.

3.2 Approach

This study adopted a largely qualitative approach to answer the research questions. Data in this study consisted of group discussions, individual interviews, and artifacts from the item writing process, which include multiple drafts and comments. Due to the exploratory nature of this study, a thematic analysis (Braun & Clarke, 2006) was employed to interpret the data from group discussions and the experiences of the individual writers as expressed in their interviews.

This qualitative approach allows for consideration of the experiences of individuals while attempting to make key generalizations about the nature of item writing in this context. The qualitative data addressing the development of LAL were corroborated with the frequency of types of comments in the writer's drafts.

3.3 Participants

Participants included the two EPT coordinators, doctoral candidates in the department of Linguistics, as well as five item writers. One writer was a doctoral candidate of Linguistics, and four were first year graduate students in the MATESL program, also affiliated with the Linguistics department. At the time that the participants were item writers, they were also teaching ESL courses as TAs, primarily in the ESL service courses offered by the MATESL program for international students. Additionally, all item writers had taken the course EIL 460 (Principles of Language Testing) in the previous semester. Therefore, the item writers had a certain degree of familiarity with test development, as well as familiarity with other writers due to taking the course together.

Table 3.1. *Item Writer Profiles*

Initials (Pseudonym)	Gender	Major and Program	Country of Origin	Role
S	Female	Linguistics (PhD)	U.S.	Test writer
E	Male	TESL (MA)	U.S.	Test writer
K	Male	TESL (MA)	Kenya	Test writer
M	Female	TESL (MA)	U.S.	Test writer
Cr	Female	TESL (MA)	U.S.	Test writer
C1	Female	Linguistics (PhD)	Korea	Test coordinator
C2	Female	Linguistics (PhD)	Korea	Test coordinator

3.4 Data Collection

Multiple types of data were collected throughout the semester to explore the item writing practices and draft revisions of the five item writers. The choice was made to exclude artifacts such as test specifications and sample prompts from the data because they were rarely mentioned in the group discussions and interviews.

Table 3.2. *Collected Data and Descriptions*

Data	Description
Drafts Comments	The weekly drafts of the first and second prompts were collected by the researcher. Drafts included the article section and lecture section of the prompts, along with comments from the coordinators and other test writers. Although the powerpoint section was also part of the test, they are excluded from the analysis due to the lack of comments. The purpose of collecting the drafts with comments was to track the development (socialization) of writers into particular genre practices of item writing.
Group discussions	The recording of the group discussions began at the beginning of the writing process for the second prompts. The purpose of recording these discussions was to explore and illuminate the features of item writing.
Individual Interviews	Interviews were conducted with the test writers. The purpose of these interviews was to elicit the writers' thoughts about what item writing entails, and perspectives of their development as item writers.

The drafts of the prompts were originally created by the writers as shared google docs to facilitate written feedback. Comments for each draft were given by both the coordinators and writers using the comment feature. The drafts and their corresponding comments were downloaded by the researcher as Word documents for ease of analysis. Each writer created two prompts, and each prompt consisted of four drafts for the article section, and two to three drafts for the lecture section (some writers required one more draft to finalize prompts).

The study received approval from the IRB at UIUC weeks after the test development process had begun. Therefore, data from group discussions include only those from the beginning of the second prompt, totaling six hours of recorded discussions. To supplement the lack of discussion data for the first prompt, data from the drafts and interviews were used to provide a better picture of the test development process from beginning to finish. Data of the group discussions were transcribed by the researcher. For the purposes of a thematic analysis, only a rough transcription was necessary. Therefore, the transcribed data for analysis do not contain detailed nuances such as length of pauses or in-breaths, but focus on capturing *what* was said rather than *how* they were said.

Three weeks after the test development process had ended, individual interviews were scheduled with four item writers (excluding the researcher). All interviews were conducted face-to-face by the researcher, and lasted on average for 30 minutes. The interviews were semi-structured, with a set of questions that targeted general aspects of the writing process with the intention of eliciting individual experiences regarding the writing process (see Appendix A). The interview guide was developed by the researcher after an initial analysis of the group discussion and draft data in order to prepare questions that would address the salient features of item writing for the writers. There were three main sections for the interview, which began with questions regarding the writers' feelings and attitudes towards writing prompts and how their strategies and processes of writing developed over time. The second section targeted the group dynamics and the role of collaboration in the item writing process. The final section targeted the development of the item writers' individual writing practices, and changes across the process of test development. All interviews were audio-recorded and later transcribed by the researcher.

3.5 Data Analysis

To answer the first research question, data from the group discussion and interview were analyzed with thematic analysis, and the themes generated from each data source were triangulated. As the first step in the analysis, a thematic analysis of the group discussion was conducted by the researcher following the six-step process described by Braun and Clarke (2006). This early stage in the process involved reading carefully through the transcribed data several times to gain familiarity with the data. Following this, initial codes were generated following a data-driven inductive approach (Boyatzis, 1998). An inductive coding approach was chosen for this stage in the coding process due to the exploratory nature of the study. The thematic analysis progressed into categorizing the codes into themes, which were then reviewed and checked to ensure themes did not overlap with other eliminations. At this stage in this process, themes that were similar or lacked key distinctions were merged to reduce the number of themes. The process of searching for themes was iterative, and the researcher compared newly generated themes with the coded extracts several times to establish an initial list of themes. A similar inductive method of coding was also conducted for the individual interviews, and a set of themes was generated separately from the group discussions. Themes were generated selectively, focusing on codes that were related specifically to the writers' descriptions of the features of item writing. A particular choice had to be made to generate themes for the interviews, which had relatively fewer codes than group interviews: if at least three writers (constituting a majority of writers) made similar comments on an interview, that was considered enough to constitute a theme. The researcher then compared the themes between the interviews and group discussions and combined themes that were similar, once again reducing the number of themes. The choice was made to discard themes that consisted of a relatively small number of codes if those themes could not be combined with others.

To answer the first research question and explore which components of LAL item writers draw upon during test development, categories were generated from the group discussions and interviews. The group discussions provided insights into the aspects of item writing that need to be addressed and negotiated during the collaborative processes of feedback and revision of individual test prompts. Additionally, semi-structured interviews were conducted with the test-writers, with some questions targeting what the item writers considered to be important aspects of item writing and the test writing processes. Codes and categories related to features of the item writing were generated from the interviews, and the categories that were similar to those of the group discussions were consolidated to form four themes related to the features of item writing. These four themes, which include developing knowledge of topic, consideration of audience, presentation of information, and writing style, reflect the collective values of the test development team that emerged from the practice of item writing and can be linked to the components of LAL. To answer the second research question, the qualitative data from the interviews were triangulated with frequency data from draft and comments. During the initial coding process of the interviews, codes related to the writers' reflection on their development as writers were isolated from the other generated codes. Themes related to item writer development were then generated from these selected codes. For the drafts, all comments were categorized by type (what to revise) and function (e.g. suggestion, praise). After the comments were all assigned types and functions, the comments with functions other than "suggestions" were excluded from the analysis. The frequencies of the types of comments were then counted for each writer, for each draft of the prompts. Drafts of articles and drafts of lectures were counted separately. It is assumed that writers that receive less suggestion-type comments in a particular feature would be relatively more competent in that area of writing. Decreasing frequencies of a type of comment across drafts could then suggest improvement/development in that feature of item writing for a writer.

Table 3.3. *Themes and codes with frequencies*

Themes	Categories	Sub-Categories	# of Categories (Discussions)	# of Categories (Interviews)
1. Developing knowledge of topic	a. Conducting research on topic	i. finding accurate topic information ii. finding sources	11	6
	b. Personal experiences and previous knowledge of topic	i. personal experiences of topic ii. previous knowledge of topic	15	2
2. Consideration of audience	a. Cultural Appropriateness	i. making prompts relatable and interesting to international students ii. cultural sensitivity towards international students	31	4
	b. Test-takers' background knowledge	i. using more common vocabulary ii. considering test-takers' familiarity with topic	21	7
3. Presentation of information	a. Examples and Explanations	i. balancing explicitness given information in prompts ii. considering amount of detail needed for prompts iii. providing sufficient examples and explanations iv. creating arguable prompts	42	8
	b. Being Objective	i. avoiding biased opinions in prompts ii. creating arguable prompts	22	3
4. Writing Style	a. Readability	i. organizing structure of writing ii. being concise with ideas and language	38	8
	b. Consistency	i. using similar writing styles as other writers	12	7

Chapter 4: Results

4.1 Knowledge of Prompt Topic

Finding information regarding the topic of the prompts was an essential task in the pre-writing stages of the test development. Before starting to write each of the prompts, the writers were asked by the coordinators to do research to find sources and background information about the topic, typically online articles or journals. As the writers began to write their drafts, finding sources of information continued to be an important task during the process, as the team members helped each other find more sources, or provided information from their own background knowledge and personal experiences. Figure 4.1 depicts the iterative processes of feedback, research, and revisions. Typically, feedback comments regarding *knowledge of prompt topic*, and occasionally *presentation of information,* required more effort to revise because the writers frequently needed to conduct more research to address these issues in the prompts.

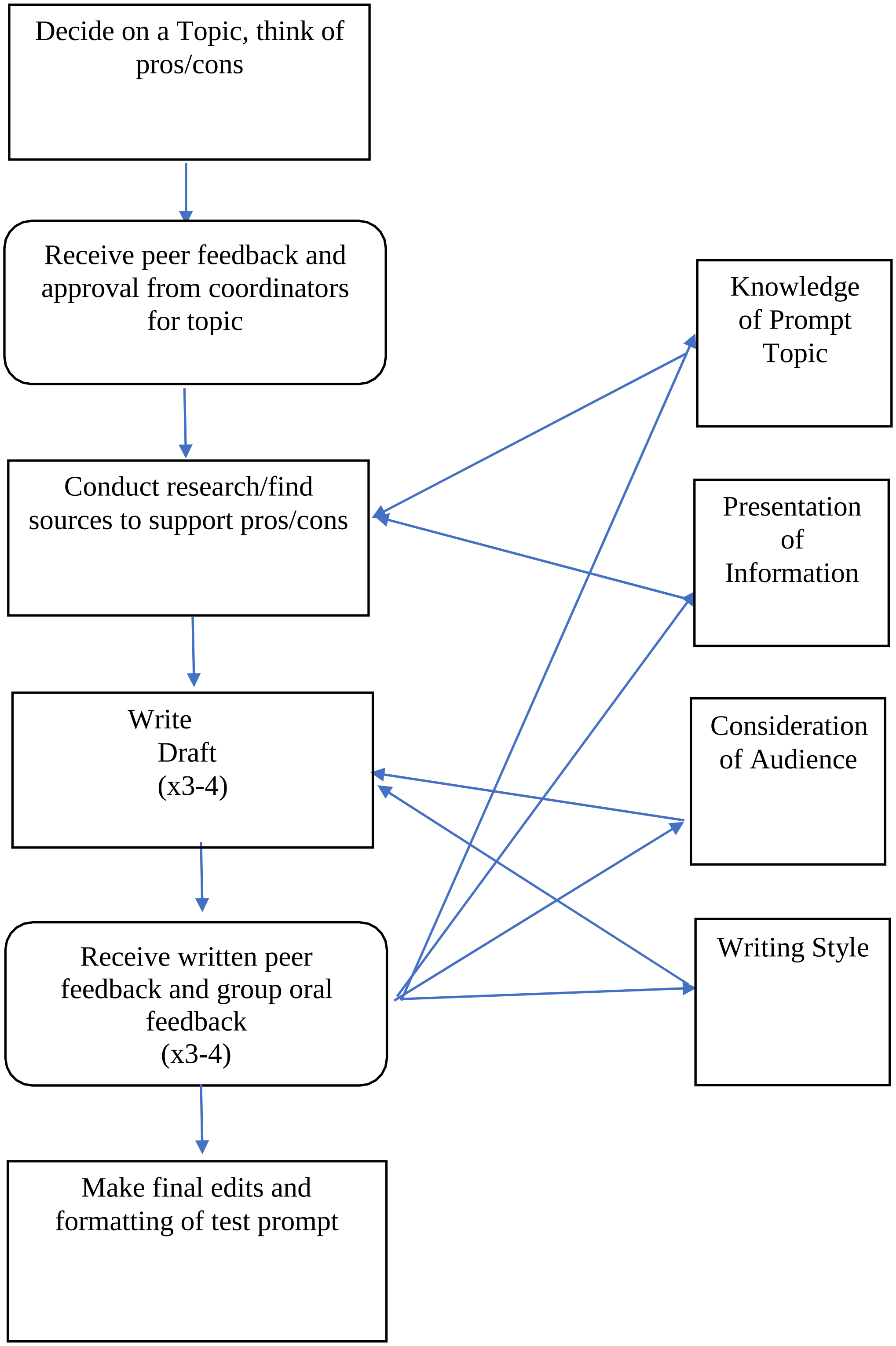

Decide on a Topic, think of pros/cons
Receive peer feedback and approval from coordinators for topic
Conduct research/find sources to support pros/cons
Write Draft (x3-4)
Receive written peer feedback and group oral feedback (x3-4)
Make final edits and formatting of test prompt
Knowledge of Prompt Topic
Presentation of Information
Consideration of Audience
Writing Style

Figure 4.1. *Item writing and feedback process of prompt articles*

4.1.1 Conducting Research

Although some of the topics were familiar to the writers already, they were encouraged to spend a substantial amount of time finding evidence from sources that could support both sides of an argument for each topic. Considering that one purpose of the EPT is to assess test-taker's argumentative skills, the members of the test development team expressed the importance of developing pro/con points that would elicit argumentative writing. Some writers expressed their difficulties in finding and using sources in creating their points for the prompt.

> S: "Well, I know I was missing one pro, I guess that was it. And, I don't know, I felt like...I don't know I didn't know if I was evidencing things well enough. Umm, or... if these were good points, I guess. I was not entirely sure." (4/3 discussion, line 6)

The importance of finding enough evidence and sources to support both sides of the arguments were also emphasized by the coordinators.

> Cr: "I think I really feel like my cons are pretty weak compared to my pros? I don't know how everybody feels, but I don't know how to make it better. Obviously, I need like more, source material but,"
>
> C2: "Overall, I thought it was good topic and the points are good, I mean, I could see that con 1 maybe needs a bit more evidence?" (4/10 discussion, line 15)

In these cases, the test development team recognized that having more evidence would make the points stronger, a concept that is common in academic writing, and also translates to writing prompts for assessment purposes. The writers described the importance of gathering enough factual information from sources, which could later be simplified or condensed as needed for the purposes of the test.

> Cr: "I feel like it's that, it's [EPT] almost mimic writing, because we're trying to create articles, which take actual information and synthesize into something that yeah, it' s pretty factual." (Interview, line 139)

> Cr: “I feel like I've lied to the students. (laughs) No, I'm kidding. You know, because we give them sort of snippets of information. And I think as a teacher that almost goes against my own personal beliefs.” (Interview, line 19)
>
> K: “It is more like just finding general information and information would have to be authentic?” (Interview, line 35)

In addition to finding sources with sufficient information, the writers also placed importance on finding unbiased sources. Sources that presented information objectively were favored over more opinionated ones.

> M: “Everytime I searched and then I started searching the library and start like looking at scholarly articles and it was just, there's too much and people have too strong opinions…” (4/3 discussion, line 140)
>
> “I was trying to figure out the best tactic to do that and explain it and try to figure out how to explain some of these things without making it too controversial exactly.” (line 146)

Finding quality sources that were both reliable and detailed with information were seen as helpful by the writers. Occasionally, the coordinators would encourage the writers to quote directly from the sources, especially if the information was good, rather than paraphrase the concepts from the sources.

> “I mean it seems like you've read a lot and you have a lot of information so maybe it's better to just cite instead saying things? Like, this person says, these are traditionally” (4/3 discussion, line 276)

4.1.2 Personal Experiences and Knowledge

Although most of the information in the prompts came from the research and sources about the topics, personal experiences and previous background knowledge of the writers played a prominent role in developing the content of the prompts. Individual writers had different perspectives to share based on their understandings of the topic, which allowed them to make suggestions and help each other develop better understandings of the topics. One of the most common displays of background knowledge was knowledge of the cultural context of the U.S. and the relationship it has with the topic at hand.

> Cr: “Yeah I think what I was going for was that, because the U.S. doesn't-like most kids don't start taking a foreign language until later when their already busy…” (4/17, line 254)

Often, the writers would discuss whether the findings from their sources were consistent with their own personal experiences or knowledge of the topic. The writers’ individual knowledge were typically used to support or challenge the accuracy of topic information. The members of the test development team tended to respond favorably when they perceived a connection between the topic of the prompts and the real-world, especially when they had previous knowledge or experience with the topic.

> C1: “I kinda liked it because I kinda understood why you can with this con 1 because I can see exactly, like the courses they teach at those schools like a really gender stereotyping like, like girls high schools in Korea for instance, they have like dense courses you know, so like I was pretty amazed.” (4/17 discussion, line 122)

In some cases, the writers’ personal experiences influenced what they felt to be important inclusions to the prompts. Those that had experiences related to the topic tended to be more sensitive to the types of examples that could be used to help make the concepts in the prompts clearer to other readers.

> K: Uh, this university has so many libraries, and they also have a storage facility off-campus.
>
> M: Right, ok. Then I think that needs, then you need to have that anecdote, to make it real. (4/10 discussion, line 281)

Although personal experiences were good resources of information and examples for the prompts, issues arose when these experiences became sources of biased opinions.

> M: Yeah, I think you-maybe, ok, maybe I'm wrong about this, but I think you, you're-like maybe writing this through your own life experiences a little bit? (4/3 discussion, line 499)

One role of the group discussions was to provide a forum where the writers could share their knowledge and experiences. Especially considering that the test-writers typically chose topics that they already had previous knowledge of or had a personal interest in, sharing their thoughts with the group allowed them to develop understandings of the information that they gathered in a more objective manner.

> M: I tried to choose something for my first topic that I knew something about, that I had, that I felt like I could find good resources on, and that I felt like I could present both sides of, and that would be interesting to me? (Interview, line 50)

> Cr: "…I don't think anybody could say that none of them influenced them, cause I feel like everybody chose topics they were either interested in or had a personal vendetta against." (Interview, line 159)

4.2 Consideration of Audience

The item writers considered the characteristics of test-takers as international students coming to study in a U.S. context. Discussions in this category focused on two main characteristics of the test-takers: culture and general background knowledge. The members of their team shared their thoughts on the types of topics and information that would be appropriate to test-takers while considering sensitivity to their cultural backgrounds and also their familiarity (or lack of) with U.S. cultural contexts. Another point of concern was addressing the diverse general background knowledge of the students. The item writers discussed what types of information would be common knowledge to the target audience of incoming undergraduate and graduate students.

4.2.1 Cultural Appropriateness

The test development team constructed the characteristics of the test-takers by drawing upon their own knowledge of various cultures, as well as previous interactions with international students within their classrooms and at the university. One element of making the topics suitable for the audience was developing ideas that would be relevant and familiar for them. A common point was that students from a particular country would have difficulty understanding concepts that were common knowledge in U.S. contexts only.

> K: I was just trying to make my passages simple enough for a diverse group, use information that is kinda general to any culture that might be taking the exams. (Interview, line 27)

It was preferable that the examples used in the prompts would be familiar to all test-takers regardless of their country of origin. In one example, the team was able to negotiate a common ground for the test-takers by taking the perspectives of internationals coming to study at the university.

> C2: I don't know if it helps, but I felt like all the students are coming here for the traditional universities?
>
> M: Yeah.
>
> C2: So I thought there should be a way to turn it more relevant to them?
> (3/27, lines 47-49)

Consideration of the international students was important not only to create prompts that would be relatable to them, but also in the manner that their identities and attitudes would affect the way they developed their arguments.

> C1: So I've seen a lot of people who are native in the U.S., like who cannot speak a, other language because they don't feel the need to do it. But the test takers are mostly international students and they come at least with one, you know, language knowledge, so-
>
> Cr: So, do you, are you saying that you think they would end up all leaning towards like one direction? (4/3, line 429)

However, consideration of the international students' cultural backgrounds and knowledge did not necessitate avoidance of the topics altogether. In cases where parts of an unfamiliar context would be presented to the test-takers, the team discussed including explanations to help their understanding.

> C2: That's a good idea, I think for your lecture, I think it's good to like kinda paint a picture, what it's like to be a typical American student here and what does typical American student lo-be in terms of his exposure to foreign language education? (4/17, line 260)

These kinds of explanations served to help test-takers succeed in the exam, and also, as earlier mentioned, served a pedagogical function for the test writers in educating the test-takers about aspects of U.S. culture. However, although the test-writers generally had shared knowledge about the U.S. context, it is important to note that there were differences among the writers.

> Cr: Have we all heard something like that?
> E: No
> K: No
> Cr: Really?!
> E: Cause credits works differently at every school. Like at my school [in the U.S.] one credit is equal to four credits here. So it's like a different system.

4.2.2 Test-taker's Background Knowledge

Another aspect of topic suitability was considering the background knowledge of test-takers, separate from their cultural backgrounds. In some cases, the information in the prompts would contain technical knowledge that most students would be unfamiliar with. As some of the test-writers were already, or became "experts" during the process of the research, it was important to consider information that would not be common knowledge to a general audience.

> C2: I think just, yeah, basically our audience is not well versed in linguistics for anything like that so-and they don't have to be, so we just want to make sure that it's um, geared toward general audience. (4/10 discussion, line 488)

Although there would be occasions where the test-takers had shared knowledge of a specialized topic, such as linguistics, generally the writers had minimal knowledge of topics outside of their own prompts. The writers reported in the interviews that they found the different perspectives helpful in determining what concepts would be unfamiliar, or what terms would require explanations for general audiences.

> K: Tablets or ebooks. Ebooks, when I was saying that ebooks are, well I didn't want to define ebooks because I thought that was really simple to everybody, yeah. (Interview, line 123)

> S: It was interesting to me how easy it is to miss something when you're writing something like that [multiple perspectives] so, that kind of helped. (Interview, line 11)

4.3 Presentation of Information

Developing the knowledge of the topic and considering the audience of test-takers leads to the next point of consideration: how to present the topic information. Before starting the drafts, the coordinators helped familiarize the writers with the format of the test, as well as the purpose of each section of the prompts. This allowed the writers to have a general idea of what would be assessed in this integrated writing task. Therefore, discussions in this area focused on what kind of information to include, the level of detail necessary for thetest-takers, as well as the types of examples and explanations to include in the prompts. Another issue was the manner in which the prompts should be presented. In order to ensure that the test-takers would produce responses based on their own interpretations of the issue, it was essential to present the information in a way that would not bias the test-takers towards one side of the issue.

4.3.1 Examples and Explanations

In some instances, the writers expressed their beliefs about the purposes of the assessment to justify the choices they made in the detail of the information they presented. For example, the writers would need to decide how much or little explanation to provide to the test-takers, as they expected that the test-takers would have a certain level of reading comprehension strategies to navigate the information in the prompts.

> C1: I tried to google how to simplify visual spatial skills, but like, it's like trying, the more I try to explain the further-
> Cr: eye space ((laughs))
> C1: It became more difficult? So yeah umm…
> C2: I mean hopefully they can figure it out…
> Cr: Yeah isn't part of this test is context clues you know, we don't want to give them simplified words for everything. (4/17 discussion, line 182)

The information provided in the prompts consisted of a combination of sources from the research process, as well as explanations and examples for the sources presented. Creating clear pro/con points in the prompts required strategic selection of the sources to present the possible positive and negative aspects of an issue.

> C2: It's really difficult, because I know there are, studies that are, opposite of this? yeah, so I'm like, if you get into too much, then you might ended up, end up sound like. you're really not giving a good pro? So, just pick and choose and just say that some research studies actually show...
> E: Cherry pick.
> M: Yeah. (4/10 discussion, line 37)

Some concepts and evidence from the sources required further explanation by the writers to shape them into pro/con points. Examples also helped to make these concepts clearer to the reader, especially when explaining more technical concepts.

> C1: Yeah, so I think that's the only concern so if you wanna make it longer, like maybe E's suggestion or C2's suggestion you can kinda change the expression to add more words, but at the same time give more graspable example? (4/17 discussion, line 224)

Discussing the purpose of the test also provided the writers with insights on the amount of detail to include in the prompts. The writers gradually began to grasp the amount of information the test-takers would need to accomplish the writing task, and recognized that too much detail could be distracting or overwhelming for readers.

> C1: I think we learned from our experiences of the topic 1, that as long as we have a clear set of ideas, and then stick to those simple ideas, do not go too deep? Then I think everything works. (4/27 discussion, line 94)

> Cr: do they need to know what features?
> M: well, personally I found it was really helpful, because she does use the phrase "features" in the article.
> S: that's what I was worried about, because I do feel it's kind like a weird use of the word? so I wanted to make sure it clear.
> C2: I actually liked it too because I think these could be- these you talk about right? in the article?
> S: yeah.
> C2: just like what M said
> Cr: so maybe just simplifying the information, cause I don't think they need to know all the features a computer grades vs. a person grades. (5/1 discussion, line 847)

The members of the team assumed that the test-takers would possess certain competencies that they would need to use to be successful in an integrated writing task, such as reading comprehension strategies to make inferences from the text. At times, the level of detail to include in the prompts needed to be negotiated to achieve of balance between being explicit with the information and assessing the test-takers' reading comprehension abilities.

> M: we're not trying to hide information from them. like, we're giving it to them, right? ((laughs)) and it's ok to say-
> Cr: I feel like last round we were trying to give it all to them, and this round we're like no, you get to figure it out. you need to earn your, whatever class you end up in. (5/1 discussion, line 581)

4.3.2 Being Objective

Early in the test development process, the writers were made aware of taking care to avoid presenting biased viewpoints of the topic. The writers reported personal interest and investment in the topic to be a source of bias in their writing, and found the discussions helpful to identify the biased aspects of their prompts.

> Cr: With Daylight Savings I almost had to learn about it before I talked about it, right? And then I had to resist the urge to put my own perspective, my own swing on it where it was very negative so, I think yeah, that definitely, the topic definitely influenced, and I don't think anybody could say that none of them influenced them, cause I feel like everybody chose topics they were either interested in or had a personal vendetta against. (Interview, line 158)

> M: Yeah so I like, I revamped some things based on our conversations.

> Cr: I really liked your, your new points. Um, I was gonna say it sounded a lot less feminist throughout.
>
> M: I tried. (4/10 discussions, lines 493-494)

Although it was clear to the writers from the beginning that it was important to avoid expressing personal opinions in the prompts, it was quite common for writers to implicitly project their values into prompts in the ways that they used stronger language to highlight certain points or the types of information they provided.

> E: What? I don't understand what I'm doing?
> C2: I don't know! like (laughs)
> C1: I think that hedging it down.
> M: Yeah, I think you-maybe, ok, maybe I'm wrong about this, but I think you, you're-like maybe writing this through your own life experiences a little bit?
> (4/3 discussions, lines 498-499)

4.4 Writing Style

Although the prompts covered a wide variety of topics, the common features that they would share in the final drafts included their organization and style of writing. Each writer had distinct writing styles in their sentence structures, word choice, and other elements of writing influenced by their previous experiences writing in other contexts. The EPT required a certain degree of uniformity among all the prompts, some which were explicitly mentioned by the coordinators at the beginning of the process, such as prompt length. However, other features of the EPT writing style were more implicitly recognized by the writers, and emerged through the processes of feedback and revision.

4.4.1 Readability

The test development team was concerned with the best writing practices to help guide the test-takers through the information they would need to develop their arguments. This was mainly achieved through distributing and organizing the information in a logical manner, with a set of paragraphs with pro points, followed by another set of paragraphs with con points. The writers made use of organizational cues, such as paragraph titles, to help the readers with previews of what the passages were about. There was also emphasis on the structure of the paragraphs and the content of topic sentences at the beginning of each paragraph to clearly state each pro and con point for the reader. Regardless of the strategies used, the purpose of the writing was to present information in an easily processed manner.

> C2: uhhh. I mean. I personally think that you need a topic sentence there, in the first paragraph. before you say something about the numbers? I think as long as you-
> K: uh. just a minute. are we talking about the first paragraph?
> Er: yeah. the first-
> C2: yeah yeah. that highlighted part. starting from like a 2012 Pew internet research. you know. instead of giving them the evidence or numbers right away. you can have some kind of topic sentence? like. it's really popular. you know?
> Er: you should ease them in because it's like "today we're gonna talk about ebooks" and then all of a sudden all of these numbers come up. ((laughs))
> C2: mmhmm. (4/24, line 357)

The focus on the organizational aspects of the prompts served the purpose of scaffolding the test-takers by providing them with the information needed to write the essays in a progressing manner. Thus, the writers expressed the importance of being efficient with their use of language, and avoiding repetitions.

> Cr: Watch out for sentences that essentially say the same thing as the previous sentence. but you're like- making it nice for them [test-takers]. Like you're explicit and you get more explicit. (5/1, line 328)

The discussions suggest that the prompts are not simply reproductions of factual information, but that there is intentional scaffolding in the manner they are written. Presenting the information in a gradually more explicit style, while avoiding repeating previous points, serves to guide the test-takers through the text, and would help them achieve more success in the assessment task. Another component of the presentation style of the prompts was the use of different registers for different sections of the test to simulate authenticity. Whereas the pro/con topics presented information in a formal manner to simulate the authenticity of reading academic articles, it was desirable for the lecture component to take a more conversational register, which had features such as second-person usage and shorter sentences.

> C1: I'm thinking that was really well-written lecture. but. we kind of make it more casual but why? and I thought the whole purpose for us to having this lecture in the EPT is kinda trying to mimic what they are going to be in the classroom setting. (4/24 Discussion, line 108)

In addition to avoiding repetition, the writers also discussed how to change sentences to be more concise through strategies of combining ideas and splitting sentences that are long, and thus difficult for students to follow. Word choice was also an important factor, as there was a preference to avoid words that are uncommonly used, or would be unfamiliar to English language learners. The choice of words also had the potential to highlight the important aspects of information in the prompt. This principle of guiding the test-takers with the language of prompts was also expressed by the coordinators.

> C1: So for instance like slide 5 then you can just talk about features, like have keywords? here and there. you know, and then guide the students. These are the topics that are discussed in the lecture for the slide and then guide people along. so that would help students to follow your lecture better. (4/24, line 12)

Attention to word choice was also important, as the writers were concerned with misleading the test-takers. They considered the ways in which the readers of the test may be confused by the language of the paragraph and misinterpret the points.

> Cr: Yes. I'm sorry. It reads to me a con, vocational schools, low employment.
> E: Low *unemployment*
> Cr: Oh!
> C2: Maybe high employment.
> Cr: Maybe high empl-yeah. (4/10, line 223)

In the interviews, the writers reported feeling that they were trying to simplify the language for an appropriate audience, to avoid making the language of the prompts too difficult for what they perceived to be the average test-taker. They also recognized the limited amount of time test-takers had to read the prompts, which further supports the need for efficiency in their style of writing.

> K: In the EPT we were trying to make it as simple as possible for maybe intermediate English students. (Interview, line 135)
>
> S: You want something people can digest in however much time they have… (Interview, line 189)

4.4.2 Consistency

Although there was the understanding among the team that the test prompts would need to have shared features, some of these features were less obvious at the beginning of the writing process and only emerged later through the discussions and written comments on the drafts. Explicit features, such as prompt length, order of paragraphs, and the use of article titles were relatively straight-forward to implement, and did not require much discussion. Sentence structures, as well as the editing of grammar and vocabulary also contributed to the uniformity of the test, and were more frequently addressed in the group discussions. Notably, the role of ensuring uniformity in the linguistic aspects of the prompts was mostly taken on by one writer, with most of the comments in the discussions coming from that writer.

> Cr: Yeah I think my biggest suggestion was like, your sentences flow in this way that rely on the these transitional words? Like take some of those off cause it starts feeling like a "which" sandwich… (4/10, line 190)

Other writers reported their recognition of the role of the “editing” writer as contributing to the uniformity of the test, because the style of the other writers would become influenced by the same writer throughout this process.

> I: So you think Cr is a major factor in having this uniformity [of the test]?
>
> M: I think so because she really, she took a lot of time doing the language editing and going through and looking at the details that I one hundred percent was not looking at, at

> all. And when you have one person doing the majority of the editing, then they're going to edit in a way that they're seeing the same things. (Interview, lines 120 – 122).

The team discussed the consistency of language and register across prompts, as well as consistency within prompts. These discussions were typically about the uses of terminology within the different sections of the prompt.

> C1: Oh yeah it's now digital texts.
> M: Yeah.
> C2: You have to be more consistent with that, that term.
> Cr: The way, yeah... Well it's something to put in the lecture, like... you know these are all the-cause there's a bunch of different ways to say it [digital texts] right? (4/10 discussion, line 230)

4.5 Test-Writer Development: Negotiating Feedback

To explore the aspects of test writing the participants developed through the course of test development, the interview questions related to this topic were coded and compared with the types of comments the writers received through multiple drafts. Three categories of comments were generated from the data, which include audience, prompt content, and writing style. Overall, the types of comments that the writers received on their drafts were consistent with the topics they discussed in the interviews and group discussions. One major topic of discussion from the interviews that were not as visible in the comments was learning to participate in the peer feedback process. The combined frequencies of the categories of comments for both prompts are presented in Table 4.1. Although some writers received fewer comments in the second prompt (suggestion some improvement), the change in frequencies across prompts were generally inconsistent. It is, however, interesting to note that the frequencies of comments in the second prompt across writers were more homogenous than in those of the first prompt (Figure 4.2). Especially for the category of writing style, the writers displayed more similar trends of comments across drafts than they did in prompt 1.

Table 4.1. *Combined frequencies of categories written comments across drafts for prompt articles and lectures*

Writer	Prompt 1 (Article)	Prompt 2 (Article	Prompt 1 (Lecture)	Prompt 2 (Lecture)
Cr				
Audience	1	2	0	2
Presenting Topic	15	36	8	11
Writing Style	12	27	18	6
K				
Audience	5	1	7	1
Presenting Topic	36	23	21	13
Writing Style	71	34	19	27
E				
Audience	4	1	0	2
Presenting Topic	20	30	14	11
Writing Style	44	24	19	15
M				
Audience	2	4	2	1
Presenting Topic	30	26	11	10
Writing Style	35	27	9	8
S				
Audience	3	5	2	4
Presenting Topic	25	33	26	10
Writing Style	36	30	16	21

Trends within prompts were more consistent (Figure 4.2). For the category of *audience*, comments were relatively infrequent and consistent in number across drafts. *Presenting topic* was more inconsistent, with fluctuating increases and decreases across drafts. Finally, *writing style* showed a more consistent downward trend across drafts, with the exception of the lectures in the second prompt.

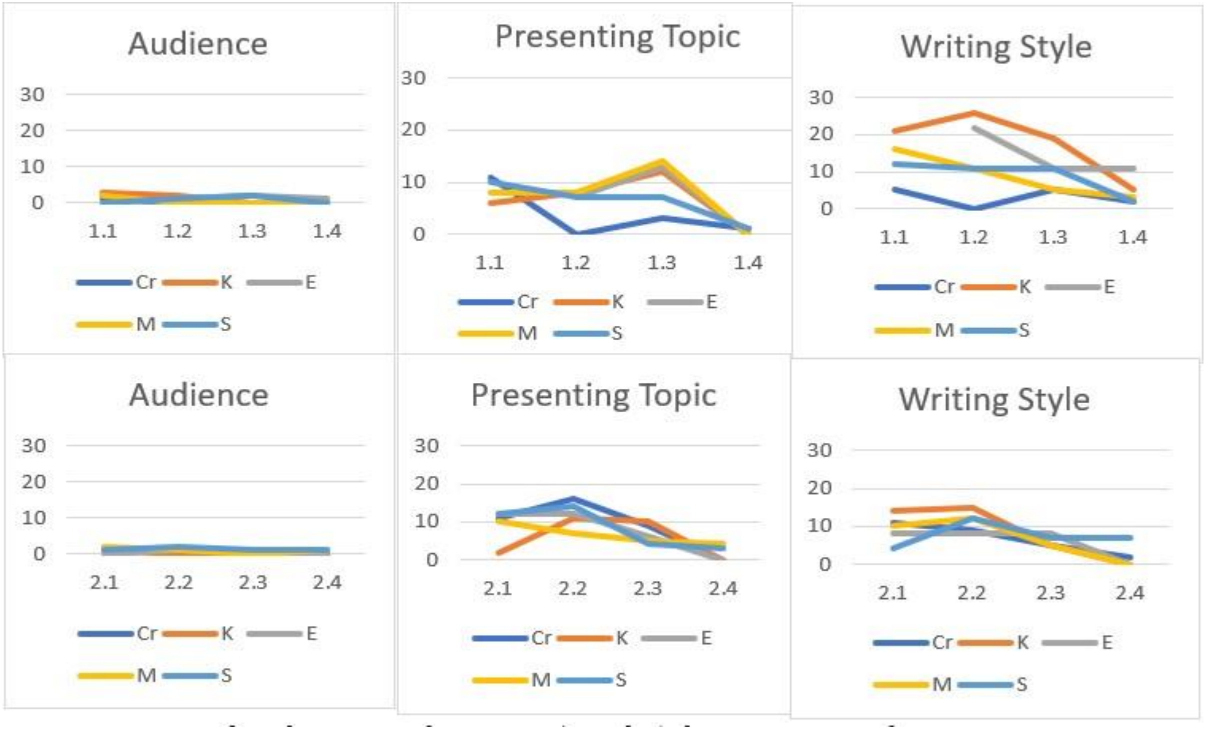


Figure 4.2. *Graphs of comment frequency (articles) for prompt 1 and 2*

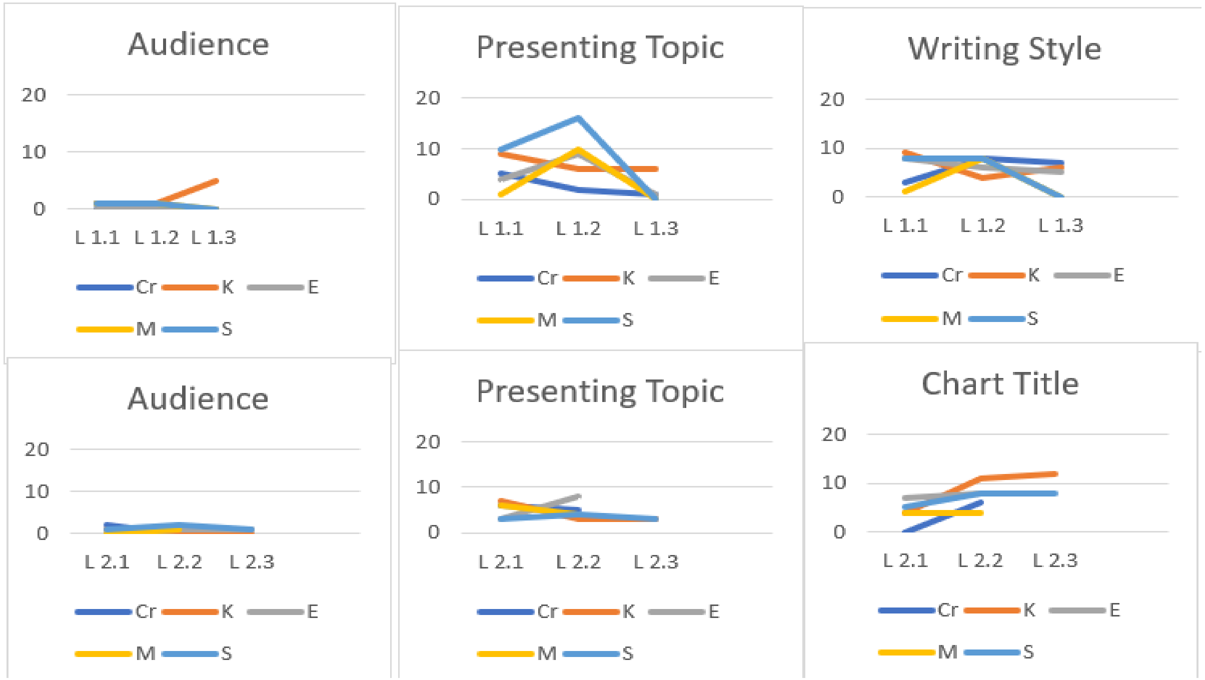


Figure 4.3 *Graphs of comment frequency (lectures) for prompt 1 and 2*

Consistent with the data from the individual interviews and draft comments, the writers seemed to easily negotiate the *audience* of their prompts. Only one writer (S) reported the benefits of the group discussions and perspective of international students in the interview, and this writer also received the most comments regarding audience. Generally, comments regarding considerations of audience were relatively rare compared to the categories of topic presentation and writing style.

> S: I think when things sound ambiguous maybe you're thinking, you're interpreting something one way and then someone comes around and they can have a completely different reading but also some of the aspects of cultural appropriateness? What might not be understood in one culture the same way it's understood how you're trying to write it, so that obviously takes multiple perspectives but I think it's interesting, that process. (Interview, line 15)

Comments related to *writing style* were mostly related to sentence structure and vocabulary usage. As mentioned previously, writing style was one of the more salient features for the writers in standardizing the prompts. Although comments related to writing style were quite frequent, they appeared easier to resolve, as suggested by consistent decreases in comment frequencies in this category with subsequent drafts.

> M: I almost always accepted the edits. Occasionally there would be one where I would be like-especially if it was in a section that I had to completely redo, where I would just end up deleting it because I was getting rid of that entire three sentences or something like that. But in terms of her [Cr] grammar edits, and sometimes sentence structure changes, I almost always accepted all of them. (Interview, line 126)

Data from the interviews and comments suggested that writing style and language-level edits required relatively little negotiation, as the writers deferred to the edits of one writer who had taken on the role of making those types of suggestions. It is also possible that writing style was seen to be less important to individual writers compared to handling the contents and topic information of the prompts.

> S: I can't help but feel like Cr with all her comments had better grammatical structures than me I have to think, oh yeah and because, I don't know, maybe I had more errors and repetitions of words just because it wasn't something I was paying attention to, especially when I was first, first drafts and stuff. (Interview, line 280)

Overall, the writers learned how to be more consistent across and within prompts by adopting features from other writers' drafts. Along the way, they used elements of writing that they would not normally use, or removed features of their writing that they felt did not match the other writers.

> K: I would look at peoples' work and borrow stuff on how to put it, like in the lecture we would say use the rhetorical questions, usually not my style. Then there's this other thing, especially the lectures, they were just very much the same. (Interview, line 175)

Presenting the topic/information was a frequent comment category that corresponds to the theme of *presentation of information* from the group discussion and interviews. Typical comments were related to suggestions for providing more examples, explanations, and creating arguable points by balancing the pro/con points and avoiding expressing personal opinions. Based on the writers' reports, the suggestions from both written comments and group discussions were vital in the selection of topic information and the manner in which to present them. Similar to the comments regarding audience, the writers also noted that having multiple perspectives were helpful in determining what kinds of concepts would be unfamiliar to a general audience, and would therefore require further explanation by the writers.

> K: Like I said before, it's really good to get different people to read your work and to see how readers perceive your work, yeah so sometimes you, I would assume that this is obvious to everybody and people would tell me no, that most people don't know this, so you have to say this and that… (Interview, line 115)

In some cases, the writers found it especially helpful to have perspectives from the group members when they were experts in the topic of their prompts, either from the research they had completed prior to writing or previous familiarity with the topics. Having the group members point out the difficulty of the concepts helped individual writers write from the perspective of someone with a more general understanding of the topic, as a test-taker would.

> S: Just cause the topic I had the second time I think that people pointed out it was a bit more, it got a bit more technical especially because I had some familiarity in the field and could get more technical about it, so I think that affected a lot, because then I kind of have to rewrite things and try to write the first time when I don't have that sort of knowledge about it. (Interview, line 66)

Although generally the writers seemed to respond positively to feedback regarding the contents of the prompts, there did not seem to be a consistent decrease in the frequency of comments in this category across drafts. Occasionally, some writers would receive more prompt content comments in subsequent drafts rather than fewer. It appeared that developing an understanding of the kinds of information to include, as well as presenting the topic to students in a suitable manner required more negotiation relative to comments related to audience and writing style.

> M: I had to keep changing my reference on how to actually present it, because it wasn't-how I was writing it wasn't conveying it in the most neutral manner or the most easily understandable manner. So it was-I had to ask for other peoples' opinions. "What do you think about this?" And it just really stre-I think if you showed me that paragraph from the first time I presented that topic, there's no way I would have guessed that that's where I would've gotten… (Interview, line 46)

4.6 Attitudes Towards Written Feedback

One of the essential aspects of the test development process was the peer feedback that was given and received by the writers. As an institutional test, the EPT was recognized as a collaborative effort, and the writers understood the importance of collaboration with others, even though they were writing prompts on different topics.

> M: Yeah, it's not, like I said this before, but it's really not about your own writing, it's about your ability and willingness to collaborate with other people and to take their criticism seriously and to really listen, and then implement what you heard. (Interview, line 150)

However, there was also a sense of understanding among the test writers that they each had their individual strengths and perspectives on certain aspects of test writing and the contents of the topics themselves. The writers took into account these individual differences when they provided feedback to the other writers to avoid offense to others.

> Cr: ...but I think with the google docs I am also very critical of the way I word things and I think that's because of my creative writing background and having to be very sensitive and hedge when you say things so you don't want to ever tell people "you have to change this", oh maybe you could consider doing this right? (interview, line 67)

Generally, the writers were able to collaborate and learn from each other during peer feedback and were able to resolve the majority of their disagreements or differences on how to write the prompts. However, this did not result in complete conformity or completely uniform test prompts. In some cases, disagreements regarding the content of the prompts were not resolved.

> S:I generally will implement whatever that feedback is, unless I have a really strong reason for disagreeing with it, which I did a couple of times, I had in the EPT process, I didn't agree with something that someone was saying and I would keep it the way it was, but the nice thing about the EPT is usually there's at least one other person that agrees with you. (Interview, line 158)

Within the group of writers, there were sub-groups that tended to share similar opinions. Therefore, not all members of the team needed to agree, as long as there was support from some of the members. Although peer feedback played a major role in standardizing the prompts in terms of style, other elements of individual writers were preserved in finalized drafts.

> S: I feel like people remained fairly distinct still, but maybe there was some similarities like maybe the structures, we started using similar structures or something? (Interview, line 288)

Chapter 5: Discussion

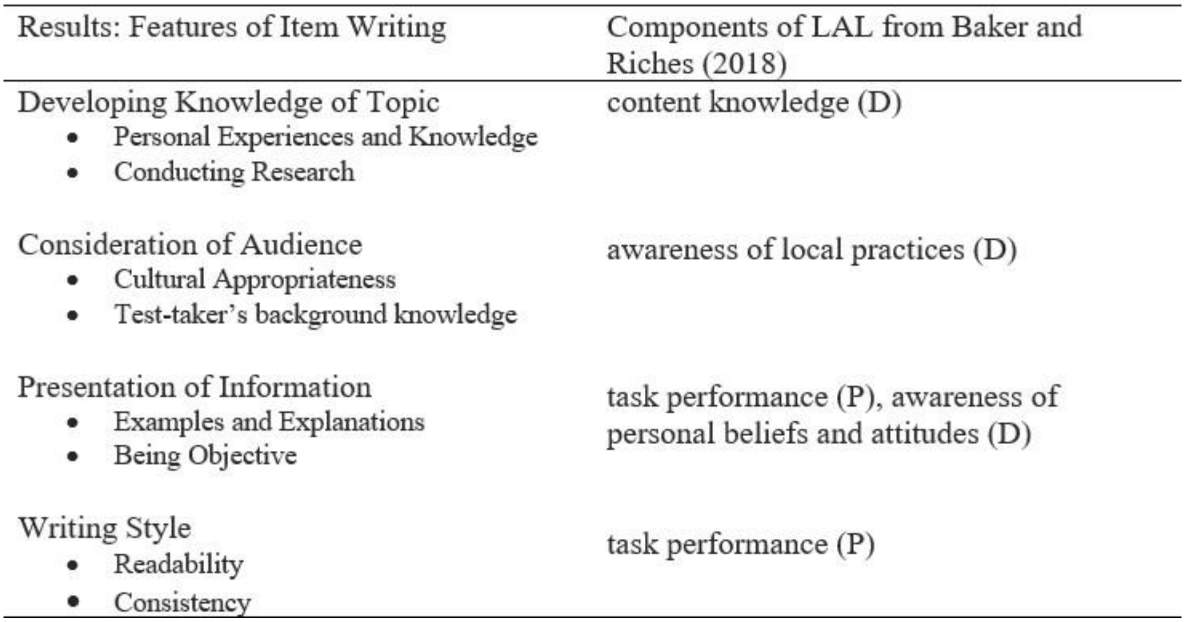
5.1 Features of Item Writing and Connection to LAL

The EPT development took place over the course of two months, during which coordinators collaborated with novice test writers in the form of written feedback and group discussions. Although there were instances when coordinators would provide explicit instruction to the writers, the majority of the valued features of the test prompts were negotiated through discussions. The purpose of observing the group discussions and keeping track of prompt drafts during this period was to explore the features of item writing that characterize this process within this context. Having a general understanding of item writing will help to make connections between the practices of this activity and the components of LAL. Another purpose of this project was to explore the potential of peer feedback and other aspects of collaborative test writing in developing item writing practices. Understanding how item writing practices develop has implications for the role that item writing plays in developing LAL.

Previous studies of the practices of item writing such as Ryan and Brunfaut's (2016) draw attention to the important role that collaboration plays in the process. This is readily apparent in this study, with the frequent interactions and various forms of feedback and exchange between the participants. To explore these social interactions further, this study conceptualized item writing as genre practices. Item writing can be seen as specialized forms of literacy practices, as writers are involved in the social production of texts for the specific purpose of assessment. As evidenced by the comments on writing style in this study, the writers developed shared practices that accomplish those purposes in this context. Undoubtedly. writers in other test development contexts form their own idiosyncratic practices through their collaborative work. As part of LAL development, item writers engage not only with their peers, but also with the written items as the texts of test development. The ways they talk about their writing and the intertextual references they make to their peers' writings contribute to the reflective practices (Kleinsasser, 2005) that lead to LAL development. Although the individual writers end up adopting shared practices through these intertextual interactions, they retain some of their own individual genre practices by negotiating with their peers. The competencies of giving and receiving feedback developed in this process, and led these writers to recognize the sociocultural values of this context so that they could adjust their writing practices to create successful prompts. These genre practices of item writing can either be explicit or implicit, as declarative or procedural knowledge. In either case, they gradually appeared more consistently across the prompts of the writers through an intertextual socialization process, and the analysis of this study generated salient features that can be linked to the components of LAL.

Table 5.1 presents the features of item writing with their corresponding components of LAL adopted from Baker and Riches (2018). Based on the results of this study, a new category of "content knowledge" was generated from the data.

Table 5.1 *Features of item writing and corresponding components of LAL*



Results: Features of Item Writing	Components of LAL from Baker and Riches (2018)
Developing Knowledge of Topic • Personal Experiences and Knowledge • Conducting Research	content knowledge (D)
Consideration of Audience • Cultural Appropriateness • Test-taker's background knowledge	awareness of local practices (D)
Presentation of Information • Examples and Explanations • Being Objective	task performance (P), awareness of personal beliefs and attitudes (D)
Writing Style • Readability • Consistency	task performance (P)

In their previous study of LAL development, Baker and Riches (2018) provided a further distinction for the categories as referring to declarative knowledge or procedural knowledge. This distinction is useful here to observe that the practices of item writing, although largely associated with the procedural nature of task performance, also involves components of procedural knowledge (awareness of local practices, awareness of personal beliefs/attitudes, and content knowledge).

The item writing features of *presentation of information* and *writing style* both correspond to the LAL component of task performance, which involves the processes involved in test design and development. The group discussions frequently addressed topics related to the technical issues of item writing style, such as adhering to the prompt template and presenting the information in a consistent manner. Writing style played a significant role in the presentation of the information and creating readable and consistent prompts that would be appropriate for assessment purposes. Similarly, the presentation of information was an important feature of creating successful prompts; good examples and explanations were needed to elicit the desired responses from test-takers. Also, it was important for the information to be presented in an objective manner so that students would not be led to argue for only one side of the issue.

The presentation of information was also related to an awareness of personal beliefs and attitudes. Although this was an institutional placement test, the item writers had the mindset of ESL teachers in their desire to support the success of the test-takers. This was realized through the process of the group discussions and expressing sympathetic views towards the test-takers. The writers were interested in providing clear explanations and examples for concepts they perceived as difficult for ESL students. Accuracy of the information provided was also an important consideration; some writers viewed the prompt articles and lectures as potential ways that international students could learn about American culture and the education system. Although the purpose of the prompts is for assessment, the item writers did not view their writing as serving only that function. Rather, the writers enacted the genre practices to create prompts that served another function of educating the test-takers about a new topic. This is similar to previous findings (Baker & Riches, 2018) of teachers realizing that language tests can be a way for teachers to support students in their learning. As pre-service teachers, the EPT writers viewed the test as having both assessment and educating functions. Overall, the item writers maintained a strong connection to their identities as teachers and felt that the role of their prompts was to provide information in a way to help the students succeed in making their arguments. They discussed the amount of *scaffolding* that would be useful for the test-takers to grasp the information within the prompts without providing too much help as to affect the validity of the test. Rather than a simple presentation of information, the writers discussed strategies of writing for the audience of test-takers by providing explanations and examples where necessary or reiterating key concepts within their topics to help the test-takers focus.

The *consideration of audience* corresponds to the LAL component of awareness of local practices. The writers frequently discussed the suitability of the prompts for international students coming to study at a U.S. university. The cultural suitability of the topics were negotiated quite often due to the perceived diversity of the test-takers, but also because of the diverse background and experiences of the item writers. The writers had varying beliefs about the characteristics of the international students and pieced together a collective conceptualization of their audience based on their experiences and knowledge of particular cultural groups. It is important to note that although item writers may be quite experienced in their local contexts, they are not homogenous in their understandings and beliefs of the local practices. Part of the consideration of audience for the item writers is the realization and awareness of different perspectives of the local practices. Therefore, the process of addressing the aspects of the U.S. cultural context and suitability of the prompts required some negotiation among the writers. Aside from considerations of cultural knowledge, the writers were also concerned with the ways that the cultural characteristics of the students may influence their pro/con opinions and arguments. It was helpful for the team to discuss the possible perspectives and attitudes that the international students have towards the topics of the prompts. Doing so allowed the writer to avoid framing their prompt in a way that would elicit only one-sided responses from the test-takers. One of the goals of the test was to have the test-takers consider both sides of the argument before writing, but topics that test-takers of a particular culture may have strong feelings towards could lead them to consider only one side and be a detriment to the validity of the test. Part of the concern for the writers also stemmed from their experiences as ESL teachers in the university writing courses. Most of them had taught the courses that the EPT would place the students into and had insights into the emphasis that the courses place on teaching the students how to construct objective arguments with evidence. References to what the students would be expected to demonstrate in the ESL courses, which included both language and rhetorical skills, affected the discussions during item writing on how to elicit good pro/con writing from the test-takers. The writers' awareness of the local practices helped with these considerations of writing test prompts that were both suitable for international students and the institutional purposes of the U.S. university.

A new component of LAL, *content knowledge,* was generated from the analysis of group discussions and interviews. In language test development, it is apparent that those involved in the process should have knowledge of the content (e.g. language) being tested in addition to the theoretical knowledge of assessment (Ryan & Brunfaut, 2016). The EPT, as an integrated writing test, is an assessment of language skills but also an assessment of the writing skills associated with academic writing, such as supporting arguments with sources. To create effective prompts, it was necessary for the writers to obtain information and knowledge of the subject matter: the topics of their prompts. Information was consolidated from sources, typically online journals and websites, as well as from other writers' personal knowledge and experiences. As the writers developed better understandings of their topic, they were able to present their ideas in the prompt in more effective ways for assessment purposes. For example, a common writing practice was to simplify the information that the writers gathered, so that the test-takers would not be overwhelmed by the amount of information, but still be able to grasp the main points. Therefore, understanding the topics was an important step before the writers could begin to write appropriate prompts. Unique to this context and type of test is also the development of *unbiased* understandings of the topic. Although the writers found a substantial amount of topical information from the sources, as well as from the personal experiences and knowledge of other writers, they quickly recognized that the information they found could be opinionated. Recognizing biased information helped the writers to present the topics in an objective manner and avoid skewing the information towards one side of the argument. This study proposes content knowledge as an addition to the previous framework of LAL proposed by Taylor (2013) and modified by Baker and Riches (2018).

5.2 Test-writer development as LAL development

This study took a CoP orientation towards the socialization processes of the writers. The writers' participation in the activities of item writing facilitated their development of genre practices, which in turn are connected to various components of LAL. The activities that lead to learning include not only the writing of the items themselves, but also learning to participate in the discourses about the items, such as critiquing peers. Although the writers did not always agree on what to include in the prompts, they developed ways to participate in the negotiation of these disagreements, which resulted in a more streamlined and effective writing process. Developing these types of competencies involved understanding what types of feedback to give (e.g. writing style, content) and also how to give feedback in a diplomatic manner.

The data from the interviews and comments revealed that collaborative writing and peer feedback played roles in developing three of the four previously identified features of item writing: consideration of audience, presentation of information, and writing style. Although the interviews and comments did not explicitly mention content as an area of development, it is likely that the item writing process contributed to this as well. This is because writing the prompts required displays of knowledge from the writers about their individual topics in order to elicit feedback on how to best present the topic information in the prompts. It is likely that having the multiple perspectives of the team members helped each writer to develop better understandings of the topics. Additionally, some writers had previous knowledge of the topics of their peers, which they shared during the group discussions. Although it is unclear to what extent the item writing process facilitated the development of content knowledge, this could be further explored as a potential area of development in future studies.

Two features of item writing were relatively easy to develop for the item writers: consideration of audience (corresponding to *awareness of local practices*) and writing style (*task performance*). Because the item writers shared several crucial characteristics and experiences, such as their role as ESL teachers and interactions with international students, the suitability of the test material was a relatively easy feature to negotiate. Although the writers had different kinds of knowledge of the local context, they tended to be open to others' opinions and readily changed their perspectives on what would be suitable for the test-takers. Generally, all writers were able to participate in these discussions and had anecdotes or examples from their interactions with their students to contribute. As such, the writers had little difficulty in understanding that international students would be a heterogeneous group, and that it was valuable for them to share their knowledge of different cultural groups to write prompts for a "general" international audience. Writing style also developed consistently throughout the drafts for several reasons. These types of revisions were easily made because they were either changes in language or organization, which involved simple fixes such as word choice or reordering of sentences. Also, the writers reported a shared value of making the prompts more uniform. Part of the process of achieving uniformity involved the writers learning to rely more on others' feedback or edits. Although initially writers wrote in the style they were most comfortable with, they were gradually socialized into values of uniformity throughout the writing process and became more accepting of writing style edits to their prompts. The writers recognized that exchanging and sharing their unique ways of writing naturally resulted in prompts that were more similar to each other in terms of language and style.

Presentation of information (*task performance, awareness of personal beliefs/ attitudes*) was a more difficult feature of writing to negotiate, as reflected by the inconsistent patterns of comment frequencies throughout the drafts. In this context, the component of task performance was more difficult to enact. Common issues included how much accommodation should be provided to the writers in terms of explaining and providing examples. Sometimes writers disagreed on how well test-takers would be able to grasp certain concepts, or perhaps had disagreements on the effectiveness of particular examples and explanations. However, despite these difficulties, the writers reported the helpfulness of discussing how to present the topic. Having multiple perspectives allowed the writers to see how the material in a prompt could be difficult for some readers and resulted in iterative fine-tuning of the prompts throughout the drafts, as evidenced by the frequency of comments in later drafts. On the other hand, *awareness of personal beliefs and attitudes* seemed to develop relatively smoothly through written comments and reports from the interviews. The writers typically reinforced their beliefs about writing prompts for the success of test-takers and appreciated peer feedback on best practices for scaffolding information. “Fact-checking” was also a common practice in the comments; writers questioned information they believed to be erroneous, or those that they believed to be more opinion than fact. These results make sense, again in light of the writers’ backgrounds as ESL teachers. Although they may have differences in opinion on how to present information, they share similar values in their student-centered orientations and desire to present accurate information.

A key aspect of the writers' development was their participation in the processes of peer feedback and revision. Collaboration, as Baker and Riches (2018) noted, has traditionally not been considered a component of LAL in the literature. Evidence from this study supports their proposal of including collaboration, a professional skill, in the framework of LAL. Within a CoP framework (Wenger, 1998), the participation and collaboration of the writers constitutes learning and development. It was observed how the writers shifted towards a more collaborative orientation as they participated more in item writing activities in the form of discussion and feedback. The writers reported being able to provide and receive criticism as essential to their success on the project. Item writing has the potential to be a source of conflict in collaborative contexts, as the abilities and personal values of the writers may be challenged by others. Notably, one writer made explicit mention of their care in hedging the comments they provided writers to avoid a prescriptive tone of suggestion. This is also evidenced by the wording of the comments from all the writers, who would often frame suggestions indirectly in the form of questions, or express their lack of understanding rather than attributing fault to the writing. Although the general atmosphere of the group was friendly, occasionally there would be moments of conflicts when members of the group held strong opinions on certain topics. As some writers noted, they had to learn to compromise some of their personal values (e.g. writing style, opinions of topics) in order to be successful in the collaborative writing process. Some reported that they had learned how to accept criticism better than they had before the test development process, especially learning to not take feedback from others personally. The writers expressed that they felt they were able to gain a better understanding of the other writers in terms of their concerns in prompts, and this helped to facilitate smoother discussions and feedback. Some also recognized the value of mutual understanding within the team and sought to help mediate feedback between writers with conflicting opinions.

Understanding the concerns of other writers allowed them to revise their prompts to the satisfaction of the group. The writers also demonstrated an awareness of their roles within the group during the interviews. As these roles became more defined, they became more comfortable in providing certain types of comments. For example, one writer was concerned with the language of the prompts and became the main commenter of that feature of prompt writing. Other members gradually recognized this role of that writer and generally deferred to that writers' opinions for language-related issues. Within this CoP of item writing, the writers developed repertoires for participating and found ways to contribute to the shared practices of item writing.

Chapter 6: Conclusion

This study examined the features of item writing to determine its relationship to the components of LAL and followed by examining the LAL development of pre-service teachers through the socialization process of peer feedback. Although producing good test items has been acknowledged as an essential skill for language teachers, the activity of item writing itself has been overlooked as a potential contribution to LAL development. Findings from this study make a case for closer investigation of this process in further studies.

Perhaps the most significant finding is the role that discussion and collaboration have on the development of item writing processes, and subsequently LAL development. Although there were instances where the coordinators gave explicit instructions on how to write items, item writing practices developed more from the activities of writing and discussion. This lends support to the idea of participation in test development as learning for pre-service teachers and invites further exploration of how item writers learn to write by examining how they participate in activities such as feedback and group discussions. The collaborative context of test development was essential for item writing and LAL development. Although initially, writers reported that they struggled to create good prompts, they gained confidence and better understandings of how to write as they engaged more with feedback from their peers.

It should be noted that some features of item writing were more easily negotiated than others and that LAL does not develop consistently. For example, writing style appeared to be the most easily negotiated feature of writing (corresponding to *skills in assessing*). After the initial conflicts or clashes in writing styles among the individual writers, the writers gradually came to be more accepting of edits of their writing and conformed to a particular style of writing for the EPT. This shared value of uniformity was negotiated through the group discussions, and the writers' sense of item writing as a collaborative effort helped to ease development in this area. Conversely, presenting the topic to the test-takers appeared to be more a difficult feature to negotiate for the writers. Interestingly, comments in this feature of item writing did not clearly decrease through subsequent prompts but remained fairly consistent. Despite the difficulties in this area, negotiation and collaboration remained a helpful and necessary aspect of test development. The disagreements between writers allowed them to reflect on their own writing practices and question their own perspectives and approaches, as well as their roles in this community of practice. This type of reflection contributes to LAL development in that it allowed item writers to recognize the value of collaboration and develop LAL as a professional competency (Baker & Riches, 2018).

Although the literature in language testing considers content knowledge an essential aspect of language test development, it seems that content knowledge has not been included as a component of LAL. One possible reason is that as content knowledge is closely associated with the domain of language teachers, it is assumed that language teachers already possess this (e.g. knowledge of English language), and do not need to develop it as part of LAL. However, there are many types of content/subject matter knowledge, especially in the domains of ESP and EAP, that language teachers may not possess. Also, existing content knowledge may need to be developed further for assessment purposes. For example, in this study, it was necessary for writers to gain a certain level of understanding about their topics so that they could understand which parts to present and simplify for test purposes. Perhaps future frameworks of LAL could include content/subject matter knowledge as an additional component.

Finally, this study reconceptualizes item writing as constituted by genre practices. Although this particular type of item writing, which involve writing prompts, readily lends itself to examination as texts, other types of tests could possibly be examined from genre theory perspectives. Conceptualizing test items as texts opens new avenues of research in LAL research by exploring how item writers and different stakeholders interact with test items. As we have observed in this study, the intertextual interactions of the writers contributed to their LAL development as a social process.

Appendix A: EPT Interview Guide

Developing as Item Writers

1. Before writing for the EPT, did you have any background in test writing?
    - did your background affect your item-writing?
2. What ideas or impressions did you have about item-writing before writing for the EPT specifically?
    - What about now that you've finished writing?
    - What did you learn from the process of writing this semester?
3. Describe your general process/strategies for writing items for the EPT in the beginning of the semester.
    - What about your process/strategies for writing now?
4. What kind of advice would you give to next years' EPT writers?

Group Interactions

1. Describe your role in the group during and outside the group meetings.
2. What were your perceptions of the group dynamics?
3. What did you learn from the group meetings?
    - How do you think the group meetings influenced you as a writer?

Writing Practices

1. How would you compare the writing you do in general to the writing you did for the EPT?
    - How did you adjust to any differences in EPT writing from your writing in general?
2. How would you compare your writing to the writings of other members of the team? What similarities or differences do you see in your writings?
3. What factors did you consider when choosing topics to write about for the EPT?
    - What impact, if any, did the topic affect your approach to writing?

4. Which aspects of writing for EPT did you focus on the most? (e.g. content, organization)

Finally, is there anything else you would like to say about the EPT writing process, or anything else in general?

References

Bailey, K. M., & Brown, J. D. (1996). L2 Language Testing Courses: What Are They?. *Validation in language testing*, *2*, 236.

Baker, B. A., & Riches, C. (2018). The development of EFL examinations in Haiti: collaboration and language assessment literacy development. *Language Testing*, *35*(4), 557-581.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007?. *Language Testing*, *25*(3), 349-383.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science* (Vol. 356). Madison: University of Wisconsin Press.

Bazerman, C. (1994). Systems of genres and the enactment of social intentions. *Genre and the new rhetoric*, 79101.

Bazerman, C. (2003). Intertextuality: How texts rely on other texts. In *What writing does and how it does it* (pp. 89-102). Routledge.

Beaufort, A. (1999). *Writing in the real world: Making the transition from school to work.* New York: Teachers College Press.

Bhatia, V. K. (2008). Genre analysis, ESP and professional practice. *English for specific purposes*, *27*(2), 161-174.

Brindley, G. (2001). Language assessment and professional development. *Experimenting with uncertainty: Essays in honour of Alan Davies*, *11*, 137-143.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007?. *Language Testing*, *25*(3), 349-383.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL students and the effects of training in test development principles and practices on improving test quality. *System*, *37*(2), 226-242.

Coombe, C., Troudi, S., & Al-Hamly, M. (2012). Foreign and second language teacher assessment literacy: Issues, challenges, and recommendations. *The Cambridge guide to second language assessment*, 20-29.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing writing*, *28*, 43-56.

Davidson, F., Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Vol. 7). Cambridge University Press.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, *25*(3), 327-347.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 419-438.

Devitt, A. J. (1991). Intertextuality in tax accounting: Generic, referential, and functional. *Textual dynamics of the professions: Historical and contemporary studies of writing in professional communities*, *336*, 357.

Duff, P. A. (2007). Second language socialization as sociocultural theory: Insights and issues. *Language teaching*, *40*(4), 309-319.

Duff, P. A. (2010). Language socialization into academic discourse communities. *Annual review of applied linguistics*, 30, 169-192.

Fulcher, G., Davidson, F. (2007). *Language testing and assessment*. London: Routledge.

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

Hills, J.R. (1991). Apathy concerning grading and testing. *Phi Delta Kappan*, *72,* 540-545.

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

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing*, *27*(4), 555-584.

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., & Roullion, V. (2010). A Case Study on an Item Writing Process: Use of Test Specifications, Nature of Group Dynamics, and Individual Item Writers' Characteristics. *Language Assessment Quarterly*, *7*(2), 160-174.

Kleinsasser, R. (2005). Transforming a post-graduate level assessment course: a second language teacher educator's narrative. *Prospect, 20,* 77-102

Kobayashi, M., Zappa-Hollman, S., & Duff, P. A. (2017). Academic discourse socialization. *Language socialization*, 239-254.

Kristeva, J. (1980). *Desire in language: A semiotic approach to literature and art*. Columbia University Press.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing*, *32*(2), 169-197.

Lave, J., Wenger, E., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation* (Vol. 521423740). Cambridge: Cambridge university press.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, *23*(2), 26-32.

Malone, M. E. (2008). Training in language assessment. In *Encyclopedia of language and education* (pp. 2362-2376). Springer, Boston, MA.

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

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension* (Vol. 1). John Wiley & Sons.

Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education*, *120*(2), 285-285.

Mertler, C. A. (2003). Preservice Versus Inservice Teachers' Assessment Literacy: Does Classroom Experience Make a Difference?. Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Columbus, OH.

Morita, N. (2000). Discourse socialization through oral classroom activities in a TESL graduate program. *Tesol Quarterly*, *34*(2), 279-310.

Myhill, D. (2005). Prior knowledge and the (re)production of school written genres. In T. Kostouli (Ed.), *Writing in context(s): Textual practices and learning processes in sociocultural settings* (pp. 117–136). New York: Springer.

Okuda, T., & Anderson, T. (2018). Second Language Graduate Students' Experiences at the Writing Center: A Language Socialization Perspective. *TESOL Quarterly*, *52*(2), 391-413.

Parkinson, J., Demecheleer, M., Mackay, J. (2017). Writing like a builder: Acquiring a Professional genre in a pedagogical setting. *Journal of English for Specific Purposes*, *49*, 29-44.

Pennycook, A. (1999). Introduction: Critical approaches to TESOL. *TESOL quarterly, 33*(3), 329-348.

Popham, W. J. (2001). *The truth about testing: An educator's call to action*. ASCD.

Popham, W. J. (2004). Why Assessment Illiteracy Is Professional Suicide. *Educational Leadership*, *62*(1), 82.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental?. *Theory into practice*, *48*(1), 4-11.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265-273.

Purcell-Gates, V., Duke, N. K., Martineau, J. (2007). Learning to read and write genre-specific text: Roles of authentic experience and explicit teaching. *Reading Research Quarterly, 42*(1), 8-45.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, *18*(4), 429-462

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

Schafer, W. D. (1993). Assessment literacy for teachers. *Theory into practice*, *32*(2), 118-126.

Schieffelin, B. B., & Ochs, E. (1986). Language socialization. *Annual review of anthropology*, *15*(1), 163-191.

Schutz, A., & Luckmann, T. (1973). *The Structures of the Life-world* (Vol. 1). Northwestern University Press.

Seloni, L. (2012). Academic literacy socialization of first year doctoral students in US: A micro-ethnographic perspective. *English for Specific Purposes*, *31*(1), 47-59.

Shohamy, E. (1998). Critical Language Testing and Beyond. *Studies in educational evaluation*, *24*(4), 331-45.

Shohamy, E**.** (2001). *The power of tests*. Harlow, England: Pearson Education.

Shohamy, E. (2017). Critical language testing. *Language testing and assessment*, 441-454.

Stiggins, R. J. (1988). Revitalizing classroom assessment: The highest instructional priority. *The Phi Delta Kappan*, *69*(5), 363-368.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*(7), 534-39.

Stiggins, R. J. (1991b). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, *10*(1), 7-12.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238.

Stiggins, R. J. (1999). Assessment, student confidence, and school success. *The Phi Delta Kappan*, *81*(3), 191-198.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758-765

Talmy, S. (2008). The cultural productions of the ESL student at Tradewinds High: Contingency, multidirectionality, and identity in L2 socialization. *Applied Linguistics*, *29*(4), 619-644.

Tardy, C. M. (2005). ‘‘It's like a story’’: Rhetorical knowledge development in advanced academic literacy. *Journal of English for Academic Purposes*, *4*(4).

Tardy, C. M. (2006). Researching first and second language genre learning: A comparative review and a look ahead. *Journal of Second Language Writing*, *15*(2), 79-101.

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

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language testing*, *30*(3), 403-412.

Uhrig, K. (2012). Cognitive Responses to Sociocultural Elements in Professional Graduate Programs: Two Case Studies. *TESOL J*, *3*(1), 87-109.

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

Volante, L., & Fazio, X. (2007). Exploring Teacher Candidates' Assessment Literacy: Implications for Teacher Education Reform and Professional Development. *Canadian Journal of Education*, *30*(3), 749-770.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge university press.

Winsor, D. A. (1996). *Writing like an engineer: A rhetorical education*. Mahwah, NJ: Lawrence Erlbaum Associates

Yin, M. (2010). Understanding classroom language assessment through teacher thinking research. *Language Assessment Quarterly*, *7*(2), 175-194.

Zappa-Hollman, S. (2007). Academic presentations across post-secondary contexts: The discourse socialization of non-native English speakers. *Canadian Modern Language* Review, 63(4), 455-485

Zuengler, J., & Cole, K. (2005). Language socialization and second language learning. *Handbook of research in second language teaching and learning*, *1*, 301-316.