Chapter 3
Revisiting the Role of Content in Language Assessment Constructs

Lorena Llosa

Abstract In 1990, Bachman first introduced his model of communicative language use, which focused not just on an individual's communicative language ability, which he defined as language competence plus strategic competence, but also psychophysiological mechanisms, the language use context, and the language user's knowledge structures. In this paper, I first review Bachman's model, specifically his conceptualization of the role of content (or knowledge structures) in language use and in construct definitions. I then examine how his conceptualizations have been taken up and built on in the language assessment practices of three key contexts in which content and language intersect: language for specific purposes, U.S. K-12 English learner education, and content and language integrated learning. For each context, I highlight examples of the latest conceptualizations of the role content plays in their assessment constructs. I conclude by arguing that, if we are to develop language assessments that yield meaningful interpretations about test takers' ability to use language in specific target language use domains, future research must focus on the role of content in language assessment constructs.

Introduction

Bachman's model of communicative language use (Bachman, 1990) has had a major impact in the field of language assessment specifically, and in applied linguistics more broadly. Bachman's model, which builds on Canale and Swain's (1980) conceptualization of communicative competence, involves not just an individual's communicative language ability, which he defined as language competence plus strategic competence, but also psychophysiological mechanisms, the language use context, and the language user's knowledge structures. At a time when the focus tended to be on grammatical and textual knowledge, this model highlighted the complexity of language use and of language competence itself. This model, updated by Bachman and Palmer (1996), together with Bachman's concept of test method facets (later

L. Llosa (✉)
New York University, New York, United States
e-mail: lorena.llosa@nyu.edu

G. J. Ockey and B. A. Green (eds.), *Another Generation of Fundamental Considerations in Language Assessment*, https://doi.org/10.1007/978-981-15-8952-2_3

referred to as the task characteristics framework in Bachman & Palmer, 1996) was the foundation of his conceptual framework, which guided his approach to language assessment research and development.

Bachman's model, which has been influential in defining the constructs of many language assessments in use today, has been the subject of much theoretical discussion, particularly in terms of the role context plays in defining language assessment constructs (see Bachman, 2007; Chapelle, 1998; Chalhoub-Deville, 2003). One component of his model that has received relatively less attention is content, which he referred to initially as knowledge structures (Bachman, 1990) and later as topical knowledge (Bachman & Palmer, 1996, 2010). The field of language for specific purposes (LSP) is one exception since, due to its very nature, it has had to grapple with the role of content in language assessment (Douglas, 2000).

I argue that content has become more important in language assessment since Bachman's model was introduced 30 years ago due to a number of changes in the nature of language education and in the field of language assessment. Language education has been moving toward approaches that integrate content and language, many of which, like bilingual education and content-based instruction, are not new. We continue to see them used in schools throughout the world to address the educational needs of students who, due to globalization and immigration, are learning content through a second or additional language. In recent decades, instructional approaches that integrate content and language have expanded further. One example is the content and language integrated learning (CLIL) movement, initially active in Europe and now also in Asia and Latin America. There also has been a rapid increase in the number of English-medium universities located in places where English is a second or foreign language (Coyle, Hood, & Marsh, 2010).

The field of language assessment research has also changed since Bachman's model was introduced. It has expanded beyond the study of high-stakes, summative tests of English proficiency to focus on classroom assessments used for summative and formative purposes. Moreover, since language classrooms are increasingly becoming spaces in which language and content intersect, assessments in these spaces have to account for the role of content.

In this paper, I first review Bachman's model of language use, specifically his conceptualization of the role of content in language use and construct definitions. I then examine how his conceptualizations have been taken up and built on in the language assessment practices of three key contexts in which content and language intersect: LSP, the education of English learners in U.S. K-12, and CLIL. I highlight examples of the latest conceptualizations of the role content plays in each context's assessment constructs. I conclude by arguing that, if we are to develop language assessments that yield meaningful interpretations about test takers' ability to use language in specific target language use (TLU) domains, future research must focus on the role of content in language assessment constructs.

Historical Perspective: The Role of Content in Bachman's Model of Communicative Language Use

One of Bachman's main contributions is his model of communicative language use, first introduced in his book, *Fundamentals Considerations in Language Testing* (1990). His model includes three components—language competence, strategic competence, and psychophysiological mechanisms—that interact with "the language use context and language user's knowledge structures" (p. 84). Content, or "knowledge structures," is defined as "sociocultural knowledge, 'real-world' knowledge." The role of content in the model is only addressed within the definition of strategic competence: "Strategic competence thus provides the means for relating language competencies to features of the context of the situation in which language use takes place and to the language user's knowledge structures" (p. 84).

In their 1996 book, *Language Testing in Practice,* Bachman and Palmer refer to content as "topical knowledge." Topical knowledge plays a role similar to that of knowledge structures in the 1990 model, and, like knowledge structures, represents a broad definition of content, ranging from the topic of a particular reading passage to a specific subject area. In the 1996 model, topical knowledge interacts with language knowledge, the test takers' personal characteristics, and the characteristics of the language use or test task situation and setting through strategic competence and affective schemata. Bachman and Palmer (1996) described this as "an interactional framework of language use" that presents "a view of language use that focuses on the interactions among areas of language ability, topical knowledge, and affective schemata on the one hand, and how these interact with characteristics of the language use setting, or test task, on the other" (p. 78). Bachman and Palmer went on to address the role of topical knowledge in defining the construct for language assessments. They questioned the commonly held belief at the time that topical knowledge is always a source of test bias or invalidity in language assessment and suggested that there are situations where topical knowledge "may, in fact, be part of the construct the test developer wants to measure" (pp. 120–121). They proposed three ways to account for topical knowledge when defining a construct: "(1) define the construct solely in terms of language ability, excluding topical knowledge from the construct definition; (2) include both topical knowledge and language ability in the construct definition, or (3) define topical knowledge and language ability as separate constructs" (p. 121). Bachman and Palmer (2010) offered the same three options, but option 2 was described slightly differently, as "topical knowledge and language ability defined as a single construct" (p. 218).

Options 1 and 3 assume that topical knowledge and language ability can be separated and either included or not as part of an assessment's construct. In option 2, on the other hand, Bachman and Palmer (1996, 2010) conceded the possibility that both topical knowledge and language ability could be a single construct (the phrasing of option 2 in Bachman & Palmer, 2010) or that at the very least they could overlap. Bachman and Palmer (1996) indicated that option 2 should only be applied when test

takers have homogeneous topical knowledge (thus minimizing its effect on performance), and they warned about inference: "The test developer or user may mistakenly fail to attribute performance on test tasks to topical knowledge as well as to language ability" (p. 124).

Bachman (2007) acknowledged that, even though his (Bachman, 1990) and Bachman and Palmer's (1996) models of communicative language use and task characteristics framework "recognize and discuss language use in terms of interactions between ability, context, and the discourse that is co-constructed, their two frameworks are essentially descriptive" and do not "solve the issue of how abilities and contexts interact, and the degree to which these may mutually affect each other" (p. 55). I would add that the frameworks do not specify how language ability and *topical knowledge* interact or the degree to which they may mutually affect each other. Understanding this relationship has become increasingly important as the field of language education has shifted toward approaches that integrate content and language, and the field of language assessment has expanded its reach to the classroom context.

Critical Issues: Grappling with the Role of Content in Language Assessment Constructs

In this section, I explore how scholars in three different contexts in which language and content intersect—LSP, U.S. K-12 education, and CLIL—have accounted for the role of content in language use and in language assessment constructs. In all of these contexts, content refers specifically to a profession or a particular discipline or subject area in school. For each context, I highlight examples of their latest conceptualizations of the role of content in their assessment constructs.

Language for Specific Purposes Assessment

The field of LSP has the longest history of grappling with the relationship between language proficiency and content in assessment. An outgrowth of the communicative language movement of the 1970s, LSP addresses teaching and learning at the intersection of language and a specific content area, often a professional field (e.g., German for business, Spanish for tourism, English for health professions). LSP assessments address the need to make decisions about individuals' performance on tasks in a specific academic or professional field. To define the construct of what he calls "specific purpose language ability," Douglas (2000) built on Bachman's (1996) model. He defined it as "the interaction between specific purpose background knowledge and language ability, by means of strategic competence engaged by specific purpose input in the form of test method characteristics" (p. 88). Douglas (2000) argued that

"specific purpose background knowledge is a necessary feature of specific purpose language ability and must be taken into account in making inferences on the basis of LSP test performance" (p. 88). This view, however, was not shared by all in the field. For example, Davies (2001) argued that "LSP testing cannot be about testing for subject specific knowledge. It must be about testing the ability to manipulate language functions appropriately in a wide variety of ways" (p. 143).

A special issue in the journal *Language Testing* provides a comprehensive illustration of the tension between these two approaches to defining the construct in LSP assessment. The special issue focuses on the Occupational English Test (OET), a test used to assess the English language skills of overseas-trained health professionals who seek licensure in Australia, New Zealand, and Singapore (Elder, 2016). The OET uses health-related materials or scenarios to assess listening, reading, speaking, and writing. The listening and reading sections are the same for all professions, but the speaking and writing sections differ by occupation. The articles in the special issue describe studies conducted to revise the speaking section of the test, which were motivated by the need to increase its authenticity. The criteria used to score performance on this section include overall communicative effectiveness, fluency, intelligibility, appropriateness of language, and resources of grammar and expression—in other words, criteria that reflect a generalized view of language, consistent with Bachman and Palmer's (1996) option 1 for defining the construct solely in terms of language ability. Many stakeholders (e.g., healthcare professionals), however, did not perceive this approach to be authentic. As Pill (2016) explained, it may be that "the test is not measuring sufficiently those aspects of performance that matter to health professionals in the workplace" (p. 176).

To address this concern, Pill (2016) turned to "indigenous assessment criteria" (Jacoby & McNamara, 1999), that is, assessment criteria derived from the TLU domain. He asked doctors and nurses to provide feedback on test takers' performance on the speaking tasks to help him understand what these health professionals (as opposed to language professionals and educators) value in spoken interactions so he could expand on the more traditional linguistic criteria in their rubric. Based on these professionals' comments, he proposed two new, professionally relevant assessment criteria for the speaking test: clinician engagement and management of interaction.

The next step was to investigate the extent to which the language professionals scoring the assessment could orient to the new criteria. O'Hagan, Pill, & Zhang (2016) explored what happened when seven OET language assessors were trained to apply these new professionally derived criteria when assessing recorded speech samples from previous OET administrations. They found that the new criteria were measuring a slightly different construct of speaking ability, one more consistent with Bachman and Palmer's (1996) option 2 of including both topical knowledge and language ability in the construct definition. The OET, however, is intended to assess only language; healthcare professionals' professional knowledge and skills are assessed by a different test. The studies on the OET speaking section thus raised an important question: Is it possible to separate language from content in an LSP assessment and still have an assessment that yields meaningful interpretations about language use in a specific TLU domain?

Cai and Kunnan (2018) conducted an empirical study to determine whether content and language can in fact be separated in an LSP assessment. Their study investigated the inseparability of content knowledge in an LSP test of nursing English. The test consisted of four texts, each addressing one topic in clinical nursing: gynecological nursing, pediatric nursing, basic nursing, and internal medicine nursing. The goal of the study was to examine whether LSP reading performance could be separated psychometrically from domain-general content knowledge (e.g., nursing) and domain-specific content knowledge (e.g., pediatric nursing). They found that "it is psychometrically possible to separate the portion of domain-specific content knowledge effect from LSP reading score assignment, but this separation is impossible for the portion of domain-general content knowledge contained in the domain-general reading factor" (p. 125). They also called attention to the importance of avoiding a simplistic understanding of content knowledge as an "either-or" paradigm in future research on the separability of content and language.

Knoch and Macqueen (2020) propose an even more nuanced characterization of content and its relation to language use in LSP assessments, specifically those for professional purposes. They suggest that the construct should be determined by sampling from various "codes of relevance" that are part of professional purposes communication. They represent these codes of relevance in the form of four concentric circles (see Fig. 3.1). The interior circle, or the intra-professional register layer, represents the professional register used by a smaller number of users with shared professional knowledge (e.g., doctors who speak to each other in "medicalese"). Language use in this circle is practically inseparable from content knowledge. The next circle is the inter-professional register layer, which represents interactions between individuals with some shared professional knowledge (e.g., a doctor interacting with a nurse or social worker in "cross-disciplinary medicalese"). The next circle, the workplace community repertoire layer, is "a confluence of community varieties with professional register" (p. 63). Interactions in this layer are between those with professional knowledge and lay people (e.g., a doctor communicating with a patient). Finally, the outermost circle represents "the array of varieties used in the broader social context of the target language use domain," including "the standard language/languages of the jurisdiction, minority languages and combinations of languages, e.g. patterns of code switching, as well as lingua francas in use" (p. 63). Knoch and Macqueen argue that this layer is essential because, by attending to it, "policy makers and test developers can see which community varieties could be helpful in contributing to decreased risk of miscommunication in the workplace" (p. 63).

Knoch and Macqueen (2020) explain that decisions about which codes of relevance to sample from when developing a language assessment for professional purposes should be determined through a careful analysis of the professional context and the purpose of the assessment. Their codes of relevance represent the latest conceptualization of language use in LSP and highlight the complexity with which language and content interact in this context.

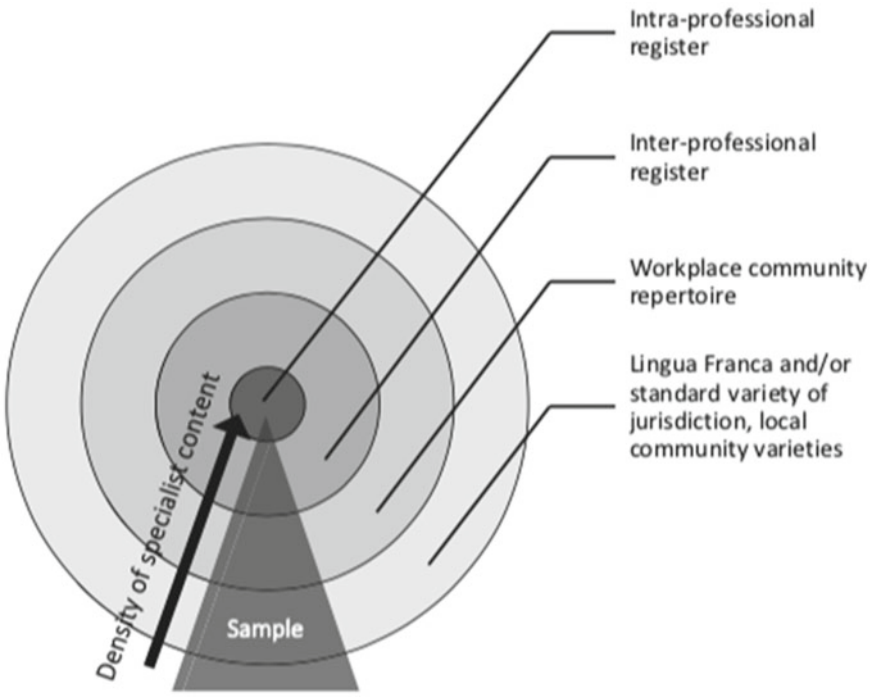

Fig. 3.1 Codes of relevance (Knoch & Macqueen, 2020, p. 61)

Assessment of English Learners in U.S. K-12 Education

Another context that has wrestled with the role of content in language assessment is U.S. K-12 education. Students who are classified as English learners are assessed every year to determine their English language proficiency (ELP) and their content learning (e.g., math, science). As Llosa (2016) explains, content in this context traditionally has been considered a source of construct-irrelevant variance in language assessments, with most assessments adhering to Bachman and Palmer's (1996) option 1—defining the construct solely in terms of language ability. However, it has become clear over time that, to yield valid inferences about students' ability to use English in school, ELP standards and assessments must focus specifically on the types of language used in school, not on general language proficiency (Bailey & Butler, 2003). ELP assessments currently in use are based on ELP standards that link language proficiency to the content areas (language arts, mathematics, science, and social studies). In fact, federal legislation requires that states adopt ELP standards that align with content standards (U.S. Department of Education, 2015). Despite being aligned with content areas, the ELP construct of most of these assessments is operationalized according to the features of academic language at the word, sentence, and discourse level (e.g., see WIDA Consortium, 2012). As Llosa and Grapin (2019) explain, this

operationalization allows ELP assessments to comply with accountability requirements to assess language separate from content and across all content areas at the same time.

However, despite the fact that ELP assessments focus on academic language and avoid assessing content, evidence suggests that the separation between the two may be difficult to achieve, especially at higher levels of performance. Romhild, Kenyon, and MacGregor (2011) investigated the extent to which ACCESS for ELLs, an ELP assessment used in 40 U.S. states (WIDA, n.d.), assessed domain-general linguistic knowledge (i.e., academic language common to various content areas) versus domain-specific knowledge (i.e., academic language specific to a particular content area). They found that the test in most forms primarily tapped into the domain-general factor, but in forms assessing higher levels of English proficiency, the domain-specific factor was stronger than the domain-general factor. Their study indicates that, even in an assessment specifically designed to assess English language proficiency, it is difficult to disentangle language from content at higher levels of English proficiency.

The latest wave of content standards in the U.S. has created an even greater overlap between language and content. The Common Core State Standards for English language arts and mathematics (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010a, 2010b) and the Next Generation Science Standards (NGSS Lead States, 2013) emphasize disciplinary practices. The Next Generation Science Standards, for example, shifted the focus of science learning from learning discrete facts to engaging in the disciplinary practices of scientists, such as arguing from evidence and constructing explanations. In this latest wave of content standards, "engaging in disciplinary practices is not simply a language skill needed to do the work of the content areas; it *is* the work of the content areas" (Llosa & Grapin, 2019). When assessing students in science means assessing their ability to argue from evidence, for example, it becomes even more difficult to separate language from content. Bachman (2002) had already identified the challenge of separating content from language in performance assessment tasks in education and had argued that "performance assessment tasks need to be based on construct definitions that include both content knowledge and language ability" (p. 16), in other words, option 2. And yet, high-stakes assessments used for accountability purposes to this day are tasked with separately assessing English learners' language and content proficiency.

In the classroom, however, the constraints imposed by accountability need not apply, yet teachers tend to adopt the same definition of ELP in terms of academic language in the content classroom. Llosa and Grapin (2019) argue that the construct of academic language at the word, sentence, and discourse level may not be a helpful way to think about English learners' ability to use language in the content classroom because it focuses teachers' attention only on *how* students communicate and not on *what* they communicate. Llosa and Grapin (2019) offer an alternative—a reconceptualization of the ELP construct that leverages the overlap between content and language for the purpose of supporting English learners in the content classroom. In this reconceptualization, the overlap between language and content is represented

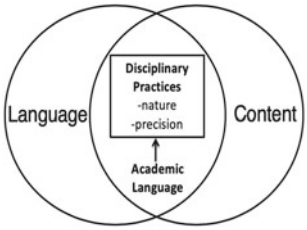

Fig. 3.2 Reconceptualization of the ELP construct for supporting ELs in the content classroom (Llosa & Grapin, 2019)

by the disciplinary practices, described in terms of (a) the nature of the disciplinary practices and (b) the precision of the disciplinary meaning communicated through the practices (Grapin, Llosa, Haas, Goggins, & Lee, 2019). As shown in Fig. 3.2, the linguistic features of academic language at the word, sentence, and discourse levels are relevant only to the extent that they contribute to communicating the intended disciplinary meaning through the disciplinary practice.

Llosa and Grapin (2019) argue that, by focusing more narrowly on the language needed to do the work of the content areas, language and content teachers can support English learners' content understanding and also help them develop the aspects of language that are most crucial to engaging in content learning. This reconceptualization of the ELP construct for the content classroom reflects Bachman and Palmer's (2010) option 2, in which "language ability and topical knowledge are defined as a single construct" (p. 218).

Content and Language Integrated Learning

CLIL is an approach to education in which academic content and a second or additional language are taught and learned simultaneously (Coyle, Hood, & Marsh, 2010). In most CLIL contexts, the additional language taught alongside content is English. Over the past several decades, CLIL has expanded from Europe to other parts of the world. It initially was implemented in secondary schools but is now the pedagogical approach used by many English-medium institutions around the world, and it has expanded to elementary education in some countries. An interesting characteristic of the field of CLIL is that, given the variety of contexts in which it is implemented, it is not (yet) subject to mandated, high-stakes assessments, and most of the research in CLIL has focused primarily on the classroom context.

Until recently, the relationship between content and language was not an area of interest in CLIL assessment. As Wilkinson, Zegers, and van Leeuwen (2006) asserted, "the fact that education takes place through a language that is not the students' mother tongue (and, in many cases, not that of the educators either) seems to have little influence on the assessment processes" (p. 30). They noted that the primary approach was to assess students as they would be assessed in a content area course in their first language. Dalton-Puffer (2013) explained that, even though CLIL

has "a dual focus on content and language," its implementation has been "driven by the logic of the content-subjects," and attention given to language in these spaces has been limited to vocabulary (p. 219).

More recently, however, significant efforts have been made to conceptualize the nature of content and language integration in CLIL (see Nikula, Dafouz, Moore, & Smit, 2016). Without a mandate to assess content and language separately (like those the LSP and the U.S. K-12 contexts are subject to), CLIL scholars have been able to focus on "how students' language can be addressed in a way which does not separate the language used from the content it expresses" (Llinares, Morton, & Whittaker, 2012, p. 187).

Recognizing that content and language teachers tend to orient to different learning goals, Dalton-Puffer (2013) identified "a zone of convergence between content and language pedagogies" (p. 216). Drawing from theories in education and applied linguistics, she proposed cognitive discourse functions (also referred to as academic language functions) as a transdisciplinary construct that captures integration in CLIL. Based on a review of the literature, Dalton-Puffer proposed seven cognitive discourse functions that subsume most communicative intentions: classify, define, describe, evaluate, explain, explore, and report. She views these cognitive discourse functions as a construct that both applied linguists and content specialists can use to inform research and development on the integration of content and language pedagogies "by making visible how transdisciplinary thought processes are handled in classroom talk" (p. 232). She claimed that, beyond its use as a research heuristic, the cognitive discourse function construct could also "function as a kind of lingua franca that may enable [content and language] educators to communicate across subject boundaries" (p. 242). Her conceptualization of content and language integration could also inform assessment constructs consistent with Bachman and Palmer (2010)'s option 2, defining topical knowledge and language ability as a single construct.

Lamenting the traditional lack of attention to language in many CLIL classrooms, Llinares et al. (2012) proposed a scale that integrates content goals with the language needed to accomplish those goals. They argued that the starting point of instruction and assessment in the CLIL classroom should be the content area. They also argued that only the language needed in that particular content area should be assessed, not general language proficiency. They proposed a content-language integrated scale with a content dimension and a language dimension. In adapting the rubric for a given CLIL classroom, the content goals at each level of the rubric are identified first. Then the language goals are identified, described in terms of the genres (text types) and registers (grammar and vocabulary) through which students will achieve those content goals at each level. The purpose of the language dimension is to bring the language CLIL learners need to use "into the open as an explicit component of the tasks they do" (p. 284). However, Llinares et al. (2012) also argued that language need not be assessed separately from content when using this rubric. They proposed that the assessment be based on the content dimension and that the language dimension be used for formative assessment purposes only. In other words, they argued that a teacher in a CLIL classroom should attend to language only to provide instructional feedback relevant to the achievement of the content goals. They view language "as an

enabler, something that is an indispensable component in the achievement of learning goals, but not targeted for separate assessment" (p. 296). This perspective is similar to Llosa and Grapin's (2019) conceptualization of English language proficiency in the U.S. K-12 content classroom and provides another example of Bachman and Palmer's option 2 for defining the construct.

Conclusions, Implications, and Future Directions

Bachman (1990) cautioned that, "for both theory and practice, the challenge is to develop tests that reflect current views of language and language use" (p. 297). Thirty years later, the language education landscape has changed and is increasingly promoting views of language and language use that are integrated with content. This change prompts a reexamination of the role of content in language use and in language assessment construct definitions. Bachman and Palmer (1996, 2010) offered us three options for accounting for content in construct definitions. Until recently, many assessments have opted for options 1 and 3, which presume that language and content can be defined as separate constructs and assessed independently of each other. This approach has been motivated in part by external requirements. As outlined in this chapter, the language assessment literature in LSP and U.S. K-12 education has focused primarily on large-scale assessments used for high-stakes purposes (e.g., licensing or certification in LSP, accountability in U.S. K-12) that specifically require language to be assessed separately from content. In these contexts, Bachman and Palmer's (1996) option 1, defining the construct solely in terms of language ability, has resulted in assessments that did not yield meaningful interpretations about language use in the TLU domain and/or were perceived as inauthentic by stakeholders in that domain. The challenge for these fields has been to find a middle ground. The large-scale ELP assessments in U.S. K-12 and many LSP assessments, such as the OET, assess specialized language; the ELP assessments assess the language of schooling across content areas, whereas the OET assesses language proficiency across a broad range of health professions. These assessments have to be specific enough to serve their purpose but not too specific (e.g., just the language of science or English for doctors), or else they cannot be used for their intended purpose. The consensus is that, in these contexts in which content and language intersect, completely separating language from content in assessment is extremely difficult. Test developers need to figure out how much overlap they are comfortable with for such high-stakes assessments.

Recently, attention to classroom assessment has opened up new possibilities for thinking about the role of content in language assessment constructs. In the classroom, where the goal is to support student learning, there is no requirement to deal with content and language separately. In fact, doing so would be both unrealistic and unnecessary. Several scholars have taken on the challenge of rethinking the language construct in ways that reflect language use in a specific TLU domain and coming up with new constructs that integrate language and content in meaningful

ways. In other words, these scholars are exploring what it would look like to truly adopt Bachman and Palmer's (2010) option 2: "Topical knowledge and language knowledge are defined as a single construct." The models proposed by Knoch and McQueen (2020); Llosa and Grapin (2019); Llinares et al. (2012); and Dalton-Puffer (2013) are examples of this effort. In all of these models, the overlap between content and language is leveraged to support students' content and language learning.

Future research could investigate the ways content and language overlap in various contexts. In so doing, future studies would benefit from developing a more nuanced understanding of content, as Cai and Kunnan (2018) point out. Future studies also could attempt to operationalize these integrated constructs of content and language and examine the extent to which assessments based on these constructs actually provide teachers with useful information that supports student learning in the classroom. Specifically, future studies could investigate the extent to which language and content teachers can orient to these new constructs and use them to provide meaningful formative feedback. Another promising direction would be for scholars across these three contexts, which have traditionally operated separately, to come together to explore new ways of thinking about and assessing language at the intersection of language and content. This type of research collaboration will be critical if we are to develop language assessments that yield meaningful interpretations about test takers' ability to use language in specific TLU domains and, as Bachman (1990) advocated, "reflect current views of language and language use" (p. 297).

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*(3), 5–17.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and context in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–72). Ottawa: University of Ottawa Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document* (CSE Technical Report No. 611). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Cai, Y., & Kunnan, A. J. (2018). Examining the inseparability of content knowledge from LSP reading ability: An approach combining bifactor-multidimensional item response theory and structural equation modeling. *Language Assessment Quarterly, 15,* 109–129.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York: Cambridge University Press.

Coyle, D., Hood, P., & Marsh, D. (2010). *Content and language integrated learning*. Cambridge, England: Cambridge University Press.

Dalton-Puffer, C. (2013). A construct of cognitive discourse functions for conceptualising content-language integration in CLIL and multilingual education. *European Journal of Applied Linguistics, 1*(2), 216–253.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18,* 133–147.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.

Elder, C. (2016). Exploring the limits of authenticity in LSP testing: The case of a specific-purpose language test for health professionals. *Language Testing, 33,* 147–152.

Grapin, S. E., Llosa, L., Haas, A., Goggins, M., & Lee, O. (2019). Precision: Toward a meaning-centered view of language use with English learners in the content areas. *Linguistics and Education, 50,* 71–83.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18,* 213–241.

Knoch, U., & Macqueen, S. (2020). *Assessing english for professional purposes*. New York, NY: Routledge.

Llinares, A., Morton, T., & Whittaker, R. (2012). *The roles of language in CLIL*. Cambridge, England: Cambridge University Press.

Llosa, L. (2016). Assessing students' content knowledge and language proficiency. In E. Shohamy & I. Or (Eds.), *Encyclopedia of language and education* (Vol. 7, pp. 3–14). New York: Springer International.

Llosa, L., & Grapin, S. E. (2019, March). *Academic language or disciplinary practices? Reconciling perspectives of language and content educators when assessing English learners' language proficiency in the content classroom.* Paper presented at the Language Testing Research Colloquium (LTRC), Atlanta, GA.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010a). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Author. Retrieved from http://www.corestandards.org/wp-content/uploads/ELA_Standards1.pdf.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010b). *Common Core State Standards for mathematics*. Washington, DC: Author. Retrieved from http://www.corestandards.org/wpcontent/uploads/.

States, N. G. S. S. L. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

Nikula, T., Dafouz, E., Moore, P., & Smit, U. (Eds.). (2016). *Conceptualizing integration in CLIL and multilingual education.* Bristol, England: Multilingual Matters.

O'Hagan, S., Pill, J., & Zhang, Y. (2016). Extending the scope of speaking assessment criteria in a specific-purpose language test: Operationalizing a health professional perspective. *Language Testing, 33*(2), 175–193.

Pill, J. (2016). Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing, 33,* 175–193.

Romhild, A., Kenyon, D., & MacGregor, D. (2011). Exploring domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency. *Language Assessment Quarterly, 8*(3), 213–228.

U.S. Department of Education. (2015). *Every Student Succeeds Act*. Washington, DC: Author. Retrieved from https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf.

WIDA Consortium. (2012). *Amplification of the English language development standards*. Madison: Board of Regents of the University of Wisconsin System. Retrieved from http://www.wida.us/standards/eld.aspx.

WIDA (n.d.). About WIDA. Retrieved from https://wida.wisc.edu/about.

Wilkinson, R., Zegers, V., & van Leeuwen, C. (2006). *Bridging the assessment gap in English-Medium Higher Education.* Maastricht, Netherlands: Maastricht University Language Centre.