CHAPTER 8

Assessments

In this chapter, we will:

- Review theoretical constructs that are important for the fair assessment of emergent bilinguals, specifically:
  - The power of assessments,
  - The difference between language proficiency and content proficiency,
  - The discrepancy between general language performances and language-specific performance,
  - The validity and reliability of tests for emergent bilinguals,
  - The fit of the assessment to the population, and
  - The match of the language of the test to the language practices of the students.
- Identify inequities in assessment practices that arise from:
  - Assessing prematurely and intensely,
  - Establishing arbitrary proficiencies, or
  - Ignoring those who need the most help.
- Consider some alternative practices, including:
  - Observing closely,
  - Assessing authentically and dynamically,
  - Enabling testing accommodations,
  - Disentangling language and content,
  - Assessing in students' home languages, using translations and transadaptations,
  - Assessing bilingually, and
  - Translanguaging in assessment.

One of the key equity issues surrounding the education of emergent bilinguals concerns the ways in which these students are assessed according to national mandates and state accountability systems. Our central question in this chapter is: *Given what we know theoretically and research-wise about assessment of emergent bilinguals, are these students being assessed according to accepted theories and research evidence about language and bilingualism?* The answer to this question, as we will demonstrate here, is an emphatic *no*. As we have previously noted, it has been widely demonstrated that as a result of inadequate high-stakes tests, emergent bilinguals experience more remedial instruction, greater probability of assignment

to lower curriculum tracks, higher dropout rates, poorer graduation rates, and disproportionate referrals to special education classes (Artiles, 1998; Artiles & Ortiz, 2002; Bach, 2020; Bertrand & Marsh, 2021; Cummins, 1984). In fact, given the negative repercussions on graduation eligibility and resource provisions to schools, Menken (2010) argues that no group has been punished more by high-stakes assessments than emergent bilinguals designated as English learners.

Because the English learner subgroup by definition cannot possibly meet the proficiency targets in these high-stakes tests, *all* programs serving emergent bilinguals are often questioned, including those that are conducted exclusively in English. Mandating high-stakes tests in English for all has acted as language policy, accelerating students' immersion in English without the advantage of home language support. Valenzuela (2005) maintains that high-stakes testing in Texas has been the most detrimental policy for Latines and emergent bilinguals and recommends that there be local control over assessment.

We agree with the assertion that all students have to be *included* in every assessment. But there are equity concerns regarding how assessments are currently being conducted and how the data they generate are being used. These equity issues have to do with misunderstandings of theoretical constructs for assessment that disproportionately impact emergent bilinguals. These include (1) the power of assessments, (2) the difference between language proficiency and content proficiency, (3) the discrepancy between general language performance and language-specific performance, (4) the validity and reliability of the tests for emergent bilinguals, (5) the fit of the assessment to the population, and (6) the match of the language of the test to the language practices of the students. This chapter first considers these theoretical constructs and then identifies the shortcomings of some ill-considered practices that surround assessment of emergent bilinguals. As in other chapters, we note that there is a gap between accepted theories regarding assessment for these students and the testing that takes place. We end the chapter by proposing alternative assessment policies and practices that are more effective for these students based on theory and research evidence.

THEORETICAL CONSTRUCTS IN ASSESSMENT

The Power of Assessments

As Foucault (1979) has indicated, assessment can be a way to exercise power and control (see also Shohamy, 2001). Foucault (1979) explains:

> The examination combines the technique of an observing hierarchy and those of normalizing judgement. It is a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through which one differentiates them and judges them. (p. 184)

Since Alfred Binet developed his intelligence testing methods in the early 20th century, tests have been used to label and classify students, often with grave

consequences. For example, the Stanford-Binet test developed by Lewis Terman was used to "prove" that "[Indians, Mexicans, and Blacks] should be segregated in special classes . . ." (Terman, cited in Oakes, 1985, p. 36). The history of assessment has been entangled from the very beginning with racism and linguistic discrimination (Wiley, 1996). Testing is used more often as a vehicle for allocating or denying educational and employment benefits rather than as a means for informing teaching and developing learning.

Bertrand and Marsh (2015) examined the use of testing data over the course of a year in six schools with large populations of African American and Latine students. Their study found that educators and administrators frequently employed data in manners that reinforced deficit perceptions of students' abilities by attributing students' performance to some aspect of the students' identity, such as being categorized as English learners, assigned to special education, or being generally and permanently a "low" student. These deficit-oriented ideas absolve educators from their responsibility, suggesting that if students are predetermined to perform poorly, then the effectiveness of the instruction they receive is irrelevant. While assessment data are supposed to inform and help improve teaching, Kim's (2017) research with secondary-level emergent bilinguals noted that assessment data sets were actually used in ways that disadvantaged these students throughout their educational journey. These students were continuously placed in remedial education and denied opportunities for high-quality curriculum, rendering them as perpetually low-achieving, struggling, and long-term English learners despite their eagerness to succeed in school. Thus, educators have to be extremely mindful of the power of tests and how they can be dangerous and discriminatory if used inappropriately.

Language Proficiency and Content Proficiency

Every assessment is an assessment of language (AERA, APA, & NCME, 2014). Thus, assessment for emergent bilinguals, who are still learning the language in which the test is administered, is not valid unless language is disentangled from content. As we have noted, English used for interpersonal communication is not the same as the more complex use required for academic tasks in English. Gottlieb (2006, 2016) describes how *academic language proficiency* is usually assessed by evaluating the comprehension and use of specialized vocabulary and language patterns in the spoken and written modes, the linguistic complexity of these modes, and the appropriate use of the sound system (phonology), grammatical structure (syntax), and meaning (semantics) of the language. *Content proficiency* refers to whether the student has actually acquired knowledge of the subject matter. When assessments use only English to test emergent bilinguals' content knowledge, both language and content proficiency are entangled.

The Discrepancy Between General Linguistic Performance and Language-Specific Performance

It is also important to keep in mind two dimensions of language performances that we call *general linguistic performance* and *language-specific performance* (García, Johnson, & Seltzer, 2017). General linguistic performance has to do with the ability of students to deploy any of the features in their language repertoire to accomplish language and content-specific tasks. Language-specific performance refers to the ability to deploy only the features considered standard and that correspond to the specific named language demanded of the task (García et al., 2017).

All speakers use features of their language system that go beyond those sanctioned in schools to perform academic tasks. In part, schools teach students to use language for tasks that are different from those they perform outside of school. In schools, for example, students are given practice in making sense of complex texts they have read or listened to; using language to make convincing arguments, especially in writing; discussing a math problem or theorem; and so on. As we have seen, emergent bilinguals come into schools with very different schooling experiences. Some newcomers arrive in U.S. schools with strong prior education. They know how to use language to perform school tasks, although perhaps they may not have the features of English demanded in a U.S. school. Let us illustrate with two hypothetical cases. Jeehyae, newly arrived from Korea, can use Korean to do all the things that the language arts standards require. In reading, and in Korean, Jeehyae can provide text evidence of key ideas, make inferences and identify main ideas and relationships in complex texts, recognize the text's craft and structure (its chronology, comparison and contrast, identify cause and effect), and associate knowledge and ideas from multiple sources and texts. In writing, and in Korean, Jeehyae can produce text types for various purposes, such as opinion, informative, explanatory, and narrative pieces. But, of course, if in an assessment Jeehyae were asked to perform these tasks using English features exclusively, she would do very poorly.

Jeehyae's linguistic performance is different from that of Moussa, a student recently arrived from Guinea in Africa whose home language practices he identifies as Fulani, but who has been schooled in French in a poor rural school. Moussa's school did not provide him with the opportunities to use Fulani or French to perform the linguistically complex academic tasks valued in U.S. schools. His school focused on developing Moussa's standard French through mechanical drills and tasks aimed to "erase" his Fulani. When Moussa entered his U.S. school, he was assessed in English only. As a result, his lack of experience using language for academic purposes was masked by his poor English-only performance.

By not differentiating between general linguistic performance and language-specific performance, assessments run the risk of obviating the important difference between Jeehyae, a student able to use language to perform academic tasks (even if that language is Korean), and Moussa, whose poor language performance is not simply a matter of English language acquisition but of learning how to use language for school tasks. Both students' language-specific performances in English language assessments are poor. However, Jeehyae can leverage her experience with ways of using Korean in school to perform academic tasks, whereas Moussa cannot perform these tasks in French, nor can he do so in Fulani. Assessments that do not differentiate between general linguistic performance and language-specific performance mask an important difference among emergent bilinguals that has

important consequences for their education. These entanglements then have important significance for test validity.

Validity and Reliability for Emergent Bilinguals

In order for test results to be equitable, emergent bilinguals must be included in the design and piloting of the instrument so that the *norming* of the test is not biased—that is, the test must have both validity and reliability for bilingual students (Abedi, 2004; Abedi & Lord, 2001).

Reliability refers to the capacity of the test or of individual test items to measure a construct consistently over time. Research by Abedi (2004); Abedi, Hofstetter, and Lord (2004); and Martiniello (2008) demonstrates that large-scale exams have differential reliability for students whose English language abilities do not match the population on whom the test was normed. Martiniello (2008) shows that emergent bilingual students and monolingual students with the same math ability perform differently on particular math test items because of unfamiliar vocabulary and complex syntactic structures.

Chatterji (2003) defines *validity* as having to do with "the *meaningfulness* of an assessment's results given the particular constructs being tapped, purposes for which the assessment is used and the populations for whom the assessment is intended" (p. 56). Given the fact that language and content are confounded in tests, there are concerns over the validity of standardized assessments for emergent bilinguals because a test may not measure what it intends. Furthermore, tests may have little *content validity* for these students because the performance of emergent bilinguals does not reveal much about their learning (Lachat, 1999). Worse still is the *consequential validity* of these tests for emergent bilinguals—the after-effects and social consequences. Emergent bilinguals often pay a price with regard to how they are taught as a result of these tests, ending up misdirected into remedial educational programs and special education (Cronbach, 1989; Kim, 2017; Messick, 1989).

Because tests are constructed for white, middle-class, monolingual populations, they always contain a built-in content bias. Tests do not always include activities or concepts from the worlds of minoritized students (Mercer, 1989). Nor do they take into account the multilingual abilities of bilingual children (Ascenzi-Moreno & Seltzer, 2021; Shohamy, 2011). These tests designed for monolingual students reflect neither the cultural practices nor the language practices with which emergent bilingual students are familiar.

Fit of Tests to Population

Criterion-referenced assessments, in which each exam is compared to a specific body of knowledge, are distinct from *norm-referenced assessments*, in which each exam is measured against the scores of other students. But criterion-referenced assessments are still not appropriate for testing emergent bilinguals. In criterion-referenced assessments, students are graded according to whether they have met a defined criterion or standard, which determines what students should know and be able to do in various subject areas. But emergent bilinguals, by definition, generally cannot meet the standard of English language proficiency; thus, they are often judged not to be competent with regard to subject knowledge.

Furthermore, because language and subject content are entangled, studies have found that there is a discrepancy between test scores and student performance in the classroom. Katz et al. (2004) studied Spanish-speaking and Chinese-speaking emergent bilinguals in San Francisco and concluded: "Test data suggested that ELL students underperformed academically compared to EO (English-only) students, but ELL students turned out to be high achievers in the classroom context" (p. 56).

It has been shown that *performance-based assessments*, tests that ask students to produce a product such as a portfolio or perform an action, are better for bilingual students because they provide a wider range of opportunities to show what they know and are able to do in both language and content areas (Abedi, 2010; Estrin & Nelson-Barber, 1995; Gottlieb, 2016, 2017; Navarrete & Gustke, 1996). Genishi and Borrego Brainard (1995) say that performance-based assessment "can be oral, written, 'performative,' as in dance, or visual/artistic" (p. 54). Goh (2004) recommends any alternative technique that can determine what a given student really knows or can do—performances, hands-on activities, and portfolios of student work—arguing that assessment must include multiple modalities. Because student problem-solving skills may be documented in different ways, performance-based multimodal assessments are less language dependent than are traditional tests, enabling teachers to better distinguish between language proficiency and content proficiency. So, although performance-based assessments tax teachers' time, they are more appropriate precisely because they require teachers' attention to the details in students' performance.

However, it is important to recognize that although performance-based assessments are a more appropriate and valid way to identify student learning, they are not necessarily neutral or objective instruments because ultimately they rely on teachers' interpretations of the results they yield. Teachers' perceptions of students and underlying ideologies play a significant role in shaping their interpretations of student performance in these assessments. For example, Ascenzi-Moreno and Seltzer's (2021) research on formative reading assessments in two different dual-language bilingual programs found that teachers tended to interpret low reading scores among Latine students differently from those of more affluent white students. While teachers analyzed low scores among white students in nuanced ways and used them to identify skills they wanted to teach, teachers of Latine students viewed their low scores as an indication of lack of literacy in their homes and discussed them as a natural result of being overall struggling learners. Thus, Ascenzi-Moreno and Seltzer argue that educators do not use assessments just to gauge students' knowledge; they may also rely on them to reinforce their preconceived beliefs.

Let us reiterate the major caveat about performance assessments: Because their interpretation relies on the judgment of those scoring the tests (Lachat, 1999), it is crucial that individuals knowledgeable about the linguistic and cultural practices of emergent bilinguals participate in the development of rubrics for scoring student work. In this way, scorers may be able to disentangle academic performance on the

assessment from language proficiency. In addition, it is important to recognize the fallacy of color-blind neutrality and engage in deep reflexivity through the interpretation process.

Matching the Language of the Test to Language Practices

The language practices of both emergent and proficient bilingual students are very different from those of monolingual students. Thus, a test constructed for monolinguals cannot match the language use of bilingual individuals who draw from a language repertoire with many more linguistic features. Bachman (2001) refers to the distinction between the language of the test and actual language practices:

> That there must be a relationship between the language used on tests and that used in "real life" cannot be denied, since if there is no such relationship, our language tests become mere shadows, sterile procedures that may tell us nothing about the very ability we wish to measure. (p. 356)

Valdés and Figueroa (1994) point to the difficulties of testing emergent bilingual students with instruments that have been normed for monolinguals:

> When a bilingual individual confronts a monolingual test, developed by monolingual individuals, and standardized and normed on a monolingual population, both the test taker and the test are asked to do something that they cannot. The bilingual test taker cannot perform like a monolingual. The monolingual test cannot "measure" in the other language. (p. 87)

Clearly, monolingually constructed and administered tests cannot validly measure the complex language practices of bilingual students.

INEQUITABLE ASSESSMENT PRACTICES

The inadequate and inequitable assessment practices for emergent bilinguals are often the product of the deficit thinking and raciolinguistic ideologies that we have identified in this book. Emergent bilinguals continue to be assessed prematurely with high-stakes instruments that confound academic language and content and that do not align with their language practices. In fact, much educational time is taken up testing with invalid instruments. And despite the fact that theory and research support the use of performance-based assessments as more valid for these students, they are rarely used as high-stakes tests.

Assessing Prematurely and Intensely

Researchers contend that the high-stakes testing of the American school population mandated by NCLB (2001) and continued by ESSA (2015) has had a negative

effect on all students (Nichols & Berliner, 2007). Much time and energy is being spent testing, even though there is no evidence that the testing is improving the education of emergent bilinguals. Although teachers are assessing students with more performance-based assessments, they often deem the scores on high-stakes standardized tests to be more important because schools and teachers are held accountable for the performance of students on such tests (Amrein & Berliner, 2002).

The intensity of testing means that less time is being spent in challenging and creative teaching or teaching subject matter that is not tested. The phenomenon known as *washback*, the process by which testing and formal assessments drive the curriculum, has been well documented, especially in the literature on language teaching (Bach, 2020; Cheng, Watanabe, & Curtis, 2004; Shohamy, Donitsa-Schmidt, & Ferman, 1996). A study by the Center on Education Policy (CEP) found that since 2002, 62% of districts reported they had increased the time for reading and required schools to spend a specific amount of time teaching reading, while 53% of districts had also done so for math (Center on Education Policy, 2005). In contrast, less time is being spent teaching science, social studies, and the arts. Crucially, the reading and math curricula narrowly follow the exigencies of the tests.

The Center on Education Policy (2016) surveyed teachers on a number of issues, assessment being one of them. Eighty-one percent of teachers responding believed that students spend too much time taking state-mandated tests. In high-poverty schools, one-third of teachers said they spend more than one month per school year preparing students for mandated exams. However, an overwhelming majority of teachers (86%) believed in formative teacher-made assessments. For state-mandated tests, less than a third of teachers surveyed wanted to eliminate them, but more than half (60%) wanted to reduce their frequency or length.

In 2015, the Council of the Great City Schools estimated that the average student in large city school systems will take approximately 112 mandatory standardized tests between prekindergarten and high school. Each year, students spend an average of 20 to 25 hours taking these tests.

In the aftermath of the Covid-19 pandemic with its attendant ongoing deficit discourses about "learning loss," test-related pressures have been exacerbated. A 2023 survey by the EdWeek Research Center, involving 870 teachers, principals, and district leaders, showed that nearly half feel heightened pressure after Covid-19 to ensure students perform well on these tests (Stanford, 2023). In addition, 41% of participants reported an increase in the time spent by teachers in their districts on preparing students for standardized tests since the 2018–2019 school year, which was the last full school year before the pandemic.

Concerns about excessive testing are not new. A 2015 national PDK/Gallup poll (PDK International, 2015) reported that two-thirds of public-school parents agree that there is too much emphasis on standardized testing. This has fueled the parental movement to opt out of mandated standardized testing. Although ESSA (2015) has continued to require testing, it has also made funds available to states to audit their testing systems and eliminate unnecessary assessments (U.S. Department of Education, 2016). During a speech to educators in January 2023, U.S. Secretary of Education Miguel Cardona noted that "too many generations of students,

particularly Black and [B]rown students, missed out on STEM, hands-on learning, experiential learning, or project-based learning because teaching was reduced to test prep. Enough is enough!" (Long, 2023). He advocated for standardized tests to serve as "a flashlight" illuminating effective educational practices rather than "a hammer" dictating outcomes. However, transforming the culture of accountability that ultimately punishes the most vulnerable students will continue to be a challenge given over 20 years of federal policy mandating annual testing and imposing penalties on underperforming schools.

Establishing Arbitrary Proficiencies

Assessment of emergent bilinguals is done particularly to determine whether or not they are proficient in English. But different states measure English proficiency differently, using tests that have diverging views of what the construct of proficiency entails. Thus, as we have said, emergent bilinguals may be deemed to be proficient in one state and not in another.

Establishing these arbitrary language proficiencies also stems from a misunderstanding of language development and bilingualism. Proficiency assessments pay attention only to the stage in which English is being learned, as if that process could be completed, thus ignoring that all speakers are constantly adding new features to our linguistic/semiotic repertoire. The point at which the student is declared to be "proficient in English" is established arbitrarily and abstractly. Gándara and Contreras (2009) note that it is a false dichotomy to say that "One is either proficient or not; one is either an English learner or a fluent English speaker" (p. 124), and they argue that this false dichotomy is imposed because of external funding and other pressures to sort students. These assessments consider proficiency only from a monolingual English language point of view, and the emergent bilingual student is considered nothing more than an English learner.

In contrast, as we have said, speakers "do language"; they language actively; they use diverse language features, practices, and additional meaning-making modes. We have argued that the bilingual continuum is not a straight linear process but rather is one that flows unevenly as students' language practices adapt to the changing social and academic contexts and interlocutors that they encounter, as well as their bodily-emotional experiences with language. The dynamic relationship among the linguistic features that they draw on means that, although students can be placed along a bilingual continuum in terms of development of both what is considered the home language and English, there is no end point by which students leave one category and enter into another because they are always translanguaging. Bilingual proficiency is dynamic. Like a river current, what García, Johnson, and Seltzer (2017) have called the translanguaging corriente, it is always flowing and always adapting to the communicative terrain in which it is being performed and the students' own emotional sense of belonging. Emergent bilinguals are somewhere along the starting point of the bilingual continuum, and developmental progress along that continuum is contingent on their opportunities for doing language and their "feel" in doing so. However, because bilinguals are constantly selecting features from their language repertoire for very different tasks and diverse interlocutors, we cannot say that they have ever finished "having" one language or the other.

The federal requirement of ESSA that students meet AYP standards in state exams drives the construction of categories of proficiency that have little to do with students' real learning and development. Assessments are then used to categorize students arbitrarily instead of to develop deep understandings and teacher knowledge about students and their learning.

Ignoring Those Who Need the Most Help

Federal educational policy's ongoing emphasis on ensuring that all students are proficient in English and pass standardized tests means that it has become common for schools to spend enormous time and energy with those who are close to moving up a proficiency level, those whom many call "the bubble kids" (Booher-Jennings, 2006). Neal and Schanzenbach's (2010) research shows that although NCLB accountability data reveal decreased disparities in achievement between white and minoritized students, when the data are disaggregated by subgroup, Black and Latine students are still likely to be left behind because schools tend to prioritize bubble kids. Unfortunately, even new strategies intended to prevent manipulation of the system are susceptible to manipulation themselves. Lauen and Gaddis (2012) found that minoritized students who were near the proficiency threshold benefited more than those significantly below or above it, indicating, again, that teachers may have focused more attention and resources on these students. In this way, while subgroup accountability measures may reduce disparities between different groups, they may also exacerbate disparities within those groups. In sum, students at the bottom were getting very little help and students at the top were also not being challenged academically and intellectually, because their positive scores would not make a difference to the schools' AYP measures.

ALTERNATIVE ASSESSMENT PRACTICES

Observing Closely

The best way to assess emergent bilinguals is for teachers to observe and listen to their students and record these observations systematically over long periods of time. Of course, to "re-view" emergent bilinguals in the close ways described by Carini (2000a, p. 56), the observer must be familiar with the linguistic and cultural practices of the student. Carini (2000b) differentiates between assessing what students learned, made, or did, and "paying close attention to *how a child goes about learning or making something*" (p. 9). Carini (2000b) continues: "[I]t is when a teacher can see this process, *the child in motion*, the child engaged in activities meaningful to her, that it is possible for the teacher to gain the insights needed to adjust her or his own approaches to the child accordingly" (p. 9). Thus, for Carini, observing closely refers to understanding the *process of learning*, not just assessing

a product. These ongoing descriptive reviews of children (Carini, 2000a, 2000b; Traugh, 2000) can develop a multidimensional portrait of bilingual learners. Rather than labeling emergent bilinguals as "limited," "at risk," or "deficient," these kinds of assessments provide avenues for understanding the capacities and strengths of emergent bilinguals. Observing closely allows the teacher/assessor to obtain valid, reliable information about the dynamics of the process of learning that then informs the teaching in a cyclical relationship.

Assessing Dynamically and Authentically With Performance Assessments

Dynamic assessment rests on the work of Vygotsky on the interactive nature of cognitive development. Its goal is to "determine the 'size' of the ZPD [zone of proximal development]" (Gutiérrez-Clellen & Peña, 2001, p. 212) and to transform students' abilities through dialogic collaboration between learners and the assessor-teacher (Poehner, 2007). Dynamic assessment and instruction mutually elaborate each other. Dynamic assessment is thus mostly formative; it simultaneously supports emergent bilinguals in the learning of language and content (Alvarez et al., 2014; Bailey & Heritage, 2014; Heritage, Walqui, & Linquanti, 2015).

Georgia García and P. David Pearson (1994), in a review of formal and alternative assessment, support the notion that emergent bilinguals be given performance-based assessments that are dynamic, in the sense that they should reveal what the student can do with or without the help of the teacher or of peers. In this way, teachers are able to evaluate the kind of support that students require in order to comprehend and complete tasks. For García and Pearson, these dynamic assessments can be conducted in English, the home language, or using practices from both languages. Through dynamic assessments that are administered bilingually, teachers may assess their students' interpretations of material and vocabulary from diverse cultural and linguistic perspectives and then use that knowledge to create further opportunities for students to learn what is appropriate. For example, through bilingual dynamic assessments, bilingual students can demonstrate their literacy understandings using their home and school language practices, and teachers can then assist them in developing academic literacy in the additional language (see also García, Johnson, & Seltzer, 2017; García, 2009, Chapter 15).

Gottlieb (2017) provides a succinct guide to assessing multilingual learners using performance assessments. She gives the following characteristics of an effective performance assessment:

- Represents students' identities, languages, and cultures,
- Consists of authentic tasks with real-life application that ideally take on social action,
- Requires hands-on student engagement, preferably in collaboration with peers,
- Exemplifies original student work that includes multiple modalities,
- Is built from features of universal design for learning,

- Connects to students' lives, interests, and experiences, and
- Offers evidence for learning based on standards-referenced criteria for success.

Portfolio assessment is an important part of these performance assessments. In portfolios, emergent bilinguals can collect artifacts of their learning and reflect on what they have learned collaboratively with one another, their teacher, and their families.

Gottlieb (2017) points to other tools that can be used in performance assessments: anecdotal notes taken by the teacher, student self-assessment, peer assessment, reading response logs, read-alouds, using translanguaging together with other meaning-making modes and the resources offered through technology (see Chapter 9). All of these are valid gauges of emergent bilingual students' academic progress, taking into account the difference between language and content proficiencies, and between general linguistic performances and language-specific performances.

Enabling Testing Accommodations

One way to improve the validity of monolingual standardized assessments is to provide students with test accommodations. Many educational authorities provide accommodations for emergent bilingual students being tested in English. As a result of the accountability-system changes associated with NCLB and continued under ESSA, emergent bilinguals must be provided with appropriate accommodations. ESSA (2015) states that such accommodations should include, "to the extent practicable, assessments in the language and form most likely to yield accurate information on what those students know and can do" (p. 129, Stat. 1826).

State accommodation policies vary substantially (Rivera & Collum, 2006). Rivera and Stansfield (2001) organize the different accommodations into five categories:

1. *Presentation:* Permits repetition, explanation, simplification, test translations into students' native languages, or test administration by a bilingual specialist
2. *Response:* Allows a student to dictate their answers, and to respond in their native language or display knowledge using alternative forms of representation
3. *Setting:* Includes individual or small-group administration of the test, or administration in a separate location or multiple testing sessions
4. *Timing/scheduling:* Allows for additional time to complete the test or extra breaks during administration
5. *Reinforcement:* Allows for the use of dictionaries and glossaries

Abedi and Lord (2001) show how *linguistic modification*, the paraphrasing of test items so that they are less complex, has resulted in significant differences in math performance among emergent bilinguals in the United States. In fact, additional

research has shown that *the only accommodation that narrows the gap between emergent bilinguals and other students is linguistic modification of questions that have excessive language demands* (Abedi & Lord, 2001; Abedi et al., 2004). Test accommodations other than linguistic modifications seem to make little difference in the scores of emergent bilinguals. However, linguistic accommodations tailored to meet the language needs of emergent bilinguals, such as glossaries/dictionaries and linguistic modifications, tend to be less commonly integrated into assessments conducted in mainstream classrooms (Yang, 2020). Because teacher beliefs and knowledge play a significant role in their assessment accommodation practices, it is important to design professional development that deliberately supports teachers in this area (Ascenzi-Moreno & Seltzer, 2021; Cho, 2019; Yang, 2020).

Disentangling Language and Content

Content and language are distinct areas of learning, but they are also interconnected, suggesting that language proficiency is necessary to convey content knowledge. Notwithstanding, Mahoney (2017) stresses the importance for educators of emergent bilinguals to consistently clarify the purpose of their assessments in order to differentiate the relevant construct. Although this disentanglement is difficult, some alternatives have been proposed. Shepard (1996) has argued that a fair assessment framework for emergent bilinguals should integrate the two dimensions—*language proficiency* and *content proficiency*. Academic performance of bilinguals should be seen as a continuum that is related to a continuum of English acquisition; in this view, the language in the tests of subject-matter content is adapted according to the place along the continuum in which the student might be situated.

Duverger (2005) has suggested that another way of disentangling the effects of language proficiency on content proficiency is to have a double scale of criteria: criteria relating to the *content* being delivered and criteria relating to the *language* being used. When content learning takes place through a student's weaker language, which is English in the case of emergent bilinguals in the United States who are the major focus in this book, subject-matter knowledge should have a higher coefficient, and language should not mask satisfactory handling of the content.

Assessing Students in Their Home Language

A much more equitable practice would be to assess students in their home languages. In this section, we address why this does not happen and the ideological and technical obstacles that stand in the way. The current debate over assessment of emergent bilinguals in the United States largely results from the lack of clarity over the goal of their education, which is whether educating emergent bilinguals should focus only on English language development or more broadly on their intellectual, academic, and social development. The education of English language learners—as the naming indicates and as we have been signaling—often focuses narrowly on the acquisition of *English language skills* and not on the acquisition of content knowledge. Houser (1995) explains that if content-area knowledge were the primary goal,

students would be assessed in their home language. However, assessing students in their home language is generally considered inappropriate because educational policy attends narrowly to fluency in English. Furthermore, developing translations and transadaptations of assessments in the home language is complicated, as we will see in the next two sections.

Test translations. Effective translation of tests requires collaboration among various stakeholders, including translators, ESL/bilingual teachers, content teachers, test developers, and sociolinguists (Solano-Flores, 2012; Turkan & Oliveri, 2014). Given how challenging this is, test translation is not always feasible or appropriate. Because the testing industry is a market-driven operation, it might be possible to develop translations into Spanish, since the numbers of Spanish-speaking emergent bilinguals would merit it, but developing them into less commonly used languages could be difficult and expensive. Furthermore, assessments conducted in different languages may not be psychometrically equivalent (Anderson et al., 1996). Maintaining construct equivalence is difficult when the test is either translated directly from one language to another or when tests in two languages are constructed. The nonequivalence of vocabulary difficulty between languages makes comparisons for content proficiency between tests given in different languages inappropriate (August & Hakuta, 1998). The Standards for Educational and Psychological Testing put forth conjointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1985) state:

> Psychometric properties cannot be assumed to be comparable across languages or dialects. Many words have different frequency rates or difficulty levels in different languages or dialects. Therefore, words in two languages that appear to be close in meaning may differ radically in other ways important for the test use intended. Additionally, test content may be inappropriate in a translated version. (p. 73)

Sometimes, emergent bilinguals are allowed to use both the home language version and the English language version of the test. But developing and validating equivalent versions of a test (two monolingual versions side by side) is difficult and costly (Anderson, Jenkins, & Miller, 1996). Furthermore, research on this issue has repeatedly shown substantial psychometric discrepancies in students' performance on the same test items across languages (August & Hakuta, 1997). This means that test items are not measuring the same underlying knowledge.

In addition, translations are only viable when emergent bilinguals have been effectively educated in their home languages. And even then, translations may privilege the standard variety of the language, which often is different from language use in bilingual contexts (Solano-Flores, 2008). If students have limited literacy in the standard variety, a translated assessment that tests content proficiency is also invalid.

An additional concern is that even appropriate translation of tests would not obviate the cultural uniqueness and register of high-stakes tests, which might not be familiar to immigrant students—the format, layout, bubbling in

multiple-choice answers, and discursive styles that are specific to tests (Solano-Flores, 2008), as well as computer-based assessments.

Furthermore, translations are only appropriate if the students have been taught through the language of the test. If the language is not used for instruction, then assessment for content proficiency in the students' home language also may be counterproductive. For instance, Shanmugam and Lan (2013) discovered that Malay-speaking students expressed dissatisfaction with Malay translation assistance because English was their primary language in school. They felt they understood mathematical concepts better in English than in Malay. In short, the language of the assessment must match students' primary language of instruction (Abedi & Lord, 2001; Abedi, Lord, & Plummer, 1997; Yang, 2020).

Transadaption of tests. Transadaptation is a strategy with promise. Transadapted tests adapt test items to fit the cultural or linguistic requirement of the students who are being tested. For example, an English test item may ask about "mountain climbing," but the Spanish test may adapt this to say "*excursión*," a more relevant cultural experience. Transadapted test items are not simple translations. They are written from the linguistic and cultural perspective of the group being tested. Transadapted tests work to eliminate cultural biases, which are prevalent in many assessments because they refer to cultural experiences or historical/social backgrounds to which many emergent bilinguals have not been exposed (Johnston, 1997). As a result, they have more validity than tests that are simply translated. They also are better able to capture content knowledge by minimizing the negative effects of language, particularly when using other accommodations such as combining transadaptation and oral delivery in math assessments (Badham & Furlong, 2022). Transadapted testing is being used in some states—for example, Texas, which uses transadapted Spanish versions of the state's test, called STAAR. However, even transadapted tests, because they are monolingual, do not take into account the full range of linguistic practices of emergent bilingual students, whose full capabilities are enmeshed with their bilingualism.

Assessing Bilingually

A valid way to assess the content proficiency of emergent bilinguals (not solely their English proficiency) is to develop large-scale assessments that build on their bilingual abilities. One example of a bilingual mode instrument designed to assess the verbal-cognitive skills of bilingual students is the Bilingual Verbal Ability Test. All subtests of the assessment are first administered in English. Any item failed is then readministered in the student's home language and the score added in order to measure the test-taker's knowledge and reasoning ability using both languages (Muñoz-Sandoval et al., 1998).

Students also could be assessed via a *bilingual mode*, a way of rendering learners' bilingual abilities and knowledge visible. For example, spoken questions can be given in one language and responses requested in another. Or written tests can have the question in English, and the responses may be produced in the home language, or vice versa. Alternatively, the written text could be supplied in English and the oral

presentation in the home language, or vice versa, thereby providing the teacher with a measure of students' productive skills across two modes and for all languages, while at the same time giving emergent bilinguals the opportunity to use all their languages. Thus, this bilingual mode of assessment would not only give educators a more accurate picture of what students really know without having language as an intervening variable, but it also would offer a clearer picture of students' language capacities.

In cases where school authorities are interested exclusively in students' progress in learning English, students might also be assessed via a *bilingual tap* mode, a way of tapping into or drawing on their home language in order to produce English (García, 2009). This type of assessment would, for example, give instructions and questions in the students' home languages and ask students to respond solely in English. In this way, the home language would be used to activate knowledge for assessment in much the same way that bilingual children use their home language and culture to make sense of what they know. This bilingual tap assessment builds on work on bilingual language processing by Dufour and Kroll (1995), Jessner (2006), and Kecskes and Martínez Cuenca (2005).

A more holistic way to assess emergent bilinguals is to integrate not only the use of bilingual approaches but also various language modalities, such as reading, writing, listening, and speaking in literacy assessment. For instance, Butvilofsky and colleagues (2021) advocate for a comprehensive biliterate assessment approach that employs writing to more accurately gauge emergent bilingual learners' literacy proficiency, such as in reading fluency and phonological awareness. In their study, they selected students whose scores on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) English assessment identified them as reading below or well below benchmark and in need of intensive interventions. Yet the authors also administered the Indicadores Dinámicos del Éxito en la Lectura (IDEL), a similar test in Spanish and collected three pairs of Spanish and English writing samples. In addition to the IDEL results, the qualitative analysis of the students' writing samples revealed biliteracy strengths across Spanish and English as well as their capacity to exhibit progress throughout their second-grade academic year that DIBELS was unable to capture. This research underscores the importance of biliterate writing assessment in comprehensively evaluating emerging bilingual learners' literacy development, offering a broader and more suitable means of assessing their literacy skills along their linguistic repertoire.

Translanguaging in Assessment

Leveraging translanguaging in assessment is the only way to level the playing field between bilingual and monolingual students, giving bilingual students the opportunity that monolingual children have—to be able to show what they know and can do using *all their linguistic resources*. Teachers who leverage translanguaging in instruction can design formative and summative assessments in which students are given the opportunity to show what they know by using their full language and communicative repertoire (Ascenzi-Moreno & Seltzer, 2021; García, Johnson, & Seltzer, 2017; Gottlieb, 2017; Mahoney, 2017).

A translanguaging design for assessments ensures that students can perform using all the features of their language repertoire; at the same time, the design tracks whether students are performing independently or with moderate assistance from other people or other resources. For just these purposes, García, Johnson, and Seltzer (2017) have developed a *Teacher's Assessment Tool for Translanguaging Classrooms*. Teachers who are bilingual can apply such a tool readily. The challenge for monolingual teachers, however, is greater, because they have to rely on other individuals (peers, parents, or other colleagues) and other resources for help in interpreting students' test responses. Even so, technology is making this easier, and electronic translations are an indispensable resource for teachers today.

Translanguaging in standardized assessments is still rare, although technology is enabling more possibilities: Alexis López and his associates at ETS have been developing middle school math and science translanguaging assessments for Latine bilingual students. Students are first asked to select a bilingual avatar who is a bilingual friend or assistant and who provides a model of how to translanguage in assessments. Next, on this computer-based platform, students have the opportunity to see or hear items in two languages, English and Spanish. This way, they can say or write responses using all their linguistic resources, either oral or in writing (López, Turkan, & Guzmán-Orth, 2017). The intent is to ensure that language and content, as well as general language and language-specific capacities, are disentangled (López, 2023b; López, Guzmán-Orth, & Turkan, 2014). Additionally, these test innovations are exploring how to enable multilingual practices in content and English language proficiency assessments (López, 2023a).

EDUCATING EMERGENT BILINGUALS: ACCOUNTING FOR FAIR ASSESSMENT

The data-driven frenzy of accountability that was first spurred by NCLB (2001) puts assessments rather than teaching at the center of education today. The issue of assessment is particularly important for emergent bilinguals, for whom some of these high-stakes tests are invalid. There are, however, ways of improving the construction of valid assessments for this population, and this chapter has considered some of these alternative practices.

In sum, we see that misunderstandings about bilingualism in the United States create obstacles to developing assessment mechanisms that are fair and valid for all students. Monolingual high-stakes tests administered to emergent bilinguals have negative consequences not only for the individual students but also for the teachers who teach them, the leaders of the schools in which they are educated, the communities in which they live, and the states in which they reside. Scores on assessments are not only driving the kinds of instruction and programmatic opportunities that emergent bilinguals can access but also the salaries that their teachers receive, the funding that their schools and the states in which they reside obtain, and the real-estate value of the communities in which they live. This creates a cycle in which the victims will surely be the emergent bilinguals, as these testing practices will result

in students dropping out of the school system, excluding them from all opportunities to learn. Developing fair and valid assessments for emergent bilinguals emerges as the most critical issue in education during this era of increased accountability. Continuing down the path we are on has the potential not only to exclude children from educational opportunities but also to undermine the entire public school system on which our U.S. democracy rests.

Study Questions

1. Why are assessments so powerful?
2. Why are assessments not always valid and reliable for emergent bilinguals?
3. How do the different kinds of tests compare in the way they assess emergent bilinguals' language and content proficiencies?
4. What are some of the complications of establishing categories of proficiencies?
5. What is the difference between observing closely and assessing students? Discuss advantages and disadvantages.
6. Describe some of the test accommodations that can be used with emergent bilinguals. Identify issues to consider and practices that work best.
7. What are the issues that surround translations and transadaptions of tests?
8. Discuss ways of assessing bilingually. In your view, what are the difficulties with implementation of such assessments?
9. What does accounting for translanguaging in assessment entail? What are the benefits and the challenges of translanguaging in assessment?