15

Assessment to Inform Teaching and Learning

Susan M. Brookhart
Duquesne University (Emerita)

Charlie DePascale
Independent Consultant

Brookhart, S. M., & DePascale, C. (2025). Assessment to Inform Teaching and Learning. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 1017–1111). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0015

In recent years, at least from the advent of the standards movement in the 1980s (Porter, 1993), expectations have risen that all assessments—not just those designed to be used formatively—will inform teaching and learning. This trend defines the overarching theme of this chapter: Educators, students and parents, and community members increasingly look to many different assessments for information about improving learning; that is, they look to assessments to serve a formative purpose. This chapter describes the variety of assessments that stakeholders press into service to inform teaching and learning, evaluates their use in this endeavor, and suggests what the future may bring as the arc of assessment continues to bend toward the formative.

CHAPTER PURPOSE AND SCOPE

Both *teaching* and *learning* have multiple definitions, depending on one's philosophical and psychological approach (Gitomer & Bell, 2016). This chapter focuses on assessment of what students know and can do in relation to learning goals that are taught and learned in school classrooms, with the intent of (a) informing teacher instructional planning and instructional moves and (b) informing students' thinking and understanding during classroom lessons, including the individual and social regulation of learning. Similarly, *assessment* has multiple definitions, with some emphasizing assessment instruments and others emphasizing assessment processes. This chapter reviews research about both assessment instruments and the processes in which they are used. As Bennett (2011, p. 7) noted, "process cannot somehow rescue unsuitable instrumentation, nor can instrumentation save an unsuitable process." Much of the research reviewed is from the United States, where teaching to learning goals is the common model of instruction; however, learning-oriented assessment is also a growing interest globally (Carless, 2007; Zeng et al., 2018). The chapter does not focus on the use of assessment information to inform curriculum revision, materials review, or policy adoption.

THESIS

Both educators and the public increasingly expect assessment to be formative and to be helpful to the learning of individuals. Since at least the 1990s, assessment users have been clamoring for more formative information, even from assessments not designed to provide such information, like grades or state accountability tests, as evidenced by educators trying to use annual accountability assessments for instructional change and by the generally unsuccessful attempts to add diagnostic scoring to large-scale tests. The logic is often some version of an argument that runs, in general, that if an assessment tells something about what has been learned, it also should have implications for what should be learned next. Growing interest in student learning theory, fairness, and social justice may be influencing this movement toward the formative as well (see Ercikan & Solano-Flores; Lane & Marion; Rodriguez & Thurlow; and Zwick, this volume). This chapter reviews a range of assessments with an emphasis on how they are being used to inform teaching and learning, whether they were originally intended for formative use or not.

Three subthemes emerge in this press toward the formative. First, the nature of the information is broad. Depending on its source and purpose, assessment that informs teaching and learning can result in information that is qualitative (e.g., feedback comments) or quantitative (e.g., scores, grades), or sometimes both (e.g., a rubric level associated with a performance-level description). Information can be about individuals or groups: Assessment that serves a formative purpose can support students working through a curriculum, and it can also help support broader decisions about instruction. Second, both information quality—the validity or trustworthiness of the assessment information—and the quality of implementation in practice are key to the effectiveness of assessment in informing teaching and learning (Bennett, 2011). Third, teacher and student assessment literacy is key to the effective use of assessment to inform teaching and learning.

ORGANIZATION OF THE CHAPTER

The chapter is organized into five sections, the first of which is this introduction. The second section describes assessments of students' current learning status: those focused on the formative, including classroom formative assessment and interim assessment, and those focused on the summative, including grading and state summative assessment. Evidence shows that even assessments intended for summative purposes are used to inform teaching and learning. The third section describes methods of assessment that assess trajectories of learning through the learning process. It begins by describing how the assessments in this section differ from conventional assessments and then describes assessments focused on trajectories of learning: curriculum-based measurement, student learning objectives, and assessment based on learning progressions. In both the second and the third sections, assessments are described according to their purpose and uses, design, measurement considerations, implementation in practice, and their impact on learning. The chapter ends with a fourth section on future directions for assessment to inform teaching and learning and a fifth, concluding section.

ASSESSMENTS OF STUDENTS' CURRENT LEARNING STATUS

More and more, educators, parents, and community members seek to leverage any available assessment information to inform teaching and learning, whether from assessments originally focused on the formative or from those originally focused on the summative.

Assessments Focused on the Formative

Assessment that has as its primary purpose informing teaching and learning is called formative assessment, and it is typically distinguished from summative assessment, which has as its primary purpose certifying or reporting learning. Bloom et al. (1971)

moved formative evaluation, originally conceived at the curriculum level, into the classroom. They emphasized assessment to inform teachers' instruction. During the 1990s, several scholars demonstrated the importance of students in the evaluation process as well (Black & Wiliam, 1998; Natriello, 1987; T. J. Crooks, 1988).

No one definition of formative assessment has yet taken its place in the field as the only authoritative one. However, definitions do converge on formative assessment as an evidentiary process, the need for student involvement, and the intent to move learning forward—just collecting information about student learning is not sufficient. Current definitions of formative assessment can be grouped into four nested categories, according to whether, to be called formative assessment, a process need only provide information about student thinking; must provide information that is potentially useful; must provide information that is used with the intention of improving student performance; or must provide information that is used and that has a positive effect on student performance (Furtak et al., 2015). The 2008 Council of Chief State School Officers' definition of formative assessment in practice (CCSSO, 2018) is used in the glossary of the *Standards for Educational and Psychological Testing* (the *Standards*; American Educational Research Association [AERA] et al., 2014, p. 219):

> Formative assessment: An assessment process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning with the goal of improving students' achievement of intended instructional outcomes.

However, beyond the glossary entry, the *Standards* do not treat this kind of formative assessment. Other definitions have been noted (e.g., Andrade et al., 2019; CCSSO, 2018).

The term *assessment for learning* is very often used as a synonym for formative assessment. Often, whether one uses assessment for learning or *formative assessment* reflects one's research and practice traditions and cultural background. Some (e.g., Black et al., 2003; James & Pedder, 2006) use assessment for learning as a broad term and reserve use of the term formative assessment for instances when assessment information is actually used to improve learning. Conversely, Stiggins (2005) considered assessment for learning as one of several approaches to formative assessment, in which students are actively involved and "inside the assessment process, watching themselves grow, feeling in control of their success, and believing that continued success is within reach if they keep trying" (pp. 327–328).

For purposes of this chapter, Black and Wiliam's (2009, p. 9) definition of formative assessment will be used:

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited.

Wiliam (2010, pp. 24–25) identified five key points in this definition:

1. “Anyone can be the agent in formative assessment,” teachers, learners, or peers.
2. “The focus of the definition is on decisions,” not intentions.
3. The focus is on “next steps in instruction,” which he defined as “any planful activity intended to create learning” (p. 25), whether by teacher or learner.
4. “The definition is probabilistic” (“likely to be better”) and acknowledges that not every well-founded decision will result in improved learning.
5. “The assessment need not change the planned instruction.” Sometimes, formative assessment evidence affirms the teacher or student’s original instructional plan.

This section describes three kinds of assessments that focus on formative uses. Informing teaching and learning is the main intended purpose of classroom formative assessment (Black & Wiliam, 1998). Informing teaching, and thereby affecting learning, is one of three main purposes for interim assessment and is the primary use claimed by most school administrators (Clune & White, 2008; Dadey & Diggs, 2019). Common formative assessments also aim to inform teaching and are considered separately from interim assessments because they are often teacher developed (Heredia et al., 2016).

Classroom Formative Assessment

Several authors (e.g., Harlen, 2005; Penuel & Shepard, 2016; Ruiz-Primo & Brookhart, 2018; Shavelson et al., 2008) have suggested typologies for classroom formative assessment. Wiliam and Thompson (2008, p. 63) presented the formative assessment framework depicted in Table 15.1.

The framework shows how five key formative assessment strategies are related to the formative assessment process and its agents. The formative assessment process, sometimes called the formative assessment cycle, can be expressed as a question

Table 15.1 Framework Relating Formative Assessment Strategies to the Formative Assessment Process and Its Agents

	Where the Learner Is Going	Where the Learner Is Right Now	How to Get There
Teacher	Clarifying learning intentions and sharing criteria for success	Engineering effective classroom discussions and tasks that elicit evidence of learning	Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	Activating students as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	Activating students as owners of their own learning	

Note. From “Integrating Assessment With Learning: What Will It Take to Make it Work?”, by D. Wiliam and M. Thompson, 2008, in *The Future of Assessment: Shaping Teaching and Learning* (1st ed.), pp. 53–82, C. A. Dwyer (Ed.), Routledge. Copyright 2015 by Taylor & Francis Group. Reprinted with permission.

loop from the point of view of students (Where am I going? Where am I now? How will I get there? [or sometimes How will I close the gap? or Where to next?]) (Assessment Reform Group, 2002; Hattie & Timperley, 2007; Sadler, 1989). This process is a scaffolded, practical presentation of the phases of the regulation of learning (e.g., forethought, performance, and self-reflection; Zimmerman, 2011). The phases overlap; for example, students can use information from any phase to do additional planning (forethought) or to further clarify the learning target in their minds (Where am I going?).

As the framework shows, student involvement is built into formative assessment. Regulation of learning can be taught and can begin as early as preschool. Self-regulation skills are important for all learners, including students with learning disabilities (Butler & Schnellert, 2015). This kind of formative assessment, where students understand what they are meant to learn and regulate their learning in pursuit of these learning intentions, helps students make sense of their own learning (Cizek et al., 2019; James, 2017). Co-regulation of learning (Allal, 2010, 2011; Andrade, 2013; Andrade & Brookhart, 2020; Hadwin et al., 2011) or interactive regulation (Perrenoud, 1998), where regulation of learning is affected by both self and other sources, is invoked when students use external information, for example, feedback from teachers or peers or information from print or electronic sources, which happens in most, if not all, classrooms.

The classroom context affects formative assessment practices. For example, in an evaluative classroom context, where it is not safe for students to be wrong, assessment can feel like quizzing to students because teachers use the information to categorize and judge them (Ames, 1992). In a classroom context more focused on learning than evaluating, where mistakes are seen as opportunities to learn, students are more likely to perceive information, even criticism, as helpful and positive (Brookhart, 2018; Jonsson & Panadero, 2018).

PURPOSE AND USES OF CLASSROOM FORMATIVE ASSESSMENT As Table 15.1 suggests, the central purpose of classroom formative assessment is to inform students' studying and learning and teachers' instruction and is characterized by a series of questions. Student use of information can range from very specific (for example, as students learn to use a capital letter and a period in a sentence, they may use a simple checklist) to more complex (for example, as students learn to organize essays, they may use a rubric to evaluate the presentation of their argument).

The two primary ways teachers use classroom formative assessment are to provide effective feedback about strengths and next steps and to decide on next instructional moves. Therefore, the main goal is not measurement of student achievement of a learning goal; rather, the primary information sought is a description of the status of student understanding (Heritage & Heritage, 2013). What is the student thinking? What is the next thought or experience the student should have? Some descriptive evidence suggests that teachers who are expert in formative assessment use student work to understand how students are thinking rather than how correct their answers are (Davis, 1997; Hattie, 2009; Hattie & Timperley, 2007; Kroog et al., 2014; Minstrell et al., 2010; Otero, 2006). Information about student thinking is useful for supporting targeted next instructional moves or feedback (for example, "You seem to have more trouble when the fractions need simplifying; let's work on that").

DESIGN OF CLASSROOM FORMATIVE ASSESSMENT As Table 15.1 shows, Wiliam and Thompson (2008) described five strategies for formative assessment: (a) clarifying learning targets and criteria for success with students, (b) engineering effective discussions that elicit evidence of student learning, (c) providing feedback that helps improve learning, (d) activating students as resources for one another, and (e) activating students as the owners of their own learning. Table 15.2 gives some examples of tools and tactics for each strategy. Notice that some of these are informal and embedded in instruction and others look like more conventional, and more formal, assessment instruments. What makes these examples of formative assessment is that they are used formatively.

Probably the most noticeable, and the most foundational, change in teaching for many of those who practice formative assessment in the classroom is the explicit sharing of learning goals and the success criteria that will indicate how students are learning. All formative assessment practices depend in some way on that foundation (Brookhart, 2020; Wiliam & Thompson, 2008). For example, feedback from teacher, self, or peers may be based on the criteria. Students' goal setting and self-regulation are based on their nascent understanding of what they are trying to learn. Questioning and other checking-for-understanding strategies monitor student progress toward the learning goal.

Feedback in Formative Assessment While all assessments provide some sort of feedback, feedback has a special place in formative assessment. Some definitions of formative assessment equate it with feedback. Other definitions of formative assessment consider feedback a major part of formative assessment but not the whole of it (Ruiz-Primo & Brookhart, 2018).

Originally, the study of feedback had its roots in behaviorism (Kluger & DeNisi, 1996), and feedback was seen as a type of reinforcement. Kulhavy (1977) critiqued this idea, pointing out that the stimulus–response environment in any learning program is constantly changing, and there is no good reason to believe that the same sequence of events that the feedback is supposed to reinforce will occur again. Further, he pointed out that feedback that is delayed for a day or more typically leads to increases in student retention, and if feedback were acting as a reinforcer, the effect would decay, not increase.

Table 15.2 Examples of Classroom Formative Assessment

Formative Assessment Strategy	Examples in Practice
Clarify learning targets and criteria for success	• Share and unpack learning target statements with students • Share and discuss examples of good work (exemplars) or examples of a range of quality of work with students • Create checklists, rubrics, or other tools that organize criteria and have students use them to monitor their learning • Co-construct criteria with students
Engineer effective discussions and tasks that elicit evidence of student learning	• Use open-ended and strategic teacher questions • Organize student response patterns so they respond to each other rather than to the teacher • Teach students how to ask effective questions • Check for understanding during learning: ○ hand signals ○ whiteboards ○ quick writes ○ concept maps ○ question boxes or boards ○ entrance or exit tickets ○ student response systems with multiple-choice questions ○ quizzes ○ performance assessment ○ structured discussions ○ predict–observe–explain tasks (Shavelson et al., 2008)
Provide feedback that helps improve learning	• Using criteria, comment on strengths and next steps • Written feedback • Oral feedback and dialogue • Video feedback • Give students opportunity to use the feedback
Activate students as resources for one another	• Peer feedback
Activate students as owners of their own learning	• Self-assessment • Reflection at pause points during learning or at the end of an episode of work

Note. Adapated from "Integrating Assessment With Learning: What Will It Take to Make it Work?", by D. Wiliam and M. Thompson, 2008, in *The Future of Assessment: Shaping Teaching and Learning* (1st ed.), pp. 53–82, C. A. Dwyer (Ed.), Routledge.

In recent years, feedback has come to mean "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007, p. 81). Interestingly, early studies of the feedback designed and executed using a behaviorist paradigm showed little to no effectiveness on learning (Harris & Rosenthal, 1985; Kluger & DeNisi, 1996), while more recent studies of feedback designed and executed using more constructivist approaches show much larger effects (Graham et al., 2015; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). However, the effects of feedback are quite variable, and—as for formative assessment in general (Klute et al., 2017)—some are negative (Kluger & DeNisi, 1996). Studies with the highest effect sizes (Hattie & Timperley, 2007, p. 84) "involved students receiving information feedback about a task and how to do it more effectively. Lower effect sizes were related to praise, rewards, and punishment." Finally, it may be as important—or even more important—to study the kinds of response processes feedback engenders as it is to study the effects of feedback on performance (Kluger & DeNisi, 1996; Van der Kleij & Lipnevich, 2021; Winstone et al., 2017).

In other words, some types of feedback are more powerful than others. Outcome feedback, sometimes called knowledge of results or verification (Shute, 2008), is the simplest and most common type of feedback (Butler & Winne, 1995). Outcome feedback tells students whether they were correct or incorrect. Formative assessment can produce this kind of feedback, and it is useful for some purposes, especially for tasks involving recall (Mason & Bruning, 2001). Outcome feedback about correctness coupled with knowledge of the correct response, as, for example, on the back of math fact flash cards or in some computer learning software, can be effective for memory tasks.

In contrast, cognitive feedback, sometimes called elaboration (Shute, 2008), contains information that students can use in their thinking (Butler & Winne, 1995; Hattie & Timperley, 2007; Tunstall & Gipps, 1996). Cognitive feedback helps students interpret the task, interpret their response or response processes, set goals and monitor progress, address particular errors, and give examples or guidance. These are the processes of the formative learning cycle, expressed in the language of self-regulation. This kind of feedback generally has more powerful effects on learning (Hattie & Timperley, 2007; Shute, 2008), although it is important to note that most of the studies in these reviews focused on current performance, not long-term learning (Soderstrom & Bjork, 2015). Teacher pedagogical questioning may be the most effective and immediate elaborated feedback, co-constructed with students in dialog (Heritage & Heritage, 2013). The distinction between outcome and elaborated feedback proves important in computer learning as well (Mason & Bruning, 2001; Van der Kleij et al., 2015). Of course, the possibility of elaborated feedback is related to the cognitive demands of the assessment task; complex tasks typically provide more opportunity for elaborated feedback.

Finally, personal feedback (e.g., "You're so smart!") is typically not effective for learning (Hattie & Timperley, 2007). This feedback does not provide information for improvement and may depress motivation.

Koenka et al. (2019) performed meta-analyses of studies investigating the impact of grades (outcome feedback) versus no feedback and grades versus comments on academic motivation and achievement in elementary and secondary school, respectively. Overall, receiving grades positively influenced achievement but negatively influenced learning, compared with no feedback. Overall, receiving comments as feedback resulted in higher achievement and internal motivation compared with receiving grades. Moderator analyses identified more complex relationships. Receiving grades demonstrated no impact on internal motivation for high achievers, as compared with no feedback, but was associated with lower internal motivation for lower achievers. Task-focused comments led to higher achievement and internal motivation, while global-affective comments (e.g., "Excellent! Keep it up!") resulted in a smaller increase in internal motivation and no difference in performance. In general, these analyses supported the notion that outcome feedback can be helpful; however, for most learning tasks, especially performance tasks that require more than recall, elaborated feedback is more effective (Koenka et al., 2019).

From a measurement perspective, what connects outcome and elaborated feedback is the fact that all types of feedback are—or should be—based on the intended learning for which one seeks information (Brookhart, 2020). The intended learning outcome functions as the construct in assessment to inform teaching and learning (Haertel, 1985). From a learning theory perspective, what connects outcome and elaborated feedback is that both can be informational to students engaged in the self-regulation of learning and to teachers who wish to facilitate student learning in their instruction.

Student Use of Feedback If feedback is to affect learning, students need to make sense of it and use it productively. To do this, students need to appreciate feedback; make judgments about the quality of their own work; manage affect, especially negative emotional responses; and take action that improves their work. The first three of these skills interact, and all three affect students' abilities to take action to improve their work or learning strategies (Carless & Boud, 2018). One important aspect of managing affect involves student beliefs about whether ability is fixed or malleable; the latter belief is known as a growth mindset. Yeager et al. (2019) found that a brief, online growth mindset intervention reduced the prevalence of fixed mindset beliefs among lower achieving ninth graders, increased grade point average in core classes, and was more effective in schools where peer norms were supportive of a growth mindset—an important finding that highlights the importance of the relationship between individual beliefs and the classroom climate.

Winstone et al. (2017, 2019) named this constellation of skills and dispositions learners' *proactive recipience* for feedback. They identified four recipience processes: self-appraisal, assessment literacy, goal setting and self-regulation, and engagement and motivation. Students' use of feedback depends on these processes as well as the characteristics of the feedback message and its content and the characteristics and behavior of the sender and receiver of the message (Winstone et al., 2017).

Jonsson (2013) found five challenges to university students' productive use of feedback. First, feedback needs to be useful. For example, one-liners and personal comments may not provide information students can use. Second, students prefer specific, detailed, individualized feedback, in part because they perceive it as helpful to them (e.g., Ferguson, 2011; Gamlem & Smith, 2013). Third, authoritative feedback may not be helpful, and much feedback in higher education has an authoritative tone. Fourth, many students do not know how to use feedback productively and may use it passively or indirectly (e.g., making a mental note for next time) or not at all, instead of revising and improving current work. Fifth, many students may not understand the feedback they are given, either because the writing is illegible (which Jonsson, 2013, found to be "a surprisingly common problem," p. 69) or because they do not understand the academic vocabulary, concepts, or criteria in the teacher's feedback. Jonsson and Panadero (2018, p. 546) extracted three conditions for productive student use of feedback from the growing literature on proactive recipience: (a) it helps if students perceive that the feedback is useful, (b) students need strategies for using the feedback, and (c) the feedback should be delivered without a grade. Similarly, writing about the K–12 context, Duschl and Gitomer (1997) argued that frequent, diagnostic assessment activities in classrooms can help students develop the thinking, reasoning, and problem-solving skills necessary for developing science learners.

In summary, feedback is integral to the learning process from both teachers' and students' points of view. Formative use of feedback is enhanced by teacher clarity and elaboration and students' understanding and productive use of feedback.

MEASUREMENT CONSIDERATIONS IN CLASSROOM FORMATIVE ASSESSMENT The validity or trustworthiness of information from formative assessment may be evaluated differently depending on the type of assessment. Informal formative assessment is usually based on some form of classroom questioning (e.g., questions used to prompt classroom discourse, for exit tickets or student reflection exercises, and so on). More formal formative assessment is based on a more formal instrument or tool, for example, an embedded performance assessment or quiz at a hinge point in a sequence of lessons.

A fairly robust literature on the quality of classroom questions predates the arrival of formative assessment in common practice (Cotton, 1988). In general, instruction that included classroom questioning during lessons produced higher achievement than instruction without classroom questioning. On average (Cotton, 1988), approximately 60% of questions asked required lower cognitive processes to answer (e.g., recall or simple comprehension), 20% required higher cognitive processes, and 20% were procedural questions. Questions requiring lower cognitive processes were more effective at building factual knowledge and memory (Cotton, 1988). However, for older students, increases in questions requiring higher cognitive processes were related to increases in on-task behavior, length of student responses, number of relevant volunteered contributions, number of student-to-student interactions, student use of complete sentences, speculative thinking on the part of students, and relevant student questions.

More recently, researchers have attended to the quality of classroom discourse, a slightly broader notion than classroom questioning. Allowing students to talk about their thinking surfaces that thinking for interpretation by teachers, peers, and the students themselves (Michaels et al., 2008), so that it can be evidence for future teaching and learning moves. Enabling productive classroom discourse involves posing a question that is relevant to the intended learning goal and extending student discourse with prompts that help students respond to each other's thinking, for example, asking a student to restate someone else's reasoning, add their own reasoning, and explain their thinking.

Ruiz-Primo and Furtak (2006, 2007) found a relationship between the quality of middle school science teachers' informal questioning in formative assessment and the amount of student learning, although their study did not allow conclusions about causality. Students learned more in classrooms where teachers deliberately elicited student responses with the express purpose of using the information to move students closer to learning goals and interpreted and acted on the information accordingly. Classroom discourse also socializes students into the classroom learning community. For example, Forman et al. (2017) showed how the use of classroom discourse in a high school science classroom moved a classroom with more didactic question-and-answer discourse toward the kind of scientific argumentation more closely resembling that of scientific communities.

These kinds of projects do not shed light on the quality of teacher-developed formal formative assessment tools like classroom quizzes and classroom performance assessments that are authored or selected by teachers. Earlier literature, mostly from the 1980s and 1990s, evaluated the quality of teachers' classroom testing (Marso & Pigge, 1993), but the authors are not aware of current studies of the quality of the tests teachers use for formal formative assessment outside research projects. Similarly, despite the fact that calls for teachers to use more assessments of authentic performance are not new (Darling-Hammond et al., 1995; Wiggins, 1998), the authors are not aware of current studies of the quality of teachers' performance assessments, which may trade off broad representation of an achievement construct for depth in one area.

IMPLEMENTATION OF CLASSROOM FORMATIVE ASSESSMENT IN PRACTICE Most definitions of formative assessment emphasize the process of implementation, not the validity or trustworthiness of the information (Wylie, 2008). Teacher inferences from student answers to questions can be key to making on-the-spot decisions to further student learning. Heritage and Heritage (2013, p. 176) described this interactional pedagogical practice as "the epicenter of instruction and assessment." In a formative teaching environment, assessment by a teacher can be likened to hypothesis testing. A teacher can make a tentative inference about student understanding from a question and then revise that hypothesis as new information comes in, leaving much more room for revising inferences than for assessment that produces a static score report.

Teacher Use of Formative Assessment Practices While there are some isolated bright spots, on balance, research on teachers' use of formative assessment information to provide formative feedback shows it is infrequent (Johnson et al., 2019; Schneider & Andrade, 2013; Yan et al., 2021). Some evidence suggests that interactional instructional decisions and follow-through that truly target a student's current level of development and move the student toward deeper understanding for a particular concept are difficult to do and that high-quality practices of this type occur rarely and require rich professional development (Furtak et al., 2008; Heritage & Bailey, 2006; Randel et al., 2016; Wylie & Lyon, 2015).

When teachers shift from a view of teaching that emphasizes what they will present to a view of teaching that emphasizes what students will be trying to learn and what they will do to learn it, the classroom learning climate necessary to support the full benefits of formative assessment can be realized. This shift is a major change, from "instructivist" (Box et al., 2015, p. 972) or "social efficiency" (Shepard, 2000, p. 4) teaching focused on delivering content to learners to teaching that supports students as they construct their own meaning. Many teachers find this shift difficult. Research results about the effectiveness of preservice and in-service teacher education in formative assessment are mixed (Brookhart, 2017). In general, at both preservice and in-service levels, where attention was paid to changing teachers' beliefs about learning and those beliefs did in fact change, there was evidence that teachers improved their formative assessment practices. These successful examples function as proofs of concept that effective formative assessment processes can be taught, but such developments are not common enough to be the rule. Other reviews of literature about teaching formative assessment practices and processes have come to similar conclusions about the necessity for, and difficulty of, teachers moving from a teacher-centered to a student-centered understanding of teaching and learning for formative assessment to be effective (DeLuca, 2012; Otero, 2006).

More than any other type of assessment, effective classroom formative assessment is advanced when teachers know both the principles of effective assessment (for example, how to construct questions and tasks that elicit evidence of learning) and how to observe or score the results. Similarly, effective classroom formative assessment is advanced when students know how to use assessment information to inform their learning. Effective classroom formative assessment also advances with the understanding and support of administrators, especially building principals (Michigan Assessment Consortium, 2017; Schneider & Andrade, 2013).

Even when questions and tasks do elicit evidence of student thinking, it appears that it is difficult for teachers to use the information appropriately (Heritage et al., 2009; Ruiz-Primo & Furtak, 2006, 2007; Ruiz-Primo & Li, 2013; Schneider & Gowan, 2013). Despite the difficulty, professional development can increase teachers' abilities to design tasks and questions that elicit student thinking, interpret the resulting information, and use it to feed students' learning forward (Furtak et al., 2016).

Ruiz-Primo et al. (2015) explored the type of teacher actions resulting from judgments based on on-the-fly, informal formative assessment. They analyzed videos from 20 science and mathematics teachers and found the same patterns in each discipline. Meaningful interactions resulting from on-the-fly formative assessment were observed in 67% of the mathematics instructional tasks observed and 71% of the science tasks, sometimes also coupled with interactions to keep students on task. However, both mathematics and science teachers responded to formative assessment information with verbal feedback more than with instructional moves, and most of these responses occurred in the context of whole-class instruction as opposed to the context of individual work. These are only a few of the response patterns teachers could have at their disposal.

Student Use of Formative Assessment Practices Students' use of formative assessment varies. For example, students need to be taught to be effective peer assessors (Topping, 2013; van Zundert et al., 2010). Classroom and school norms affect students' formative beliefs (Yeager et al., 2019) and therefore their use of formative information. Boekaerts et al. (2006) emphasized that students pursue multiple goals in the classroom simultaneously, including wanting to be entertained, to belong, and to feel safe and valued, in addition to their academic learning goals, and that the instructional climate affects each of these. Similarly, Leighton et al. (2013; Leighton, 2019) proposed that the instructional climate, learners' mental models of emotion and cognition, and students' performance—their use of feedback—are related in predictable ways.

IMPACT OF CLASSROOM FORMATIVE ASSESSMENT ON LEARNING As students develop greater facility in regulating their learning, they become more powerful learners. As Black et al. (2006, p. 126) summarized, "Pupils' learning is more productive if it is reflective, intentional, and collaborative, practices which may not come naturally but which can be taught and can lead to pupils taking responsibility for their learning." Students who understand what they are trying to learn can be more intentional about their learning than those who do not, who simply comply with teacher directions. Looking at examples of work is one way that students come to develop an understanding of what high-quality work looks like and, by inference, the understandings and skills that underlie high-quality work. As Sadler (1989, p. 121) wrote, "In other words, students have to be able to judge the quality of what they are producing and be able to regulate what they are doing during the doing of it."

Increased achievement and increased student self-assessment capabilities result when students understand the criteria for good work (Andrade & Valtcheva, 2009). This has been found in many subject areas and grade levels, for example, primary projects (Higgins et al., 1994), elementary and middle school writing (Andrade et al., 2008, 2010; Coe et al., 2011), middle school mathematics (Ross et al., 2002); middle school special education (E. Lee & Lee, 2009); secondary school social studies (Panadero et al., 2012; Ross & Starling, 2008), and college biology (Hafner & Hafner, 2003).

Reviews of research on rubrics, which are one way to codify and communicate criteria for students, have suggested that one of their main functions is to clarify expectations and quality criteria (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010). It is important that the criteria promote understanding and not a mechanical criteria compliance for a short-term performance boost (Balloo et al., 2018; Gitomer & Duschl, 1996; Sadler, 2014).

Estimating the size of the effect of formative assessment on student learning has proved more difficult than it sounds because formative assessment, as the framework in Table 15.1 shows, encompasses so many different strategies. In addition, fidelity of treatment can be an issue. Meta-analyses of the effects of formative assessment on achievement have shown generally positive effects (Kingston & Nash, 2011; Klute et al., 2017; H. Lee et al., 2020; Wiliam et al., 2004), with some of the effects being negative. Studies of the effects of formative assessment practices in specific disciplines have also found generally positive, but variable, effect sizes (Andrade et al., 2019; Burkhardt & Schoenfeld, 2019; Decristan et al., 2015; Graham et al., 2015; Herman et al., 2015). The variability may reflect the fact that formative assessment studies are not necessarily all investigating the same thing. When studies are screened to focus on a single important aspect of formative assessment, effect sizes can be higher. For example, H. Lee et al. (2020) found that formative assessment interventions that featured student-initiated self-assessment had an effect size of 0.61. Hattie and Timperley (2007) found an average effect across meta-analyses of feedback studies of 0.79.

EVALUATION OF CLASSROOM FORMATIVE ASSESSMENT INFORMING TEACHING AND LEARNING Classroom formative assessment is the type of assessment most clearly associated with informing teaching and learning. While elements of it such as feedback have a long history, classroom formative assessment has made a relatively recent arrival into assessment research. Theory is in development (e.g., Table 15.1) and research agendas have begun.

Classroom formative assessment, especially the more informal types, stretches conventional measurement thinking in terms of the nature and quality of the assessment process and tools. This stretching is necessary to bring the benefits of sound assessment squarely into the learning community, typically a classroom, where learning takes place and where students and teachers can gather and use information to inform teaching and learning as it happens. Each classroom is a unique community of learners based on a give-and-take between a particular teacher or coteachers and their students. Classroom information requirements differ from the information requirements for standardized assessment—where context should not matter—in important ways. Formative assessment foregrounds the role of the student and often invokes contemporary learning theories that view the student as an active participant in assessment and learning. This changes the role of students from test takers to learners. Formative assessment is heavily dependent on the quality of the process, not just the quality of the tools, instruments, or questions, for its effectiveness. Formative assessment works best when both students and teachers know how to use it effectively. While research suggests formative assessment can have more impact on teaching and learning than other types of assessment, these challenges and others must be met, even if gradually, for classroom formative assessment to reach its full potential.

Interim and Benchmark Assessment

Many educators and districts use the terms interim and benchmark assessments as synonyms, and the terms have been used interchangeably in the research literature as well. Sometimes a distinction is made: Interim assessments can be parallel test forms for an external accountability test, covering an entire year's worth of content and administered two or three times during the school year to track student learning and achievement growth; and benchmark assessments can be nonparallel test forms covering a portion of the year's content (e.g., the first report period) and administered at a specified point in the school year and curriculum (Ferrara et al., 2019). This chapter uses the term interim assessments for both.

Many interim assessments are sold by test vendors. Interim assessments can also be sold as item banks or as item banking and testing software that can be populated with locally written items or items from other sources. Burch (2010, p. 147) defined interim assessment technology as "the software that is sold to schools and districts in order to gauge students' progress towards high-stakes summative tests and to comply with the test reporting requirement of the No Child Left Behind Act (NCLB) of 2001." She argued that the upsurge in use of interim assessment technologies represents an intensification of educational privatization brought on by NCLB (Burch, 2006, 2010) and that this is at least partly because of pressure on schools to be "more efficient, more compliant, and more equitable" since NCLB (Burch, 2010, p. 147).

PURPOSE AND USES OF INTERIM ASSESSMENT Interim assessments are used to serve one or more of three general purposes: instructional, predictive, and evaluative (Perie et al., 2009). For example, an instructional purpose might be to identify particular content or standards for remediation or reteaching. A predictive purpose might be to determine each student's likelihood of attaining a proficient score on the end-of-year state accountability test. An evaluative purpose might be to provide information about the effectiveness of an instructional program or some curriculum materials. In practice, most districts use interim assessments for several different purposes simultaneously, most commonly for instructional and evaluative purposes (Dadey & Diggs, 2019; Davidson & Frohbieter, 2011). Even when districts use interim assessment for predictive purposes, the intent of the prediction is to inform instruction before the state accountability test, not simply to make the prediction.

DESIGN OF INTERIM ASSESSMENT As the above definitions imply, most interim assessments were designed primarily to give predictive information. They often use a small sample of items across many standards and at the standard level are most reliable for decisions about groups. However, the logic used by district administrators when they purchase interims is that an investment in gathering interim information will support alignment of teaching and learning in the district, provide practice for state accountability tests, and provide information for teachers on the needs of both current and future students (Clune & White, 2008)—again illustrating that trend to want to add a student-level diagnostic purpose to all assessments. The fact that one assessment would not serve these purposes equally well is not typically considered (Perie et al., 2009); however, it is an important consideration in evaluating research on the effectiveness of interim assessments (Immekus & Atitya, 2016).

Interim assessments may be computer-administered or paper-and-pencil tests and usually result in one or more of the following kinds of scores: scale scores, percentiles, and proficiency-level ratings. The results are meant to be used by teachers and administrators to assist individual students (for grouping, remediation, and so on) or to assist with group planning (to identify strengths and weaknesses in current instruction and plan for future instruction). The intention is that educators will set goals, measure and evaluate progress, and use data to make changes. This process is focused mainly on educator learning and is relatively silent on student learning. The intended connection between interim assessments and student learning comes from the claim that measures are linked to learning standards. The most common responses to interim data are reteaching and review (Penuel & Shepard, 2016).

MEASUREMENT CONSIDERATIONS IN INTERIM ASSESSMENT Some commercially produced interim assessments have technical manuals that present standard kinds of information about reliability, validity, and scaling (e.g., Northwest Evaluation Association, 2019). Some, of course, do not. Many of the more established commercial products are reviewed in the Buros Mental Measurements Yearbook series (https://buros.org/mental-measurements-yearbook). Validation differs depending on the intended use of the interim: whether instructional, evaluative, or predictive (see also Lane & Marion, this volume).

Arguably the predictive use requires the simplest and most readily available validity evidence, to answer the question of how well performance on the interim predicts state accountability test scores. R. S. Brown and Coughlin (2007) investigated the availability and quality of predictive validity information for a selection of interim assessments used in the Mid-Atlantic region. They considered four assessments (the Northwest Evaluation Association's Measures of Academic Progress, MAP; Renaissance Learning's STAR; Study Island; and CTB's Terra Nova). While all four provided concurrent or predictive validity information of some sort, only one, Terra Nova, presented a truly predictive study.

Pereira and Tienken (2012) undertook a study in four middle schools to support the validity of using a computer-based interim assessment tool as recommended by the New Jersey Department of Education. The tool had pre- and posttests in mathematics and language arts. The product was marketed to predict achievement on the state accountability test. Findings indicated that the pretest and posttest had almost identical predictive power, prompting the authors to question the need for a posttest (Pereira & Tienken, 2012, p. 11). Further, Babo et al. (2014) reported that the odds ratios for passing the state test for pretest and posttest scores were close to 1, indicating little predictive power. This finding is consistent with broader finding of the large contribution of general mental ability to assessments of educational achievement (Deary et al., 2007).

IMPLEMENTATION OF INTERIM ASSESSMENT IN PRACTICE Dadey and Diggs (2019) identified and summarized 20 studies of interim assessment use. They found the most frequent reported uses were instructional, usually in the form of broad claims that interim results would be used to modify or improve instruction. All 20 of the studies they reviewed listed instructional purposes, for example, general instructional planning, grouping students, and providing remedial support. Sixteen studies also mentioned evaluative purposes, citing evaluation of teachers, programs, and curriculum and supporting resource allocation at the staff, school, and program levels. Eight studies cited predictive use of interim results and always the same one: predicting student performance on the state accountability test.

The process of gathering interim assessment information and using it to make decisions is usually framed as some sort of data-driven decision-making process (Penuel & Shepard, 2016), often more about management than about learning. Acknowledging this shortcoming yet wishing to harness the formative power that can be gained by gathering relevant data for specific learning needs, Wiliam (2018a, p. 47) called for "decision-driven data collection."

Marsh (2012) described a data use theory of action with five steps: (a) access and collect data; (b) turn data into information by organizing, filtering, and analyzing; (c) turn information into knowledge by combining it with teacher understanding and expertise; (d) apply the knowledge to arrive at a response and action, and (e) assess the effectiveness of the outcomes of that action. This is a very linear expression of what is inherently an interpretive process (Blanc et al., 2010; Farrell & Marsh, 2016). Accordingly, the research studies in their Philadelphia project (Blanc et al., 2010; Christman et al., 2009) reframed the teacher data use process as sense making. They described three types of sense making in which teachers engaged with interim data. They described strategic sense making as the most common; it focused on short-term adequate yearly progress targets—things like identifying students who could, with just a little help, become proficient. Affective sense making, also common, focused on teachers' professional responsibilities and things they could do to motivate students and improve learning. Reflective sense making was least common; it focused on questioning and evaluating one's instructional practices.

The Urban Data project of the Council of the Great City Schools (Faria et al., 2012; Heppen et al., 2010, 2011) situated teacher data use within a broader theory of action with five key dimensions: (a) the district's assessment context, (b) supports for data use, (c) working with data, (d) instructional responses, and (e) improved student achievement. This theory of action considers contexts and supports as foundational to data use and achievement outcomes. Heppen and colleagues (2010) found that the assessment context in schools predicted the other key dimensions and that teacher attention to data was a significant, positive predictor of instructional responses. Bulkley et al. (2010) concluded that interims could serve instructional purposes, but only if strong district supports were in place.

The use of interim data is limited by time constraints, especially lack of time for remediation and the presence of pacing guides for instruction (Abrams et al., 2016), and also by the data culture or context in schools (Goertz et al., 2009; Heppen et al., 2010), as well as teachers' own facility with and attention to data (Faria et al., 2012; Heppen et al., 2010). One study using the web logs behind a district's information dashboard (Tyler & McNamara, 2011) found that, on average, teachers spent 2.3 minutes per week looking at the class roster page of their interim results and 30 seconds per week looking at individual student pages—or less than 3 minutes per week on the two results pages that contained information relevant to instructional adjustments. Another study of the use of interim assessments in an urban charter high school found that the tests, based as they were on grade-level standards, did not provide much useful information on students who were often more than 2 years behind in grade-level achievement. In addition, there were too many separate and varied standards on the tests to help teachers make decisions about what to do, since the students needed assistance in most or all of them. For these reasons, that testing program was abandoned after just over 3 years (Bancroft, 2010). Another study, using randomized controlled trials, found little evidence of changes in teacher practices as a result of having access to an interim assessment tool (Chojnacki et al., 2013).

Several studies describe the instructional decisions teachers report making on the basis of interim assessment results, and these studies also note great variability by teacher, including some teachers who do not use the information. Teachers who do use interim results tend to use it to identify topics to teach and students with needs in those topic areas (Abrams & McMillan, 2013; Abrams et al., 2015; Goertz et al., 2009; Oláh et al., 2010; Shepard, 2010), although these decisions are influenced by teachers' perceptions of the difficulty of particular content and students' past performance (Goertz et al., 2009) and their perceptions of the assessment's design and item quality (Abrams et al., 2015; Farrell & Marsh, 2016). Most teachers have not been found to use interim results to change the way they teach content or to look for underlying reasons students may have developed misunderstandings (Christman et al., 2009; Goertz et al., 2009; Oláh et al., 2010). For example, a study of teachers' use of mathematics interim assessments found teachers mostly used results to group students or reteach procedural knowledge, rather than making sense of students' conceptual understanding (Oláh et al., 2010). Teachers report using interim data to identify students who could potentially move from *Basic* to *Proficient* or *Below Basic* to *Basic* by the time of the state test, to identify skills and concepts to reteach, to identify students with similar areas of weakness or strength, to change some classroom routines (Blanc et al., 2010), and, in mathematics, to locate and diagnose errors (Goertz et al., 2009).

Several studies discussed teachers' use of interim data as part of a larger assessment system that included classroom formative assessment, teacher-developed classroom summative assessments, teacher-developed common formative assessments, interim assessments, and state accountability assessments. Wilkerson et al. (2021) surveyed teachers in Nebraska and found that some teachers did not use data at any level (state summative, interim, or classroom formative), but those who did used data from interim assessments once or twice a month to inform instruction. Farrell and Marsh (2016) found teachers had the most mixed opinions of interim assessment data, which they often found less useful and less trustworthy than other kinds of assessment data. In the span of a district assessment system, data from state accountability tests sometimes influences grouping and instructional decisions at the beginning of the year (Abrams et al., 2016). Interim assessment data, as noted above, is most often used to decide *what* to teach to whom, often looking backward at what content needed reteaching (Abrams et al., 2016; Farrell & Marsh, 2016; Riggan & Oláh, 2011). Classroom formative assessment, in contrast, is used to help get students to explain their thinking (Abrams et al., 2016; Riggan & Oláh, 2011) and support decisions about *how* to teach or reteach going forward. Sometimes teachers looked to classroom formative assessment information to help explain or deepen their understanding of interim assessment results to make instructional decisions (Martone et al., 2018).

To use interim assessments well, teachers and administrators need to be able to read, interpret, and use a wide range of both norm- and criterion-referenced scores, as well as understand the concepts of standard error of measurement and standard error of the mean or other aggregated scores, the limited reliability of change scores, and the like. They also need the skills to communicate assessment results to parents and students in an understandable fashion. Many test vendors provide teacher resource material and professional development as part of their interim programs, but as reported previously, the evidence suggests that the inclusion of these resources does not necessarily lead to deeply conceptual differentiated instruction or to improved student achievement.

IMPACT OF INTERIM ASSESSMENT ON LEARNING A small number of studies have evaluated the effects of using interim assessments on student achievement. Faria et al. (2012) found that teacher data use and student achievement were statistically significantly but weakly correlated in middle school mathematics and elementary reading but not in elementary mathematics or middle school reading. Principal data use was statistically significantly correlated with student achievement only in middle school mathematics.

Three randomized studies were found, each investigating different interim assessments. One looked at the effects of using interim assessments (mCLASS in K–2 and Acuity in 3–8) in the state of Indiana in the 2009–2010 school year on the state ISTEP+ test (Konstantopoulos et al., 2013, Konstantopoulos, Li, Miller, & van der Ploeg, 2016; Konstantopoulos, Miler, van der Ploeg, & Li, 2016). Effects in mathematics were statistically significant in Grades 3–8 (not in K–2); effects were larger in the lower (10th and 25th) quantiles. Effects in reading were smaller and not always significant. The exact effect size depended on the grades included in the analysis and whether the analysis examined intent to treat or treatment on the treated (Konstantopoulos et al., 2013). A second study looked at the effects of a program implemented by the Johns Hopkins Center for Data-Driven Reform in Education, including the use of their 4Sight benchmark tests and associated professional development (Carlson et al., 2011), in 500 schools in seven different states, again using the state tests as the dependent variable. They also found significant weak effects in mathematics but not in reading. A third study (Cordray et al., 2012) looked at the effects of using the Northwest Evaluation Association's MAP program in 32 elementary schools in Illinois in Grades 4 and 5 reading. MAP teachers were not more likely than control teachers to use differentiated instructional practices in their classrooms, and there was no significant effect on reading achievement in either Grade 4 or Grade 5. Another study in Massachusetts, using a matched comparison design, compared the 2001 to 2006 Massachusetts Comprehensive Assessment System (MCAS) scores for high-poverty middle schools that were in a pilot program using benchmark assessment with the scores of schools that were not. The intervention occurred in 2006, when the pilot schools used interim assessment technology software to develop quarterly benchmark tests. There were no differences between treatment and comparison schools, either in 2006 (Henderson et al., 2007) or 1 year later (Henderson et al., 2008).

EVALUATION OF INTERIM ASSESSMENT INFORMING TEACHING AND LEARNING The popularity of interim assessments grew from a perceived need for more frequent information for educational accountability assessment. However, the popularity of interim assessments and the pressure to use them for diagnostic decisions they are not designed to support presents interesting evidence of the movement toward insisting that assessment inform teaching and learning—the trend toward the formative use of assessment information.

On balance, then, two arguments are supported. First, teachers do not make much effective use of data from interim assessments. It does seem that educators' use of interim data can be described by one or more of the posited theories of action for data-driven decision-making. However, these processes alone, at least as used to date, offer only a very partial explanation of what happens as educators use data. Many teachers do not engage in these processes, or they do so with inadequate assessment literacy or with instructional adjustments that are topical or procedural rather than adjustments intended to support students' conceptual change.

Second, there is little evidence that using interim assessments improves student achievement. There are only a small number of studies of the effects of interim testing programs on student achievement. These studies yield only small and intermittent effects or no effects. Studies show teachers use item analysis as a main strategy for decisions about reviewing topics and procedures rather than seeking to understand student thinking, and items are only samples from an achievement domain and not the domain—the achievement standards or learning goal—itself. Therefore, even if teachers did make better use of interim assessment data, it is not likely that their effectiveness would be dramatically improved.

Common Formative Assessments and Other District-Developed Assessments

Some districts design what they call *common formative assessments* to be administered across classes in a specific subject area and grade level at pause points in instruction, for example, after an instructional unit or series of related units. The intention is typically to inform teacher instructional planning.

PURPOSE AND USES OF COMMON FORMATIVE ASSESSMENTS Ainsworth and Viegut (2006, pp. 2–3) defined common formative assessments as "assessments collaboratively designed by a grade-level or department team that are administered to students by each participating teacher periodically throughout the year." Teachers use information from common formative assessments for instructional planning for current students and for future planning for when they teach the same subject matter again.

DESIGN OF COMMON FORMATIVE ASSESSMENTS Common formative assessments are usually, although not always, tests and are sometimes performance assessments created or curated by teachers who teach common curricular material in a building or across a district. All of the literature, both professional and scholarly, consulted by the authors assumed quantitative scoring of some kind: right/wrong for selected-response tests and the use of a rubric, typically arranged as a proficiency scale on the standard, for constructed-response test items and performance tasks.

For Heredia et al. (2016), collaborative teacher design is part of the common formative assessment process. Ainsworth and Viegut (2006) added collaborative scoring. Ideally, the use of a common tool for formative assessment will decrease the variations in teacher formative assessment practices and potentially therefore maximize student learning (Heredia et al., 2016, p. 698). Some test vendors produce assessments they call common formative assessments and market them to school districts, which circumvents the teacher design work and its benefits in terms of developing consensus on both formative assessment (Heredia et al., 2016) and supporting student learning (Furtak, Circi, & Heredia, 2018) and makes these products similar to interim assessments.

Common formative assessments are distinguished from other school-based tests, for example, departmental tests in secondary education, in that their use is intended to inform teacher instructional planning and not student grading. At their best, common formative assessments can inform both instructional planning and teacher professional development (Furtak & Heredia, 2014). Common formative assessments are a cornerstone of the professional learning communities movement (DuFour et al., 2010; Myers, 1996) that has gained momentum in school districts around the United States.

MEASUREMENT CONSIDERATIONS IN COMMON FORMATIVE ASSESSMENTS The professional literature recommends that teachers, typically in grade-level or subject area teams, study collaboratively standards, curriculum documents, and other instantiations of what students should know (e.g., textbooks, examples of student work, or previous student performance data) to establish what they will design their common formative assessments to measure (Ainsworth & Viegut, 2006; DuFour et al., 2010). To date, the authors of this chapter have found no studies of the quality of teacher team–developed common formative assessments. However, the heavy emphasis on teacher teams identifying standards or learning progressions and aligning common formative assessment items and tasks to them, in regard to both content and level of thinking, suggests that the content-related evidence for validity—and a match between the tested and taught curriculum—may be a strength of common formative assessments.

To make sure information from locally created common formative assessments is of high quality and can support instructional decisions, the professional literature simply advises district teams to look at recommendations from assessment experts; released items from other testing programs, including on websites such as that for the National Assessment of Educational Progress (NAEP; https://www.nationsreportcard.gov/nqt/); and examples of performance assessments and rubrics from local or other sources (DuFour et al., 2010). District teams are advised to refine assessments after use according to established standards for evaluating assessments (Ainsworth & Viegut, 2006). Given the documented variability of teachers' assessment practices in other areas (see, for example, the section on "Grading"), it is reasonable to expect that this advice will be followed more in some places than in others and that there will be variability in the quality of common formative assessments.

IMPLEMENTATION OF COMMON FORMATIVE ASSESSMENTS The process of implementing common formative assessments may be quite loose or it may be more structured, depending on how teams in any given school operate. It is typically part of a school improvement process. For example, the Team Learning Process (DuFour et al., 2010, pp. 26–33) calls for a process in six steps: (a) identify essential outcomes all students must learn; (b) create common pacing guides and curriculum maps all teachers will follow; (c) develop multiple common formative assessments; (d) establish a target score for proficiency in each skill on each assessment; (e) administer common formative assessments and analyze results; and (f) celebrate strengths and identify and implement improvement strategies. This is a standards-based approach.

The Formative Assessment Design Cycle (Heredia et al., 2016; Furtak & Heredia, 2014) uses a learning-progressions approach, in five steps: (a) reflect, (b) explore student ideas, (c) design tools, (d) practice using tools, and (e) enact tools. Then the cycle repeats, beginning with reflection; typically, the tool (the common formative assessment) is also revised. Reflection is based on current teaching practices as well as student work. Exploring student work can include looking at aggregated data as well as examples of individual student work, with a focus on student thinking.

When done well, the work involved in aligning content standards with the assessments can serve as professional development for the teachers (Frey & Fisher, 2013). Gallimore et al. (2009) conducted a 5-year quasi-experimental study of teacher professional development using an inquiry protocol and found that teachers' attribution of improved student performance changed from implicating external causes to attributing learning to their own teaching. Part of the protocol involved looking at indicators of student learning, designed by the teachers together.

Assessment literacy requirements for teachers creating tests or performance assessments aligned with standards, to be administered across classes, are high. Teacher assessment literacy rises as part of the process (Furtak et al., 2014, 2016; Heredia et al., 2016), and assessment validity rises because the work is informed by collaboration with teachers and access to their pedagogical content knowledge (Ball et al., 2008). Even with support, however, teacher growth in understanding of learning progressions and the assessment strategies and tools that will elicit evidence about them takes time and involves struggle (Furtak, 2012; Furtak & Heredia, 2014).

IMPACT OF COMMON FORMATIVE ASSESSMENTS ON LEARNING Little published evidence of the impact of common formative assessments on learning was found. Gallimore et al. (2009) reported evidence that when teachers used inquiry and followed a learning problem long enough to understand it, assess it, and address it, student achievement improved. This professional development included the use of some assessments that could be described as similar to common formative assessments. Frey and Fisher (2013) found that the use of common formative assessments helped increase student achievement in writing.

EVALUATION OF COMMON FORMATIVE ASSESSMENTS FOR INFORMING TEACHING AND LEARNING Common formative assessments involve teachers in a collaborative process to check on student progress on learning standards in their curriculum. In principle this is a sound idea, and there is some evidence that when done well, teachers and students benefit. However, there is also evidence that assessment design is difficult for teachers to do well, and because of recommendations in the teacher professional literature it is likely that many common formative assessments of unknown quality exist. One key to effective use of this type of assessment is increased literacy on the part of teachers regarding the construction and interpretation of assessments. Increased literacy may also help teachers select, rather than design, appropriate assessments.

Assessment Focused on the Summative

When assessment informs teaching and learning, it is being used for a formative purpose. Therefore, it may seem that assessments designed for summative purposes should have no place in this chapter. However, there is evidence that both effective teachers (Wiliam et al., 2004) and successful students (Brookhart, 2001) use information available from summative assessment to inform future planning. Moreover, as assessment trends toward the formative, it is more and more common that educators are expected to use all available data sources, including summative, to inform teaching and learning (Farrell & Marsh, 2016; U.S. Department of Education [U.S. DOE], 2009; Wilkerson et al., 2021).

In effect, because learning is ongoing, most K–12 summative assessments fulfill their immediate summative purpose, but in the larger picture serve more as periods at the end of sentences rather than as the end of a story. This chapter will consider the use of two common forms of summative assessment, grading and state accountability testing, for formative purposes.

Grading

Grading refers to the symbols assigned to individual pieces of student work or to composite measures of student performance on student report cards. A long history of research has led those who study measurement to question the reliability and validity of grades and, sometimes, therefore, to dismiss their importance. However, grades loom large in the educational experience of all students and are the activity on which teachers spend most of their assessment time, at least at present. They are perhaps the most universally used assessment to inform teaching and learning (Brookhart et al., 2016).

PURPOSE AND USES OF GRADING Grading has a wide array of purposes and uses. Grades are used to inform teachers, students, and parents/caregivers about teaching and learning at many different levels, for example, to support next steps in instruction or studying for a small amount of instructional content or to make broader decisions about next year's teaching or course taking. Grades are used to rank and sort students for various placement and selection decisions, for example, as part of decisions about placing students with special needs in classes or programs, qualifying students to participate in extracurricular activities, and college admissions. Grades can be used to evaluate teachers, schools, and educational programs. Informally, students and teachers use grades to do all sorts of things they were never meant to do, for example, support classroom behavior management or students' sense of self-worth (Covington, 1992; Thomas & Oldfather, 1997).

A recent grading reform in the United States, called standards-based grading, reports student achievement relative to content standards established for a particular grade level. Reformers claim this gives students, parents, and educators more useful feedback (Peters et al., 2017). In other words, in standards-based grading, grades are assumed or intended to inform teaching and learning and that purpose is assumed to be primary. Standards-based grading, therefore, makes explicit that grades are to be used formatively as well as summatively, something that traditional grading practices typically do not do. Standards-based grading principles call for reporting academic achievement separately from behaviors, effort, and attendance; referencing both assessment and grading to standards; using report card grades to report current achievement status by prioritizing the most recent evidence of learning; allowing students to edit and resubmit work; using proficiency-based rubrics and decision rules for aggregating them that take their ordinal nature into account; and using quality formative assessments as well as summative assessments (Guskey & Bailey, 2001; Peters et al., 2017).

DESIGN OF GRADING Grades come in many forms. For traditional report cards in the United States, the most common is an ordinal scale of letters (often ABCDF). Many high schools report grades as percentages (0–100), and many schools at all levels that use the ABCDF scale for report cards grade individual pieces of student work on the percentage scale. Many other countries (e.g., Botswana, Canada, Sweden) use some version of a categorical scale, typically letters. An ordinal scale of proficiency levels (e.g., *Advanced, Proficient, Developing, Emerging,* or 4, 3, 2, 1) is often used with standards-based grading, a practice that evaluates student achievement against individual standards rather than subject areas overall. The process of assigning grades is at present mostly a teacher function, although students can be involved, and there is evidence students are fairly accurate in their appraisals of their own work (Sanchez et al., 2017).

MEASUREMENT CONSIDERATIONS IN GRADING Grades have the distinct disadvantage that the general public thinks they understand their meaning ("I'll give that restaurant an A+"; "C is average"), when often such meanings no longer apply, if they ever did. Different stakeholders may view grade categories differently. For example, Waltman and Frisbie (1994) found that on average—and with much variation—parents and fourth graders thought the average mathematics grade was a C+ when in fact it was a B.

Validity In the early 1900s, in what would now be called validation research, studies sought to describe the relationship of grades to intelligence, assuming that intelligence was the construct grades were supposed to measure, just as the purpose of schooling was to sort students so that the bright students received more education and the duller students left school for more menial jobs (e.g., Banker, 1927). As the philosophy of teaching shifted from sorting students to a more achievement-based approach, studies of the validity of grades shifted to comparing grades with tested achievement, as measured by a standardized test (e.g., Moore, 1939), expecting strong relationships between these two measures of achievement. However, this is not and has never been the case (Bowers, 2011; Brookhart, 2015; Brookhart et al., 2016). Over half a century ago, Miner (1967) demonstrated that graded achievement and tested achievement are not the same construct.

There is ample evidence that teachers use nonachievement factors in grades (Brookhart, 2013), against the recommendations of measurement professionals (Stiggins et al., 1989), which may explain part of the difference. Bowers (2009) termed the variance in grades not accounted for by standardized tests of achievement a "success at school factor" (p. 623). His study replicated the existence of the nonacademic factor in U.S. students' grades that Lekholm and Cliffordson (2008) had found in Sweden and named it, arguing that this measure of success at school should be useful for educational decision-making beyond the information available from standardized test scores.

That high school grades are a slightly better predictor of first- and second-year college grades than ACT scores (Westrick et al., 2015) also supports the notion that high school grades include some sort of success-at-school factor. Galla et al. (2019) showed that the incremental predictive validity of high school grades (beyond test scores) for on-time graduation from college was explained by student self-regulation, while the incremental predictive validity of test scores (beyond grades) was explained by cognitive ability. This study is important because it adds a theoretical framework, self-regulation, to the practical idea of a success-at-school factor. (See also Camara et al., this volume, for a review of research related to admissions and predictive validity.)

There is some evidence that grades can yield higher quality information if certain effective grading practices are followed (Pollio & Hochbein, 2015; Welsh et al., 2013; Willingham et al., 2002). There is also ample evidence that high-quality grading practices are not always followed (Brookhart et al., 2016). Cumulative grades, but not individual grades, are valid for purposes of predicting and sorting (Anderson, 2018). Evidence is not strong for the validity of either cumulative or individual grades to provide information about student achievement of specific learning outcomes.

Reliability In the early 1900s, the accuracy of the 100-point scale was questioned (e.g., Starch, 1913), and categorical grading was adopted to increase reliability—or at least as a trade-off between perceived precision and reliability. Most grades currently, in the United States and around the world, use a categorical scale of some length. The categories may be based on cutoffs on a percentage or other point scale or on performance-level descriptions for each of the categories.

Recently, attention has turned to the reliability of grade point averages. Beatty and colleagues (2015) found an overall estimate of reliability of .86 for first-year college grade point average and .93 for overall grade point average. Westrick (2017) reported the range of reliability values for fourth-year cumulative grade point averages at each of 26 four-year institutions was .89 to .92. Thus, while grades on single assessments may be unreliable, cumulative grades are reasonably reliable (Anderson, 2018). Again, though, the reliability in question is reliability for relative ordering. Although this is an important quality, it pertains more to some of grading's many purposes than to others. Reliability of proficiency classification, for example, was not studied.

For some grading purposes, most notably the sorting and ranking ones, the added precision that comes with aggregating, especially up to the level of a grade point average, adds reliability. For other grading purposes, most notably the feedback-related purposes of conveying information about student achievement of specific subject area content or standards, the added precision may muddy interpretation. In the case of a borderline grade, for example, there may not be a meaningful difference in achievement between a very high B and a very low A, but the categories proclaim such a difference and students and teachers may act on the basis of that difference.

When categorical (ordinal) grades are used, nonparametric statistical techniques are best suited to aggregating component grades on students' tests, projects, and other summative assessments into composite grades for reporting. In the early 21st century, that aggregation is most often accomplished with grading software, much of which was developed by programmers who have not studied the nature of grades and treat them parametrically (e.g., using means, standard deviations), with calculations often carried to several decimals and imputing a precision that was not there to begin with. For better or worse, grading software removes some of the decision-making about grading from teachers' hands.

IMPLEMENTATION OF GRADING IN PRACTICE Research on teacher grading practices, for example, what assessments teachers count in a report card grade, has been much studied since the early 1900s and reviewed periodically (Brookhart, 1994, 2013; Brookhart et al., 2016; A. D. Crooks, 1933; Kirschenbaum et al., 1971; Smith & Dobbin, 1960). All these reviews concluded that that the most effective grading system would be to report student achievement on established standards, using explicit criteria, and all found that then-current grading practices typically did not do that. Standards-based grading reforms, beginning in the 1990s and gaining more traction in the 2000s, are intended to remedy this situation. The movement is gaining momentum, but definitive studies are not yet available.

Much research on teachers' grading has used a practical, rather than theory-based, approach and simple survey methods. General findings include the following (Brookhart, 1994, 2013): Teachers try hard to be fair to students, including informing them of what will contribute to their grade. Achievement measures, especially tests, are the major components in grades; teachers commonly also consider effort and ability. Elementary teachers use more informal evidence and observation than secondary teachers, who base grades more on paper-and-pencil tests and other written activities.

Finally, teacher grading practices vary widely, partly because different teachers view the meaning and purpose of grades differently (Frary et al., 1993; Vanlommel & Schildkamp, 2019).

Two recent research agendas (McMillan, 2001; McMillan et al., 2002; Randall & Engelhard, 2009, 2010) support the conclusion that while teacher grading practices are much maligned, in fact academic achievement is the factor teachers consider most heavily. Teachers reported considering academic enablers (McMillan, 2001), such as effort, in students' grades as a way to evaluate students' academic engagement, which some teachers value as part of school learning. Randall and Engelhard (2010) also found that academic achievement is the factor teachers primarily consider when assigning academic grades. They found interactions among ability, achievement, behavior, and effort: with high effort and behavior, low-achieving, low-ability students receive an average grade of C+, and in general, high effort and behavior boosts the grades of students of any ability somewhat. Thus, studies continue to find that teachers' grading practices mix achievement and nonachievement factors, but it seems clear now that achievement is the main component of grades, and teachers' use of other factors is generally reported to be motivated by what in their view is good for students.

If, in general, improving teacher assessment literacy is difficult (C. Campbell, 2013), in grading it is even more difficult because of the multiple conflicting viewpoints teachers hold about grades. For example, Olsen and Buchanan (2019) investigated changes in the understanding and grading practices of teachers involved in year-long professional development designed to reform grading practices in two secondary schools. Change did occur, but it was partial and did not necessarily have the intended effects. Teachers sometimes adapted recommended strategies and then refined them to fit their classroom context, in the process using a belief system that the grading reforms were meant to change.

Students, too, need some basic assessment literacy to understand and benefit from the feedback inherent in the grades they receive. Almost by definition, successful students navigate this well, eagerly wringing any information they can get from their grades to learn more and continue to do well in school (Brookhart, 2001). Students need to learn that grades apply to their learning and not to them as people, that grades are changeable, and that, at least to a large degree, that change is under their control. These lessons, of course, only hold true in situations where teachers' grading practices are supportive of learning.

IMPACT OF GRADING ON LEARNING Grades constitute a form of summative feedback that is used for reporting and certifying student learning. However, successful students also use grades as part of their self-regulation of learning, as they learn to become self-monitoring; as part of their self-efficacy development, as they learn how grades sum up their performance; and as part of their metacognition, as they use grades as one piece of evidence to judge the quality of their own understanding (Brookhart, 2001).

Students consider grades a type of feedback. For example, Harris et al. (2014, p. 121) asked a sample of New Zealand upper primary and lower secondary school students to create a free-response drawing of their experience of feedback—44% of the drawings included grades or scores.

Lipnevich and Smith (2009a,b) pointed out that the most common type of feedback students receive in class is grades. Nevertheless, the U.S. college students in their focus groups deemed grades unnecessary if the purpose of the activity was to learn. In addition, they exhibited negative emotions when receiving a low grade; some became angry at the instructor and others reported dissatisfaction with themselves or embarrassment. In a survey of teacher education students in Australia, "stated grade" was rated as the least useful type of feedback with the exception of group verbal feedback; "written summary" and "brief comment throughout" were rated most useful (Ferguson, 2011).

College students want feedback even on graded work (Pitt & Norton, 2017), using the term *feedback* to mean comments. This study of UK university students showed they interpreted grades and any additional feedback differently depending on their feelings of competence, their feelings about their lecturer, and their level of emotional maturity. One of the students in Pokorny and Pickford's (2010) focus groups of college students in the United Kingdom commented that feedback (comments) accompanying final graded work was moot, since "it's way too late to do anything about it" (p. 24).

The effects of grading practices on students' motivation to learn has also been studied. T. J. Crooks (1988) and Covington (1992) both showed the effects of school evaluation practices on students. Covington (1992) called for practices that fostered what he called "motivational equity" (p. 21); not all students can learn the same things, but all students should experience evaluation practices that support student learning (p. 21):

> Everyone can experience feelings of resolve and a commitment to think more, and to dare more; feelings of being caught up in the drama of problem solving, and of being poised to learn and ready to take the next step.

He showed that low grades do not accomplish this. Lohbeck (2019) found that for a sample of German fourth graders, self-reported grades in mathematics and German predicted academic self-concept in the respective domain (but not the other), which in turn predicted self-perception of effort in the domain. In other words, students whose grades give them evidence that they are effective learners in a domain may be motivated to continue to expend effort in learning in that domain.

EVALUATION OF GRADING INFORMING TEACHING AND LEARNING Grades are used for wide range of purposes, some of which they serve better than others. Grades are multidimensional measures of situated school learning and correlate only moderately with standardized achievement measures. Grades have a long history as assessment information used to inform teaching and learning. There is evidence that grades can be supportive of student learning if certain grading practices are followed, although there is great variation in grading practices. Grading may be one of the most difficult areas in which to support and improve teacher assessment literacy because many people (teachers, parents, students, etc.) have strongly held opinions and feelings about grading, based on their own grading histories.

State Summative Assessment

High-stakes, large-scale assessment has affected teachers' instructional practices (Faxon-Mills et al., 2013; Pedulla et al., 2003); that is, state summative assessment has been and is being used to inform teaching and learning. However, the ability of such assessments to inform teaching and learning is limited, at best.

PURPOSE AND USES OF STATE SUMMATIVE ASSESSMENT State summative assessments are typically administered beginning in March or April and are not intended to provide information to inform the instruction of students within the current academic year. Rather, state summative assessments, also referred to as state accountability tests or simply state tests, are intended to serve state and federal accountability purposes, including district-, school-, teacher-, and student-level accountability. State summative assessments may provide information that is useful to teachers to evaluate their instructional practices or useful to schools and districts to evaluate their curricular and instructional programs.

DESIGN OF STATE SUMMATIVE ASSESSMENT Beginning in the 1990s, many state assessments were designed to serve as examples of the type of assessment that should be occurring in the classroom. The increased use of constructed-response items, direct writing prompts, and performance tasks within traditional end-of-year assessments and alternative formats such as portfolio assessments is evidence of this desire to give students a test worth teaching to (Resnick & Resnick, 1992; Rothman, 1995; Shepard, 1991). For example, one state wrote that the purpose of state assessments was "to improve classroom instruction by (a) providing useful feedback about the quality of instruction and (b) modeling effective assessment approaches that can be used in the classroom" (Massachusetts Department of Education, 1999, p. 4).

The implementation of NCLB in the early 2000s brought both increased stakes in terms of school accountability and increased testing with the requirement of annual testing of all students in Grades 3 through 8 plus one grade in high school (NCLB, 2001). The 2010s further increased the stakes and footprint of state summative assessments in the classroom with longer tests, more complex standards and tasks, the declaration of an "honesty gap" between national test scores, state test scores and teacher grades, and the use of test results for teacher evaluation (Achieve, 2013). These factors converged to create an expectation that state summative assessments can, should, and must provide "actionable information" to inform teaching and learning (CCSSO, 2013;

U.S. DOE, 2015), another instance of the pressure to extract such information even from assessments not designed for it.

The primary information from state summative assessments is an indication of student performance in relation to state achievement standards in terms of scale scores and achievement levels. State summative assessment operationalizes the state's content standards and performance expectations, showing teachers how the state is interpreting the content standards. In many state programs, this information is supplemented by sets of released items with annotated samples of exemplar student responses to constructed-response items or essays. All of this information is designed to help increase teachers' and administrators' understanding of the state's interpretation of the content standards and expectations for students at each grade level. Teachers and administrators are expected to use such information to calibrate their expectations for student performance with those of the state (Madaus et al., 2009; Parsi & Darling-Hammond, 2015).

MEASUREMENT CONSIDERATIONS IN STATE SUMMATIVE ASSESSMENT Federal peer review has ensured that the process of gathering data and producing scores on state summative assessments is typically sound. Operational best testing practices employed routinely by assessment contractors in conjunction with federal requirements and peer review for state assessments systems, including requirements regarding accessibility and the availability of accommodations, and the commonplace practice of convening technical advisory committees for state testing programs help ensure the quality of the process and the validity of the results, although the availability and quality of validation studies related to state assessment programs remains a concern. State assessment programs, however, are not immune from real and perceived errors or irregularities in test administration, scoring, or reporting, which are often consequential and highly publicized (Rhoades & Madaus, 2003). In addition, there are long-standing concerns about the generalizability of results from state summative assessments and test score inflation due to test preparation and a narrowing of instruction to the particular standards, or portions of standards, included on the state summative assessment (Koretz, 2017).

The process of designing state summative assessment programs for school accountability, including the selection (or rejection) of items based on their statistical properties, however, limits the usefulness of the information gathered through state summative assessments to inform teaching and learning because the assessments' design focuses on the summative purpose. State assessments are designed to produce information about overall student achievement of, or proficiency on, state standards, focused appropriately on the validity of inferences related to these broad student outcomes. In contrast, most instructional decisions require more fine-grained diagnostic information. In addition, measures to inform teaching and learning need to provide timely information about student performance and the instructional process, and they should be based on an explicit theory of student learning and a theory of action related to improving instruction (Ing et al., 2021).

Measures that inform improvement in instruction need to be explicitly linked to a theory of improving instruction (Bryk et al., 2015). In contrast to measures for the purposes of accountability, measures for improvement provide timely information about the improvement process to address questions such as, How can processes central to the theory of improvement be altered to be improved? (Solberg et al., 1997). That is, rather than indicating whether a student has learned something, measures for improvement speak to the processes that support student learning instead of focusing solely on the outcomes of student learning. Not surprisingly, multiple measures are needed for this ambitious purpose of improving instruction at scale (see, e.g., Takahashi et al., 2022). No single measure can provide all the necessary information for what is required to increase the quality of instruction (Bennett, 2015; Bennett & Gitomer, 2009; Newton, 2010). Instead, a comprehensive system of measures is needed to gather the necessary information to inform instructional improvement efforts, including outcome measures such as indicators of student learning (ranging from daily analysis of students to student performance on annual standardized achievement measures) as well as measures for improvement that inform the daily work of those in the system who are working toward the same improvement goals. In such situations, the same measures may be used for different purposes by different users (Gitomer & Duschl, 2007).

Increased pressure to provide more information and more actionable information from lengthy state summative assessments has led to the reporting of subscores for individuals and groups of students. Given that most state summative assessment programs are designed to fit unidimensional item response theory models and often contain a small number of items measuring specific content standards or strands, subscores provide little, if any, useful, reliable information to inform teaching and learning (see Zenisky et al., this volume).

IMPLEMENTATION OF STATE SUMMATIVE ASSESSMENT IN PRACTICE Expectations that state assessment and accountability data could or should affect teaching and learning in some way still exist. That rhetoric abounds on state and district education websites. The U.S. DOE has encouraged the use of district data systems, including state summative test results, for informing teaching and learning and suggested data literacy supports for teachers (U.S. DOE, 2009, 2011). At the beginning of the NCLB era, there was concern about state test data having negative effects on instruction. For example, Pedulla et al. (2003) classified states according to the severity of the state summative test's consequences for districts, schools, and teachers and for students. They surveyed classroom teachers and found that teacher responses varied with the severity of stakes and with level (elementary/middle/high school). Teachers reported feeling pressure to raise scores on the state-mandated tests. A large percentage of teachers reported pressure to deliver instruction that ran counter to their own beliefs about good teaching practice.

There is some evidence that some teachers use state assessment results to inform changes in curriculum content and emphasis, changes in pedagogy, and changes in teacher interactions with individual students (Faxon-Mills et al., 2013). Changes in curriculum included changes in sequence and emphasis of topics, a focus on basic skills, and a focus on higher order skills, depending on the state and study. Changes in pedagogy included a focus on test preparation and changes in instruction and assessment practices. Changes in interactions with students included individualizing instruction and focusing on students who were close to the proficiency cutoff and could potentially cross it and increase pass rates; these students have been referred to as "bubble kids" (Faxon-Mills et al., 2013), an unintended consequence of NCLB accountability requirements that has been alleviated by the use of indicators that consider student performance across the achievement continuum. Similarly, Au (2007) found that the primary effects of high-stakes testing on teaching were curriculum narrowing, fragmenting subject area knowledge into test-related pieces, and increased use of teacher-centered instruction; however, in some cases the high-stakes tests led to curriculum expansion, synthesizing subject area knowledge, and student-centered instruction. Some of these changes might have potentially positive effects on learning, for example, focusing on student thinking or integrating knowledge, and some might have negative effects, for example, focusing on some students more than others or increasing lectures and other teacher-centered instructional methods.

More recently, some studies have reported that while some teachers use state test results, at least in a general way, classroom assessment—both summative and formative—is more closely linked to decisions about changes in instruction. For example, Wilkerson et al. (2021) found that Nebraska teachers' use of state summative, interim, and formative assessment data to inform instruction was consistent with the frequency of administration of each type of assessment, with data from state summative tests being used least often.

Farrell and Marsh (2016) studied five high-needs middle schools in three districts. They found that state assessment results were used mostly at the beginning of the school year for grouping students but were not associated with other changes in instruction.

IMPACT OF STATE SUMMATIVE ASSESSMENT ON LEARNING Any impact of state summative assessment on student learning would be mediated through its impact on curriculum and teaching, as described in the previous section. Given this distal connection to teaching and learning and its intended purpose as a summative tool for accountability, one might argue that it is inappropriate to even ask questions about the impact of state summative assessment on learning. A primary purpose of federally mandated state assessment programs, however, is to support Title I, which is intended "to provide all children significant opportunity to receive a fair, equitable, and high-quality education, and to close educational achievement gaps" (Every Student Succeeds Act, PL 114-95, 2015, Sec. 1001. Statement of Purpose).

Therefore, one way to examine the impact of state summative assessment on student learning that might be appropriate is in the way that it was intended to be used: that is, in terms of the extent to which test scores indicate that progress has been made in providing equity in educational opportunity and closing achievement gaps. Results from NAEP and international assessments indicate that the achievement gap by socioeconomic status has remained large and stable since the 1970s (Hanushek et al., 2019). Results from the NAEP Reading and Mathematics tests at Grades 4 and 8 since 1992 showed modest impact on the pace of reductions in the gap in performance between Black students and White students through 2005 (Braun et al., 2010). NAEP results showed little change in achievement and in the size of the White–Black and White–Hispanic achievement gaps from 2010 to 2020 (U.S. DOE et al., 2020).

EVALUATION OF STATE SUMMATIVE ASSESSMENT INFORMING TEACHING AND LEARNING State summative assessment programs are designed to provide summative information for accountability and administrative purposes. The growing clamor for more formative information from all assessments perhaps causes more discomfort in the area of state summative assessment than for other types of assessment because its intended purpose and design are decidedly summative. It remains to be seen whether state summative assessments can be redesigned to better provide desired formative information.

Some states are considering the use of through-year assessment models, which emerged as an alternative to a single end-of-year test during the 2009–2010 Race to the Top assessment competition, in an attempt to provide both formative (i.e., timely and instructionally useful) information and the required summative information from a single assessment program. To date, many of those efforts have been limited to replacing the single end-of-year state summative test with the traditional interim assessment model in which all standards are assessed on a short test administered three or four times per year, but some states are attempting to better synchronize the sequence of standards tested throughout the year with instruction or to closely connect the state assessment with the taught curriculum (Powell et al., 2022).

Using information from a single assessment event for both formative and summative purposes poses a significant challenge to the interpretation and use of through-year assessment models for summative accountability purposes (Dadey et al., 2023; Gong, 2021). Some states and state assessment programs, such as Smarter Balanced Assessment Consortium, have responded to this not by changing the summative assessment, but by offering optional interim assessments of varying lengths (full comprehensive assessments, interim assessment blocks, or focused interim assessment blocks, depending on how many standards are assessed) aligned to the state summative assessment. Teachers can choose to administer and use these interims for formative purposes and even create their own from item banks. In other words, Smarter Balanced has responded to this pressure on summative assessment by shifting resources to interim assessment (Smarter Balanced, 2019).

ASSESSING LEARNING TRAJECTORIES FROM CLASSROOM LESSONS TO SUMMATIVE

The assessments described in the previous section were measures of performance or learning status at single points in time, and this evidence about achievement status was what was leveraged to inform teaching and learning. In contrast, the assessments described in this section offer information for informing teaching and learning by looking at student achievement trajectories. This work represents a distinction from assessments in the previous section in two ways: It looks at performance over time and it is theory based—in contrast with any growth modeling using the more traditional measures.

Learning implies growth or change. Growth or change can be measured or indicated in several different ways: as trend lines, pre–post (or pre–mid–post) designs, or progress on a map of development in a domain of learning. Different assessment methods have been applied to show learning spans or trajectories, and three of these are described in this section. In this chapter, the generic term *assessment of learning spans or trajectories* is meant to include any assessment method based on trajectories. The term *learning progressions* has come to be applied to one particular method, described below in the section "Assessment Based on Learning Progressions." However, the idea of learners progressing from naive to expert, or from novice to mastery, in a particular domain of content underlies all three of the assessment methods in this section. Assessments of trajectories spanning from classroom lessons through summative (outcome) achievement rely on individual measures taken over time. Inferences are then made from students' trends or growth, as well as the final outcome or level achieved, and this trend information is used to inform teaching and learning.

Assessing trajectories of student achievement fits with the formative assessment cycle as described by the question loop in Table 15.1: Where are we going? Where are we now? How will we get there? Assessments that focus on learning trajectories ask those same questions, but in terms of a bigger developmental picture than when those questions are used for classroom formative assessment of small sequences of lessons. In both, the emphasis is on the formative, on "getting somewhere." Unlike state standards or classroom learning goals, learning trajectories describe both the destination and the way to get there. This makes learning trajectories potentially a rich source of formative information.

This section considers three assessment methods focused on learning spans or trajectories. Curriculum-based measurement (CBM) has the longest history of the three, in both research and practice. Student learning objectives (SLOs) date from the recent standards-based accountability era and arose in a practical context. Assessments based on learning progressions seek to connect theory about how students learn various topics or concepts with their assessment. Learning progressions are the subject of several ongoing research agendas.

Curriculum-Based Measurement

CBM is an approach to measuring the academic growth of individual students during instruction toward a general outcome of interest (Deno, 2003a). The measurement process is based within the curriculum, and the measures themselves are carefully constructed measures of that general outcome (i.e., the endpoint), rather than being drawn from or based on a particular curriculum. Repeated measurements of student achievement of the general outcome provide information to the teacher regarding a student's progress toward the ultimate learning goal, signaling whether changes to instruction are needed.

Figure 15.1 shows a sample graph of CBM measures collected over 27 weeks of instruction, with an instructional change occurring after Week 16 (Espin et al., 2018). Key components of the CBM progress-monitoring graph include: (a) baseline data, representing the student's current level of performance; (b) peer data, representing typical performance and reflecting the discrepancy between the student and peers; (c) a goal line, representing the expected rate of growth and end-of year level of performance; (d) data points, representing the number of correct and incorrect responses on weekly probes; (e) slope or growth lines, representing the student's rate of growth over time; and (f) solid vertical lines, representing instructional changes. The slope of the student's performance during a particular instructional intervention provides feedback to the teacher of the success of that instructional approach for the student.

Purpose and Uses of CBM

CBM was developed in the 1980s to support the instruction of individual students with disabilities (Deno, 1985, 2003b). CBM was developed as a means "for teachers to use technically sound, but simple, data in a meaningful fashion to document student growth and determine the necessity for modifying instructional programs" (Stecker et al., 2005, p. 795).

Since its initial development as a tool for elementary school teachers of students with disabilities, the applications and use of CBM and progress monitoring have expanded to other grade levels and multiple content areas, general education, and other uses, such as program evaluation (Allen & Smith, 2022; Jenkins & Fuchs, 2012). A major application of the principles of CBM has been as part of the large-scale Response-to-Intervention program (Fuchs & Fuchs, 2007; National Center on Response to Intervention, 2010). The use of CBM for program evaluation and large-scale screening programs such as Response to Intervention is not directly related to the use of assessment information collected during instruction to inform instructional decisions and therefore is not a focus of the discussion of CBM in this chapter.

Design of CBM

From its inception, CBM has focused on the frequent collection of student performance on a general outcome measure linked to students' overall proficiency on long-term goals (Deno, 1985, 2003a; Stecker et al., 2005). The purpose for using the general outcome measure, rather than more discrete measures tied to individual standards or particular knowledge and skills, was to provide teachers with clear evidence of student progress toward the desired goal, that is, the level of proficiency desired at the end of the year or other defined instructional period (Fuchs & Deno, 1991). The goals are specific to the content area assessed and the grade level of the students. The general outcome measures are intended to focus on those academic performances that represent "vital signs of educational development" (Deno & Mirkin, 1977, p. 14).

The vital signs measured on individual assessments used in CBM are to be regarded as indicators of student performance within a domain, rather than as direct measures of specific knowledge and skills within the domain, in some ways consistent conceptually with the manner in which state summative tests are designed to provide an indicator of students' overall proficiency and not a measure of student mastery of individual standards (Fuchs, 2016). Central to the vital signs approach to measurement in CBM is the belief that (a) teachers, in many cases, have no simple indicators with which to monitor the effects of instruction and (b) information that teachers are more likely to have, student mastery of individual standards or skills, is not necessarily evidence of increasing student proficiency on the general outcome (Deno, 1993).

A progress-monitoring graph depicting student performance over time, like the one in Figure 15.1, is one of the fundamental components of CBM (Espin et al., 2018).

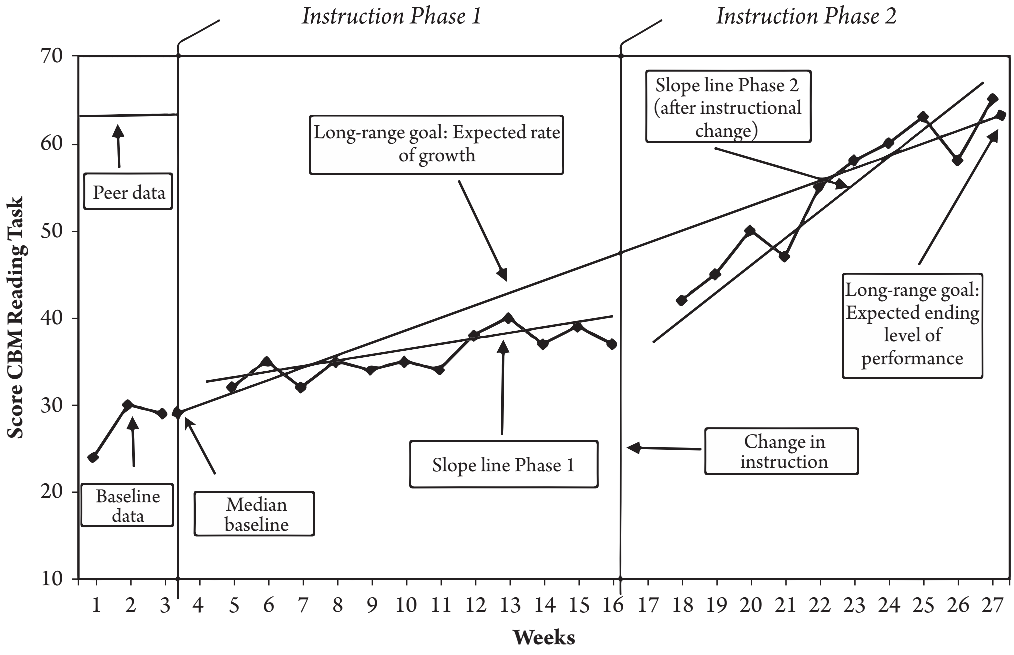

FIGURE 15.1

Sample Graph of Curriculum-Based Measurement (CBM) Measures Collected Over 27 Weeks of Instruction, With an Instructional Change Occurring After Week 16

Note. From "Curriculum-Based Measurement Progress Data: Effects of Graph Pattern on Ease of Interpretation," by C. A. Espin, N. Saab, R. Pat-El, P. D. M. Boender, and J. van der Veen, 2018, *Z Erziehungswiss, 21*(4), 767–792. https://doi.org/10.1007/s11618-018-0836-9. Copyright Creative Commons Attribution 4.0 International License. Reprinted with permission. (http://creativecommons.org/licenses/by/4.0/).

To evaluate the effectiveness of instruction for a particular student, the teacher examines the graph to determine whether the student is progressing at the desired rate of growth and whether the student will achieve the goal (van den Bosch et al., 2017).

By comparing the slope of the line depicting student progress to the line representing anticipated growth over time toward the long-term target, teachers can determine whether an instructional change is warranted (Stecker et al., 2005). Consistent below-target performance is regarded as evidence that a change in instructional program might be needed. Consistent above-target performance may suggest that the teacher consider raising the long-term goal. In addition to information about an individual student's progress toward the long-term goal, teachers may also consider normative information (e.g., student progress relative to others in the class) or group-level information (e.g., median student progress in a class).

Measurement Considerations in CBM

The quality of CBM information is enhanced by the use of technically sound assessment instruments developed for particular grade levels, content areas, and skills (Stecker et al., 2005; Marston, 2012). Each assessment probe is developed to be a measure of the same overall performance and equal in difficulty to all other probes measuring that area of proficiency. A critical component of CBM, therefore, is the selection or development of multiple alternate forms of an assessment instrument, each of which is a sufficiently reliable indicator and supports valid inferences of students' overall proficiency. Consequently, true applications of CBM cannot rely on teacher-made assessments unique to an individual teacher or group of teachers or on the ad hoc selection of commercially developed assessment instruments. CBM requires the use of assessments that have been empirically selected to meet the technical requirements outlined in this paragraph (Deno, 2003a,b).

An extensive body of research describes the type and technical quality of measures used in various content areas: reading (Fuchs, 2004; Miura Wayman et al., 2007; Tindal, 2013); writing (McMaster & Espin, 2007; McMaster et al., 2011), and mathematics (Foegen, 2008; Fuchs, Fuchs, Compton, et al., 2007; Fuchs, Fuchs, & Hollenbeck, 2007; Gersten et al., 2011). Research also describes measures used with various groups of students, for example, students at secondary levels (Espin et al., 2018; J. Shin & McMaster, 2019), English learners (H. Campbell et al., 2013), and students who are deaf or hard of hearing (Lam et al., 2020).

Fuchs (2004, p. 189) described two approaches to the task of identifying "measurement tasks that simultaneously integrate the various skills required for year-end performance" as (a) identifying tasks that "correlate robustly (and better than potentially competing tasks) with the various component skills constituting the academic domain" and (b) a "systematic sampling of the skills constituting the annual curriculum to ensure that each weekly CBM represents the curriculum equivalently." Tasks at early elementary grades are more likely to reflect the first approach and tasks at upper elementary and secondary grade levels are more likely to reflect the second approach (Allen & Smith, 2022).

Although linked directly to students' overall proficiency on the desired construct, the vital signs approach of CBM relies heavily on the use of the predictive validity of the measures and has given less weight to issues related to alignment of the individual measures to specific content standards (Deno, 2013), raising concerns and criticisms about the appropriateness of the measures. CBM assessment probes used at higher grade levels in general education classrooms and in content areas other than reading have focused more on alignment to content standards (Allen & Smith, 2022; Fuchs, 2004; Morton, 2013). Recent research has focused on the alignment of CBM measures to state content standards and has used performance on state assessments as a criterion for evaluating the technical adequacy of CBM measures (Shriner & Thurlow, 2012; Tindal, 2013).

Implementation of CBM in Practice

After the selection of assessments, the key steps in the CBM process are regular administration of the assessments, scoring and graphical depiction of student performance, and appropriate interpretation of student progress toward the long-term goal. From its beginnings, an important feature of CBM was the use of assessment instruments that were short, easy to administer, and easy for teachers to score. Such features were intended not only to increase the reliability of scores, but also to increase the utility of the process; that is, such features were considered critical to enhancing the likelihood that teachers would administer the assessments and collect data on student performance on a regular basis.

CBM arose from a different view of teaching and learning than classroom formative assessment, one more focused on affecting teacher instructional adjustments and thereby enhancing student achievement (Stecker et al., 2005), similar in focus to formative evaluation for mastery learning (Bloom et al., 1971) rather than informing students' regulation of learning directly. Although described as a tool to support data-based or data-driven decision-making (Fuchs & Deno, 1991; Hosp & Hosp, 2012) the development and implementation of CBM is not a product of the accountability-centered data-driven decision-making models described as a policy theory of action (Shepard et al., 2017). Rather, with its foundations in special education and the instruction of individual students by special education resource teachers, CBM is more aptly classified as an application of single-subject research design or time series design (Marston, 2012).

In contrast to the data-driven decision-making models assumption that teachers will know how to help students (Shepard et al., 2017), CBM is based on the assumption that a teacher does not know in advance whether, or how well, a particular instructional intervention will work for an individual student. In their Data-Based Program Modification manual, a precursor to CBM, Deno and Mirkin (1977, p. 22) offered the following as two fundamental assumptions of Data-Based Program Modification:

> Assumption 1: At the present time we are unable to prescribe specific and effective changes in instruction for individual pupils with certainty. Therefore, changes in instructional programs which are arranged for an individual child can be treated only as hypotheses which must be empirically tested before a decision can be made on whether they are effective for that child.
> Assumption 2: Time series research designs are uniquely appropriate for testing instructional reforms (hypotheses) which are intended to improve individual performance.

The theory of action for CBM, therefore, can be described as experimental teaching or data-based instruction in which the teacher takes an active role in verifying that an instructional strategy is effective for the individual student (Fuchs & Bradley, 2012). The fundamental assumption of CBM is

> successful intervention required that teachers receive clear and unambiguous feedback regarding the general effects of their instructional efforts. If teachers are either uncertain about the overall effects of their efforts, or believe they have been successful simply because a student learns the specific content that has been taught, their efforts to improve growth will be unsuccessful. (Deno, 2003a, p. 6)

In early applications of CBM, individual teachers were required to create their own progress-monitoring graphs, but with the assistance of computer-based applications to generate the graphs, the focus of teacher assessment literacy needs has shifted to the ability to read and interpret the information provided by the graphical depiction of student progress (Fuchs & Fuchs, 2007). This requires a basic understanding of how to interpret fluctuations in student performance over shorter time intervals in relation to changes in performance over larger time intervals (van den Bosch et al., 2017). With training, teachers, on average, become fairly proficient in reading the data on CBM graphs, but there is variation in the ability of individual teachers (Espin et al., 2017). Further, teachers have greater difficulty in interpreting the data on the graphs, particularly ambiguous data, and linking it to instructional decisions (van den Bosch et al, 2017). Although studies have shown that students make more progress in achievement when teachers use graphs than when they simply record data (Fuchs & Fuchs, 1986), the recent research on teachers' ability to read and interpret graphs highlights the need for professional development in interpreting CBM data and using it in conjunction with other data about instruction and student performance to inform instructional decisions. A meta-analysis of the impact of professional development on teachers' knowledge, skill, and self-efficacy in data-based decision-making (Gesel et al., 2021) found that professional development supports could have a positive impact on teachers' data-based decision-making skills.

Impact of CBM on Learning

Relatively few studies have directly examined the impact of the use of CBM on student learning (Deno, 1985). Stecker et al. (2005, p. 795) stated, "The hope was that by responding instructionally to students' poor patterns of performance, teachers should be able to enhance student achievement." Two studies that synthesize the research studies that have been conducted in this area indicate that the proper use of CBM by teachers as part of instruction can have a positive impact on student learning (Jung et al., 2018; Stecker et al., 2005). Both studies, however, describe significant challenges to teachers implementing CBM with fidelity and using the results of measurement to inform modifications to instruction.

Stecker et al. (2005) described research with students with disabilities from the initial development of CBM in the 1970s and research conducted with students in general education settings that began in the 1990s. Critical variables examined in the studies reviewed include the use of data-based decision rules, the inclusion of a skills-analysis component to the reporting of CBM results, and the provision of recommendations for next steps by teachers. Across the studies for students with disabilities and in general education settings, Stecker et al. (2005) found that the use of CBM alone did not have a significant positive impact on student learning (i.e., progress monitoring without making use of the data), but that significant gains in student achievement were attained under the following five conditions: (a) teachers respond to the CBM data to tailor instruction to student needs; (b) a preestablished data-decision framework is adhered to; (c) computer applications are used to improve the efficiency of all aspects of the CBM process; (d) skills analysis, in conjunction with consultation, is provided; and (e) some form of recommendations for next steps are provided. Studies conducted in general education settings also demonstrated that instructional modifications based on class-wide CBM data, rather than data for individual students, could be used to produce gains in student achievement and that with training, CBM data could be used successfully by students in peer-tutoring contexts to improve student achievement.

Jung et al. (2018) conducted a meta-analysis of the effects of CBM, under the generic label of data-based individualization (DBI), on the performance of students with intensive learning needs. They examined the effect of teachers' use of DBI on student achievement and the factors that influence the strength of effects of DBI. Of the 14 studies included, 6 were conducted after the original Stecker et al. (2005) review. Overall, Jung et al. (2018) found that the use of DBI alone and the use of DBI with additional information or supports, DBI Plus, had significant positive effects on student achievement compared to a "business-as-usual" control. Consistent with the findings of the earlier review, they found that providing teachers with some type of support for the interpretation and use of the progress monitoring data to make instructional modifications had a significant impact on the efficacy of using DBI to improve student achievement.

The findings from both reviews on the impact of CBM on student learning is consistent with the definition of formative assessment adopted by the authors for this chapter, that is, that a practice is formative to the extent that the evidence about student achievement is not only elicited, but also interpreted and used by teachers and students to make decisions about instruction. Research on the impact of CBM on student learning is also consistent with the findings elsewhere in this chapter that additional research is needed on teacher assessment literacy, and additional ongoing support is needed for teachers' use of data to inform instructional decisions.

Evaluation of CBM Informing Teaching and Learning

CBM is supported by an extensive body of research compiled over more than 3 decades since its inception (Allen & Smith, 2022; Espin & Wallace, 2004; Stecker et al., 2005; Tindal, 2013). Much of the research on CBM can be classified into three categories that are critical to developing a validity argument for its use to inform teaching and learning (Fuchs, 2004): (a) research on the technical adequacy of the measures and scores produced from each assessment, (b) research on the technical features of the slope (i.e., student performance over time) to determine whether increasing CBM scores are in fact associated with improvement in overall competence in the academic domain, and (c) studies concerning instructional utility to determine whether practitioners can use the CBM information to improve instructional decisions and student achievement. This third category of research is particularly relevant to the use of CBM to inform teaching and learning.

Tindal (2013) described the dual focus of the CBM research as grounded in a "measurement paradigm" and focused on a "training in data use and decision-making paradigm"; although the former is an essential foundation for the use of CBM to inform teaching and learning, the latter is critical to ensuring that it can and will be used appropriately to inform teaching and learning. As with formative assessment processes, a long-acknowledged threat to the use of CBM to improve teaching and learning is that when the data indicate that a change in instructional intervention is needed, teachers do not know how or what to adjust in their instruction (Fuchs et al., 1989; Stecker et al., 2005). Recent research has identified teachers' aforementioned difficulty in reading and interpreting CBM graphs (Espin et al., 2017; van den Bosch et al., 2017) and teachers' insufficient professional development in making instructional decisions based on data as two additional threats to the use of CBM to inform teaching and learning (Espin et al., 2021).

Student Learning Objectives

SLOs arose at the local level in the context of teacher evaluation policy, giving teachers some say in the measures for which they would be held accountable. The locus of SLO shifted to the state level when teacher evaluation based on teachers' impact on student learning became a federal requirement for states receiving Race to the Top funding or NCLB waivers. Using SLOs meant teachers of subjects that were not represented in state accountability tests could be held accountable for student progress in their subjects. This brief section on SLOs is included in the chapter for the sake of completeness because the intent of SLOs is that student learning will be affected. However, there is little evidence to date that student learning is in fact affected.

Purpose and Uses of SLOs

SLOs—also referred to as student learning goals, student learning outcomes, student learning measures, and student growth objectives—delineate a process designed to integrate assessment and instruction for the purpose of improving student learning. Typically, teachers and building principals together negotiate the measures to be used for the SLO process in teacher evaluation (New York State Education Department, 2014; Texas Education Agency, 2021). The SLO process is intended to include (a) developing carefully planned learning goals for what students will learn over a given period of time and (b) evoking critical, evidence-based thought about student progress and growth through instruction (Community Training and Assistance Center [CTAC], 2018).

Initially used as part of innovative pay-for-performance compensation programs (CTAC, 2004, 2013), SLOs were developed as a school improvement tool to be implemented collaboratively by administrators and teachers at the local district and school levels (Diaz-Bilello & Thompson, 2019). SLOs gained popularity as they were adopted by many states as the primary method for evaluating teachers' impact on student learning in content areas and grade levels in which there was not a state assessment (Marion & Buckley, 2012) in response to federal requirements under Race to the Top and NCLB waivers (Briggs et al., 2019).

Design of SLOs

The use of the SLO process as a tool for school improvement has been described as a school-based application of the management-by-objectives model (Chapman, 2014). The design and implementation of individual SLOs by teachers under the CTAC (2013) model is driven by six elements: (a) learning content (what will be learned), (b) instructional strategies (approach to teaching), (c) interval of instruction (time frame for teaching), (d) assessment (measures of learning), (e) student growth targets (numeric goals for learning), and (f) student population (who is included).

Those elements are then implemented in a cyclical process that includes periods of instruction, assessment, and adjustment. A common SLO cycle might include the following steps (Diaz-Bilello et al., 2016; Diaz-Bilello & Thompson, 2019): identify the learning goal, develop or select assessments, analyze baseline data, establish targets, analyze a cycle of instruction and assessment, revise targets when appropriate, analyze a cycle of instruction and assessment, and complete the SLO evaluation. Each cycle of instruction would also include the use of formative assessment procedures and adjustments to instructional strategies, as needed.

Measurement Considerations for SLOs

A critical distinction between SLOs designed for accountability and SLOs designed for improving instruction is the focus on comparability across districts, schools, and teachers, as evidenced by the level of autonomy allowed in two critical aspects of the SLO process: the development and selection of assessments and the establishment of growth targets (Cushing & Meyer, 2014; Diaz-Bilello & Thompson, 2019). It is less important for SLOs designed to improve teaching and learning than for SLOs for teacher evaluation to be comparable across contexts (Lane & DePascale, 2016). Even when SLOs are used for teacher evaluation, however, there is a large degree of variation across states in how SLO performance targets have been established (Crouse et al., 2016).

Using SLOs requires a basic understanding of reliability of observed scores, in general, and difference scores, in particular. Decisions made about individual students' meeting their growth target will be affected by error in the pretest and posttest scores. For example, a common approach to setting growth targets for SLOs is referred to as the "half-the-distance" method, in which the growth target on a posttest is defined as the score that is halfway between the student's baseline score on a pretest and either a specified criterion score or the maximum score on the posttest (i.e., a perfect score; Austin Independent School District, 2015; Missouri Department of Education, 2015). The use of approaches such as half-the-distance can produce different results based on the level of student performance and the design and type of assessments used (Austin Independent School District, 2015) and can produce results that are not acceptable within the given context, such as target scores that are below the accepted passing score on the posttest (English et al., 2015).

Implementation of SLOs in Practice

When implemented with fidelity, the SLO process should provide feedback to teachers, administrators, and students about the specific learning content and growth targets expected to be achieved during the interval of instruction. It also should provide feedback to teachers and administrators about the instructional strategies teachers will use to support student learning and attainment of growth targets. At the end of the SLO process, teachers and administrators receive summative feedback on whether a SLO or multiple SLOs implemented by individual teachers have been met. That decision is usually based on the percentage of students included in the SLO process who met their individual student goal (Hall et al., 2014).

The SLO process requires a high level of training and support at the district and school levels and is difficult to implement without an extended commitment to the process (Schneider & Johnson, 2018). In addition to general assessment literacy, specific teacher needs include the ability to clearly specify measurable objectives that are also primary curriculum goals and to select or create assessment tools whose results are valid, reliable, and fair indicators of those goals. Teachers and administrators need the ability to analyze and interpret assessment results against both student learning goals and teacher professional goals for purposes of high-stakes decision-making.

Impact of SLOs on Learning

Few research studies have been conducted examining the extent to which the use of SLOs provides feedback that teachers use to inform their instructional strategies and decisions during the implementation of the SLO process. The limited number of surveys that have been conducted show mixed and/or conflicting results about the type of feedback received and used by teachers from the SLO process (Diaz-Bilello & Thompson, 2019; Lachlan-Hache et al., 2015). In preparing this chapter, the authors found no studies on the effects of using SLOs on student achievement.

Evaluation of SLOs Informing Teaching and Learning

SLOs are a process designed to integrate assessment and instruction. The process involves teachers and administrators agreeing on specific learning goals that the teacher will monitor and on selecting assessments to use as measures over time of student progress on those goals. The process was largely overtaken by the teacher accountability movement. Little research to date has focused on the effectiveness of SLOs for improving teaching and learning.

Assessment Based on Learning Progressions

The National Research Council (2007, p. 219), focusing on science education, defined learning progressions as "descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn." Corcoran et al. (2009, p. 8) used a more expansive definition: "Learning progressions in science are empirically grounded and testable hypotheses about how students' understanding of, and ability to use, core scientific concepts and explanations and related scientific practices grow and become more sophisticated over time, with appropriate instruction." In mathematics, the term *trajectory* has also been used, stemming from Simon's (1995, p. 135) proposal of a "hypothetical learning trajectory" that referred to mathematics teachers' expectations for how student learning might proceed for certain mathematical content. Work in learning progressions has developed in other countries as well, beginning as early as the 1980s.

Although more work on learning progressions has been done in mathematics and science than in other disciplines, the idea has been applied to many different disciplines. The appeal is that learning progressions can form the basis for the teacher's instructional decisions (Deane et al., 2015), improving teaching and learning. In fact, some definitions of learning progressions include not only progressions in students' thinking but also progressions in instructional experiences that might help students move along, and the relative lack of attention to including the latter in progressions is one of the current criticisms of learning progression research (Mosher, 2022).

All definitions of learning progressions envision learning as growth in sophistication of understanding, not solely a body of content to be covered, and envision progress as growth in understanding not referenced to age or grade level (Heritage, 2008). Learning progressions describe typical development of expertise in a domain for most, but not all, students. They are not rigid; a student's level on a learning progression may fluctuate as learning is taking place. In many domains, learning progressions may describe building blocks that lead up to more complete understandings. In science, lower levels on a particular learning progression may also describe commonly held facets or naive conceptions that students must abandon as their learning progresses (Alonzo, 2018). Since students learn in the context of instruction, learning progressions are not independent of curriculum sequence (Wiliam, 2018b).

Many—if not most—sets of content standards and curricula are built on a logical, developmental progression of knowledge, skills, and competencies across grade levels (e.g., Common Core Standards Writing Team, 2019; Confrey et al., 2014; National Centre, 2020) and certainly qualify as progressions in the plain English sense of the word. However, the empirically developed and validated learning progressions defined and described in this section represent something beyond this; while acknowledging a relationship with students' educational experiences, the assessments based on learning progressions described in this section also seek to detail cognitive growth in a more theoretical way.

Purpose and Uses of Assessment Based on Learning Progressions

Learning progressions are a hypothesized vehicle for making assessment more useful for informing teaching and learning, whether at the classroom or the large-scale level, and especially to guide teachers' formative assessment practices (Alonzo, 2018). Teachers need to have a conception of what comes next in learning to respond effectively to students' work. It is for this reason that informing teaching and learning seems the most salient purpose for learning progressions and assessments based on them. However, a learning progression framework can also be used to describe student performance summatively.

Design of Assessment Based on Learning Progressions

There are different approaches to constructing learning progressions as a priori, theoretical statements about how learning takes place in a domain. In general, such approaches are based on the work of human panelists (as opposed to data mining) and can be categorized as top-down, beginning with the ideas of experts in the domain, or bottom-up, beginning with the ideas of curriculum content experts and teachers about what is best taught when (Heritage, 2008). Learning progressions can be big, encompassing K–12 learning in a broad domain (e.g., writing), or relatively small, describing the acquisition of understanding on a single topic.

A true learning progression—as opposed to, for example, a curriculum scope and sequence or a simple rubric—requires empirical evidence or validation, or what Graf and van Rijn (2016) have called *empirical recovery*. If learning progressions are a representation of a theory about how children learn and develop in a domain, they should be empirically verifiable (Denvir & Brown, 1986; Graf & van Rijn, 2016).

Therefore, design concerns for assessment based on learning progressions include both verifying the progression and designing and validating measures to assess them. Confrey (2019) cautioned that studies need to distinguish between the learning progression and its measures. Evidence for the former must verify the constructs in the progression, while evidence for the latter must validate a scale for measuring student progress on the progression. When the two are conflated, what results is much like the validity evidence for any other educational measure. When the two are distinguished, research on the progression can inform research on the measure and vice versa.

Learning progressions design work is done in a content area, because learning progressions are content specific, and most often requires partnerships between researchers and educational practitioners. The result is often assessment that is neither completely classroom based and teacher developed nor completely externally developed, as for large-scale assessments. This hybrid development is a hallmark of assessments based on learning progressions, at least to date, and foreshadows some future possibilities as technology makes possible the coordination needed for this kind of assessment.

Measurement Considerations for Assessment Based on Learning Progressions

Some learning progressions are constructed to help teachers create or select classroom-based tasks and progressions, and some are constructed to serve as the basis for externally developed tasks and questions that teachers can use in the classroom. Content-related evidence for the validity of the progression has been reported for the former (Furtak, Circi, & Heredia, 2018; Hess & Kearns, 2010, 2011). A wider range of validity evidence has been reported for the latter, including content-related evidence as well as evidence of whether measurement models fit assessment data as predicted. It is worth noting that studies do not always support hypothesized learning progressions, or do not support them completely (e.g., Pham et al., 2021). Examples of how several different programs of research have treated validity and validation follow.

One of the more influential frameworks for the design of learning progressions (Duschl et al., 2011), especially in science, is the BEAR Assessment System (Berkeley Evaluation and Assessment Research Center; Wilson, 2009; Wilson & Sloane, 2000). The BEAR Assessment System is based on four principles, each with accompanying building blocks for the assessment system (Wilson et al., 2016, pp. 7–8):

> (1) Assessment should be based on a developmental perspective of student learning; the building block is a *construct map* of a progress variable that visualizes how students develop . . . (2) There must be a match between what is taught and what is assessed; the building block is the *items design* . . . (3) Teachers must be the managers of the system, with the tools to use it efficiently and effectively; the building block is the *outcome space*, or the set of categories of student responses that make sense to teachers . . . (4) There is evidence of quality in terms of reliability and validity studies and evidence of fairness, through the data analytics; the

building block is an algorithm specified as a *measurement model* that provides for a visual representation of the students and the items on the same graph (called a "Wright Map") and a number of other data analytic tools.

The BEAR Assessment System, then, creates a priori learning progressions based on research and uses researcher-developed measures in a system managed but not authored by teachers (Black et al., 2011; Wilson & Scalise, 2016; Wilson & Sloane, 2000). The measures may be based on multiple-choice or open-ended items. Classroom teachers use the system during instruction. Validity can be supported if the instruments reflect the expected characteristics of learning described in the progressions. Construct maps can be used to study this match (Black et al., 2011), and statistics from unidimensional or multidimensional item response theory analyses can be offered as evidence for appropriate internal structure and reliability. Assessment results are used to adjust both the learning progression and the items (e.g., Schwartz et al., 2017). Table 15.3 shows an

Table 15.3 Construct Map Showing a Learning Progression for Student Understanding of Earth in the Solar System

Level	Description
5 8th grade	Student is able to put the motions of the Earth and Moon into a complete description of motion in the Solar System which explains: the day/night cycle the phases of the Moon (including the illumination of the Moon by the Sun) the seasons
4 5th grade	Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that: the Earth is both orbiting the Sun and rotating on its axis the Earth orbits the Sun once a year the Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the sky the Moon orbits the Earth once every 28 days, producing the phases of the Moon Common error: Seasons are caused by the changing distance between the Earth and Sun. Common error: The phases of the Moon are caused by a shadow of the planets, the Sun, or the Earth falling on the Moon.
3	Student knows that: the Earth orbits the Sun the Moon orbits the Earth the Earth rotates on its axis However, student has not put this knowledge together with an understanding of apparent motion to form explanations and may not recognize that the Earth is both rotating and orbiting simultaneously. Common error: It gets dark at night because the Earth goes around the Sun once a day.

(continued)

Table 15.3 *(continued)*

Level	Description
2	Student recognizes that: the Sun appears to move across the sky every day the observable shape of the Moon changes every 28 days Student may believe that the Sun moves around the Earth. Common error: The Sun travels around the Earth. Common error: It gets dark at night because the Sun goes around the Earth once a day. Common error: The Earth is the center of the universe.
1	Student does not recognize the systematic nature of the appearance of objects in the sky. Students may not recognize that the Earth is spherical. Common error: It gets dark at night because something (e.g., clouds, the atmosphere, "darkness") covers the Sun. Common error: The phases of the Moon are caused by clouds covering the Moon. Common error: The Sun goes below the Earth at night.
0	No evidence or off-track

Note. From "Measuring Progressions: Assessment Structures Underlying a Learning Progression," by M. Wilson, 2009, *Journal of Research in Science Teaching, 46*(6), p. 720. Copyright 2009 by John Wiley and Sons. Reprinted with permission.

example of a construct map that functions as a learning progression for student understanding of the Earth in the solar system (Wilson, 2009, p. 720).

Confrey and her colleagues (Confrey & Toutkoushian, 2019; Confrey et al., 2020) created Math Mapper 6–8, a digital learning system used in middle-grades mathematics classrooms. Multiple forms of assessment are used, including multiple choice, multiple select, numeric, one letter, and open ended. Like the BEAR system, Math Mapper 6–8 is based on a priori learning progressions derived from research, but it uses a different validation framework. The validation process is iterative: Results from the digital learning system are used to adjust the learning trajectory, and vice versa. Evidence is sought that the assessment taps domain knowledge (cognitive evidence), that the assessment is aligned to the curriculum and opportunity to learn and is used to inform teaching and learning (instructional evidence), and that the assessment reliably yields model-based information about student performance (inferential evidence; Pellegrino et al., 2016). Both quantitative (e.g., from empirical studies of model fit) and qualitative (e.g., observation-based design studies) evidence are sought.

Pham et al. (2021) studied three learning progressions for formative assessment in middle school mathematics. They, too, used a priori learning progressions derived from research to describe three related progressions—equality and variable, linear functions, and proportional reasoning—and created items to measure them. They used item response theory to investigate whether student response evidence supported claims about the order of levels within each progression and multidimensional item response theory to investigate whether student response evidence supported predicted relationships across pairs of progressions.

In language, Bailey and Heritage (2014) use a more descriptive and qualitative approach to validation of learning progressions in the Dynamic Language Learning Progression project. The project collected a body of elementary school–age children's oral and written language (including from English learners and children with a range of language experiences) and from this corpus created learning progressions illustrating how children's use of language in school becomes gradually more sophisticated over time. Validation included a study of students' oral explanations (both procedural and conceptual) of mathematics problems. The language samples are searchable online. The language learning progressions are intended for use with both regular and English learner students and eventually are intended to inform interim and summative assessment, as well as formative assessment and classroom instruction (Bailey & Heritage, 2014) and student self-assessment (Goral & Bailey, 2019).

One critique of the learning progressions approach is that cognitive understanding is not quite as linear as learning progressions may suggest (Shavelson, 2009). Several research agendas have used nonlinear cognitive models or maps that relate knowledge components within a content area. Automated methods for cognitive diagnostic modeling are often used to create these cognitive maps (Goldin et al., 2016). Wilson et al. (2016) saw a place for exploratory automated methods as well as confirmatory methods such as those used in the BEAR system, partly depending on how much is already known about learning in an area. However, some argue that even in well-researched areas such as mathematics learning, automated methods help with the diagnostic precision needed for teachers to make decisions about next instructional moves (Kingston & Broaddus, 2017; R. Liu & Koedinger, 2017).

For example, R. Liu and Koedinger (2017) showed that most students who can calculate the area of a two-dimensional shape given its sides can also calculate the sides, given the area, and human experts most often consider calculating area of a shape to be one skill. However, cognitive modeling showed that this is not true for all shapes: for example, calculating from the radius to the area of a circle is a different skill from calculating the radius given the area. R. Liu and Koedinger (2017) showed that is probably because of the use of the square root. This additional precision about student understanding has obvious utility for informing teaching and learning.

Another example of using automated techniques for cognitive mapping is the Dynamic Learning Maps project, which uses diagnostic classification modeling to create the learning maps (Kingston et al., 2016). Originally applied to alternate assessment of state standards, Dynamic Learning Maps can support formative assessment and instructional decisions about typical educational standards as well (Kingston & Broaddus, 2017) and allows for representing finer grained, more diagnostic information for content standards than that produced by large-scale, standards-based tests.

In sum, to date, much of the measurement consideration for assessment based on learning progressions has focused on the development and validation of the learning progressions themselves. Relatively little research has focused on traditional measurement considerations related to their use by teachers, schools, and districts for various formative or summative purposes. Learning progressions may eventually provide a framework for the design of assessment systems, for example, district assessment systems, that include both classroom and large-scale assessment (Duschl et al., 2011). The most effective assessment systems will relate formative and summative measures in coherent, comprehensive, and continuous ways (Wilson & Draney, 2004) that most current assessment systems do not.

Implementation of Assessment Based on Learning Progressions in Practice

Learning progressions and assessment based on them are still very much in the developmental or small-scale implementation phase. Such development and implementation almost always involves a research–practitioner partnership of some sort. Even so, teachers do not always use learning progressions as intended (Alonzo & Elby, 2019).

Australia has been an early adopter of this approach, beginning work on what they called "subject profiles" in 1988 (Rowe & Hill, 1996). Subject profiles are vertical maps of achievement in a domain that can be used as a framework for assessing student work. The work in Australia was conceived in a summative context, looking for ways to record and report student growth. Currently, it is also used formatively, illustrating the aforementioned trend toward assessment being called on to inform teaching and learning that is happening all over. Australia currently uses a learning progression approach to student growth in reform efforts aimed at both curriculum and assessment, for both formative and summative purposes (Commonwealth of Australia, 2018; Confrey, 2019).

Furtak and Heredia (2014; Heredia et al., 2016) showed that the design and use of common formative assessments using a learning progression can inform how teachers interpret student thinking (Furtak, Circi, & Heredia, 2018) and adjust instruction. Furtak, Circi, & Heredia (2018) showed that a group of science teachers were able to design formative assessment tasks generally aligned with multiple learning progressions in high school biology, specifically in the area of natural selection, with some gaps and variations and associated uneven student learning gains across the learning progressions. In this research agenda, implementation can also proceed through the external development of curriculum-embedded assessments of learning progressions (Briggs & Furtak, 2020), including performance-based tasks and labs, phenomenon-based item clusters, and conceptually oriented multiple-choice questions.

Math Mapper 6–8 (Confrey & Toutkoushian, 2019; Confrey et al., 2020) provides teachers with assessments designed to give immediate, actionable feedback to both teachers and students. Teachers are allowed to use the assessments as they see fit during their instruction. Confrey and Toutkoushian (2019) found that teachers vary in their use of the assessments; most measure students' progress about two thirds of the way through an instructional unit, when there is still time to adapt instruction.

Black et al. (2011) described a broad model of pedagogy based on teachers' understanding "road maps" or progressions toward curricular learning goals and using classroom assessment strategies that provide feedback to students. Wilson and colleagues applied this model in the BEAR Assessment System by working with content area researchers in various disciplines. Work has been conducted in the areas of science (Wilson & Draney, 2004; Wilson & Sloane, 2000), scientific argumentation (Yao et al., 2015), structure of matter (Morell et al., 2017), data science (Arneson et al., 2019), and statistical reasoning (Lehrer et al., 2014; Wilson & Lehrer, 2021). The BEAR Assessment System embeds assessment in the classroom teaching and learning context. Teachers are involved in managing the way the assessment system is used in their classrooms, collecting and scoring student work, interpreting the results in instructional terms, and using the results in instructional decision-making. To facilitate this, the BEAR Assessment System provides a single scoring guide for each variable in the progression; teachers do not need to use a new scoring guide for each assessment. In so doing, the assessment gives direct evidence of students' location on the progression and implications for instruction. Looking beyond the development of assessments, Lehrer et al. (2014) developed an approach that integrates the learning progression for an instructional program in middle school statistics with predeveloped items and teacher-managed formative assessment. This approach has been expanded to incorporate a mobile device–based system for recording teachers' daily observations of their students in alignment with the postulated learning progression (Lehrer, 2021; Wilson, 2021, 2023).

Impact of Assessment Based on Learning Progressions on Learning

Most studies of the effectiveness of learning progressions have studied the learning progressions themselves, seeking to confirm that students do in fact demonstrate thinking as described in the levels, or studied their measures, seeking to validate that performance on a measure reflects the levels described in the learning progressions. Some have studied the effects of using learning progressions, or of professional development using learning progressions, on the alignment of teacher-created formative assessment tasks, teachers' interpretation of student work, or instructional decision-making (Alonzo & Elby, 2019; Furtak, Bakeman, & Buell, 2018; Heredia et al., 2016); in other words, these studies investigated how learning progressions might inform teaching.

Furtak, Circi, & Heredia(2018) showed that students whose biology teachers participated in professional development using learning progressions, designed formative assessment, used it with their students, and reflected on next steps in instruction did increase along the learning progression as measured against their own baselines.

A few studies have used experimental designs. Wilson and Sloane (2000) used an experimental design to evaluate the initial work of the BEAR Assessment System in the context of middle school science. Results demonstrated that students whose teachers used the assessment system improved their learning—with both statistically and educationally significant effects—beyond that of students whose teachers were in either of two control groups: teachers who participated in professional development but were not required to use the assessment system and comparison teachers who taught the regular middle school science curriculum. More recently, BEAR Assessment System work in the context of middle school statistics also showed statistically and educationally significant results in a controlled experiment using a change score approach (Gochyyev & Wilson, 2021; Wilson & Lehrer, 2021).

Some experimental studies have investigated the impact of using cognitive tutoring systems based on cognitive maps and found the effect to be strong and positive (Ritter et al., 2007). R. Liu and Koedinger (2017) compared achievement of students who used a cognitive tutor with that of students who used the cognitive tutor redesigned with adjustments to the cognitive map based on the results of data mining. Students in the redesigned condition had significantly higher posttest scores than control students, even after controlling for pretest scores, which they interpreted to mean that the quality of the cognitive map had improved and made a difference in the effectiveness of the tutoring system.

Evaluation of Assessment Based on Learning Progressions Informing Teaching and Learning

The idea of learning progressions has both theoretical and intuitive, practical appeal. Logically, learning progressions should help address the known difficulty teachers have deciding precisely what feedback or next instructional moves would help learners the most in moving toward learning goals (Heritage et al., 2009; Ruiz-Primo & Furtak, 2006, 2007; Ruiz-Primo & Li, 2013; Schneider & Gowan, 2013; see the "Formative Assessment" section above). Teachers demonstrably do need some way to conceptualize "what's next" to adjust their teaching.

The current state of the field of assessments based on learning progressions shows some promising proofs of concept. Most of these projects leverage the recent increase of technology in the classroom to support classroom-based formative assessment with deep pockets of external research and methodological capability. Several projects leverage teacher professional development and external research support, some with and some without large-scale measurement methodology. At present, it is difficult to imagine enough of these resource-intensive projects to cover the curriculum (also see Baird et al., 2017). When these projects talk about going to scale, they usually mean involving more classrooms in a computer-based tutoring or assessment system, rather than being generally or routinely used in the educational landscape.

Work on assessment based on learning progressions does illustrate the major theme of this chapter, the trend toward the formative. Learning progressions research continues the shift from the focus on groups and aggregate scores to the focus on individuals. Yet several of the research programs described in this section show how difficult it is to apply a particular learning progression to an individual student, even if it is clear what "typical development" looks like.

Work on learning progressions also highlights the role of learning theory in assessment and assessment design. Most of the learning progression research described in this section is theory driven, with the progression distilling research on how students learn particular content into a developmental path. This section shows that personalized learning and assessment system design fueled by data science also has the potential to be driven more by data than by theory. The balance of theory and data science in future learning progression development is likely to be quite interesting.

FUTURE DIRECTIONS

The concepts of teaching and learning continue to evolve, and the pace of change appears to be increasing (Bennett, 2014). In a historical sense, the inseparable connection between teaching and learning is still relatively new, with research on teaching that focused on the student dating only to the last quarter of the 20th century (Peterson, 1979; Rosenshine, 1979; Shulman, 1986) and the current focus on student-centered assessment emerging around the same time (Chappuis, 2022; Stiggins, 1994). The definition and scope of learning—and student-centered education in particular—continues to expand, redefining nearly all aspects of teaching and learning: what is taught, how it is taught, why it is taught, who is taught, and by whom (Carter & Darling-Hammond, 2016; Cochran-Smith & Lytle, 1999; Faltis & Valdés, 2016). As previously described, this shift from "instructivist" (Box et al., 2015, p. 972) or "social efficiency" (Shepard, 2000, p. 4) teaching focused on delivering content to learners to teaching that supports students as they construct their own meaning has fueled the press for all assessment to be formative. The argument will be summarized in more detail in the "Conclusion" section; here, it is enough to note the growth in interest and use of formative assessment methods (classroom formative, interim, and common formative), the use of summative assessment methods (grading and state accountability assessment) for formative purposes, and the development of assessments focused on trajectories that describe the process as well as the final outcome of learning.

In addition, the introduction of technology into almost every aspect of schooling, from record keeping to communication, to curricular and instructional resources, to instruction itself, is changing the nature of teaching and learning (Bennett, 2015, 2018; see also Bennett et al., this volume). Information technology is advancing rapidly, including the availability of computers and other devices in the classroom, as well as Internet and communications technology linking students in classrooms to external information sources and to other students. The rapid growth of information technology has generally changed most aspects of schooling—for better or worse—and has the potential to both support and shape the ongoing transition to more learning-oriented assessment.

It is within this fluid environment that the future of assessment to inform teaching and learning will be determined. The argument throughout this chapter has been that usefulness of assessment in informing teaching and learning is directly related to its proximity to learning, factors such as the timing and nature of the content assessed and the manner in which it is assessed, the nature of the information the assessment provides as feedback to teachers and students, and the type and level of pedagogical skill needed to effectively use the information provided. It is inevitable that the nature of assessment information will change as the understanding of teaching and learning evolves (Mislevy, 2018a; Pellegrino, 2014; von Davier et al., 2019; Wilson & Scalise, 2016) and the constructs that are valued and measured shift to meet societal needs (L. Liu et al., 2017; Stoeffler et al., 2020). It is likely that prevailing measurement models and measurement principles will be adapted or replaced as the field expands its view of what is important to measure (Mislevy, 2018b; Taylor et al., 2018). And it is possible that the very nature of assessment will change as technological advances, including advances in artificial intelligence, increasingly reshape education and all aspects of life (Bennett, 2018; von Davier et al., 2021; Wilson et al., 2015).

The Future of Assessment Information

In the third edition of *Educational Measurement*, Bunderson et al. (1989) outlined a four-generation framework to describe the effect that ongoing advances in technology would have on educational assessment.

- *Generation 1. Computerized testing*: administering conventional tests by computer.
- *Generation 2. Computerized adaptive testing*: tailoring the difficulty or contents of the next item or an aspect of the timing of the next item on the basis of test takers' responses.
- *Generation 3. Continuous measurement*: using calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student's achievement trajectory and profile as a learner.
- *Generation 4. Intelligent measurement*: producing intelligent scoring, interpretation of individual profiles, and advice to learners and teachers, by means of knowledge bases and inferencing procedures (Bunderson et al., 1989, p. 368).

Further, Bunderson et al. characterized a "shift in emphasis from institutional purposes [of educational measurement] to individual purposes" as characterizing "the distinction between the first two and last two generations of educational measurement" (p. 369). It is such a shift in emphasis in the purpose of educational measurement that is likely to move the field from Generation 1 and 2 assessments, designed to provide information about groups of students to support policy decisions about institutions, to a much greater focus on Generation 3 and 4 assessments, designed to produce information about individual student performance and growth to inform teaching and learning. For example, Zheng et al. (2019) showed that formative assessment information embedded in MATHia, an intelligent tutoring system for middle and high school blended learning, can predict state test scores beyond the ability of a prior-year test.

As demonstrated throughout this chapter, in the time between the third and fifth editions of *Educational Measurement*, the shift in demands on assessment from serving institutional purposes to serving individual purposes has begun and is exerting growing pressure on Generation 1 and 2 assessments and their stakeholders. While many classrooms currently use Generation 1 and 2 assessments, Generation 3 and 4 assessments are being gradually introduced in research projects such as the ones described in the section on "Assessment Based on Learning Progressions" and require partnerships between schools and external organizations. Generation 3 and 4 assessments are not yet the common practice—but then, not that long ago, neither were Generation 1 and 2 assessments.

Although the field as a whole has progressed through the framework over the 30 years since the framework was published, the past decade has seen significant movement in K–12 assessment. Advances in the capabilities and availability of technology have fueled widespread adoption of computer-based assessment and rapid acceleration from Generation 1 (computerized testing), through Generation 2 (computerized adaptive testing), to the cusp of Generation 3 (continuous measurement). Computerized adaptive testing is well established as the primary mode of testing for interim assessments and has been accepted as a viable option for state summative assessment (Bennett, 2015).

It is in the classroom, however, that there is most likely to be an emergence of Generation 3 assessment—the use of calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student's achievement trajectory and profile as a learner (Bennett & Gitomer, 2009; Scalise & Wilson, 2011; Wilson & Sloane, 2000). This brings with it implications for privacy rights and responsibilities, teacher assessment literacy, and school technological capacity and policies.

Constructs

The constructs measured by assessments designed to support teaching and learning are likely to shift in two areas. The first shift involves measuring content in currently measured academic areas to a much greater depth and measuring additional areas that are not currently a primary focus of K–12 educational measurement. The second shift involves measuring process as well as outcomes in each of those areas.

The first shift—driven by academic content, curricular, and instructional changes—involves the use of assessments designed to measure students' higher order thinking skills via performance on complex tasks. The first shift also includes measurement of other skills not currently a primary focus of K–12 educational assessment. Those include so-called 21st-century skills such as communication, teamwork, and collaboration, as well as constructs such as social-emotional learning and student engagement. While these types of changes in assessment are already being driven by the development of new content standards in the core areas of English language arts, mathematics, and science, it is likely that similar changes will also be seen in other content areas. These changes may bring with them long-standing generalizability issues related to performance assessment because responses to complex tasks are subject to a large person-by-task interactions (Gao & Baxter, 1994; McBee & Barnes, 1998).

The second shift involves increased interest in and capacity to measure process in addition to student outcomes. Computerized assessment has provided the opportunity to efficiently collect process data such as time spent completing an item or progress through an instructional unit, intermediate steps or pathways, and changes to responses, as well as student outcomes. Including such process data in measurement has the potential to improve significantly the quantity and quality of information about student thinking that can be derived from assessment. As discussed previously, for formative assessment the main goal is not measurement of student achievement of a learning goal, but rather a description of the status of student understanding on the way toward a learning goal; the main information sought is an understanding of the nature of student thinking, not the correct answers that thinking will generate (Heritage & Heritage, 2013).

The availability of additional data alone, however, is likely to do little to improve the ability or use of assessment to inform teaching and learning. It will be necessary to process that wealth of data to produce Generation 3 and 4 type information that helps teachers understand students' achievement trajectories and interpret individual student profiles or offers instructional advice to learners and teachers. Generating that type of information will require the expansion of current measurement models, as well as the development of new models, and will likely incorporate long-standing and emerging work and modeling in the areas outside traditional psychometrics and assessment, such as computer-assisted instruction, intelligent tutoring, and data mining (Gobert et al., 2013; Koedinger et al., 2018; Ritter et al., 2007).

Measurement and Other Models

Since the NCLB era began in 2001, the design and development of K–12 large-scale summative and interim assessments has relied on the application of unidimensional measurement models such as the Rasch model and a variety of one-, two-, and three-parameter item response theory models. The use of these models has served the field well for accomplishing the institutional purposes of educational measurement (i.e., providing comparable estimates of students' overall ability across a variety of alternate test forms administered within and across years). As discussed in the sections of this chapter on state summative and interim assessment, these models have been less useful in providing information to inform teaching and learning, particularly at the individual student level.

As the focus of educational measurement has shifted to the classroom and individuals, new models and different applications of current models have emerged to inform the design of assessments intended to inform teaching and learning and support formative assessment. (See Cai et al., this volume, for a discussion of modeling for different types of assessment.) Theory-driven applications of the traditional unidimensional and multidimensional Rasch models have been used to help define, describe, and measure student performance along learning progressions with the specific intent of providing information to inform teaching and learning (Black et al., 2011; Briggs & Alonzo, 2012; Shepard, 2018; Wilson, 2012). As discussed in the section on learning progressions, to date, much of the innovative research in this area is still being conducted in close partnership between researchers and local educators. Widespread application of this emerging work at scale has yet to be accomplished.

While the unidimensional and multidimensional applications of the Rasch model have been useful in describing relatively simple linear progressions, new models or modeling approaches are needed to support the development of assessments to measure and support the development of complex learning progressions/trajectories or learning maps (Confrey et al., 2014; Toutkoushian, 2019; Kingston & Broaddus, 2017; Kopriva et al., 2016). In addition, latent class models such as cognitive diagnostic models and hierarchical diagnostic classification models have applied a multidimensional approach to describe more fine-grained profiles of student performance and estimates of mastery (Bradshaw et al., 2014; Templin & Bradshaw, 2014). In addition to these models focused on outcomes for individual students interacting with items and tasks, new measurement models are being developed to provide information for new types of assessment designed to measure 21st-century skills such as collaboration.

Computational psychometrics is described as

> a blend of stochastic processes theory, computer science–based methods, and theory-based psychometric approaches that may aid the analyses of complex data from performance assessments, including collaborative assessments. This discipline organically developed around the complex next-generation learning and assessment systems that include performance tasks, such as collaboration, games, and simulations. (von Davier, 2017, p. 3)

Computational psychometrics combines theory-based psychometrics and computer science–based methods to develop a model that incorporates process and outcome data into information about student performance on complex tasks.

The computer science–based methods employed in computational psychometrics, such as machine learning and data mining, can also be applied on their own without theory-based psychometrics to develop models of student learning and achievement. Such models fall under the broader category of learning analytics, which may incorporate data mining, machine learning, artificial intelligence, and other data science techniques in an attempt to provide information to inform instruction. Although instructional theory, learning theory, and measurement theory are not necessary aspects of learning analytics, it is expected that the most effective use of learning analytics to inform teaching and learning will involve an iterative process in which theory and empirical data are used to guide the collection and analysis of data and transform data into understanding that is useful to teachers and students (Wong et al., 2019). As Wilson and Scalise (2016, p. 2) noted, "applying learning analytics to educational assessment therefore requires negotiating a key intersection, which is at the interface of measurement technology and information technology." At a minimum, it should be expected that the dissemination of information and feedback to teachers and students from learning analytic models and any of the models described above will be based on a solid theoretical foundation.

Nature of Assessment

As described throughout this chapter, the nature of formative assessment to inform teaching and learning is distinctly different from the nature of summative and interim assessment designed primarily to serve institutional purposes. By their nature, the process of formative assessment, and therefore the tools to support formative assessment, are most effective when they fit the Generation 3 description of measures that are embedded in a curriculum to estimate, continuously and unobtrusively, dynamic changes in the student's profile as a learner. At this time, however, the field of educational measurement has not yet reached the point where technology can support continuous and unobtrusive assessment in the classroom. The role of the student in technologically based measurement also bears investigation. As advances in technology progress, it can be expected that some aspects of teacher and student activities and interactions in the classroom will be more likely candidates for continuous and unobtrusive data collection than others. As assessments and measurement models are built from that data, it will be critical to ensure that the data collected from the classroom are the most relevant data to inform teaching and learning and that they are representative of the interactions among teachers and students.

Assessment Literacy Requirements in the Future

Teacher assessment literacy requirements for the effective use of assessment information will remain centered on deep content knowledge and the attendant understanding of how to select learning targets and success criteria, communicate them to students, and accurately recognize how students are thinking from the evidence in their work. Teachers will benefit from deeper content and pedagogical knowledge as content, curriculum, and instructional changes focus teaching and learning on more complex cognitive processes and greater individualization and personalization.

An additional assessment literacy demand is likely, namely, that teachers will have the skills needed to filter more data and information than is currently normally available for individual students. At this time, teachers are often asked to make sound instructional decisions based on sparse evidence. As information technology makes the collection, processing, and storage of data from the classroom feasible, teachers will have to evaluate multiple sources of information and data, identify relevant information, and combine multiple pieces of data to make informed decisions. They will need supports to do this (U.S. DOE, 2009). An alternative view, from the artificial intelligence research, is that the teacher's job becomes simpler, with the technology doing the interpretation and presenting the teacher with a simple set of options (Hamilton et al., 2023). The authors doubt that this will be the case. As teaching has become more and more complex, the central role of the teacher as the most influential factor in students' learning has been demonstrated over and over (Hanushek, 2011; Ker, 2016; Opper, 2019).

As the field transitions from Generation 3 to Generation 4 measurement and the information provided from assessment becomes more prescriptive, some assessment literacy demands placed on teachers may be reduced; however, demands for skills in other areas are likely to increase. Teachers will need to be able to judge the quality of the information provided and determine the basis of the information: Was it customized for an individual student or generic for a wide variety of students? Was it based on learning theory or solely on probability?

Student assessment literacy requirements will remain focused on the ability to understand learning targets and criteria, compare work to criteria, and use different types of feedback to improve. As instruction becomes more personalized through the use of technology, students are likely to be required to process feedback from sources other than their teacher. Similarly, an increased focus on collaborative skills may require students to process real-time feedback from their peers in the course of completing performance-based activities.

Advances in ways researchers and educators can work together to determine what data to collect, what information to convey to teachers and students, and the best ways to convey that information will be critical to advancing the assessment literacy of teachers and students and improving the usefulness of information from assessment to inform teaching and learning.

CONCLUSION

The chapter has explored eight general types of assessment—classroom formative assessment, interim assessment, common formative assessments, grading, state summative assessment, curriculum-based measurement, student learning objectives, and assessments based on learning progressions—each used in different ways to inform teaching and learning. These types of assessment vary in the nature of the assessment information they provide, their effectiveness in terms of improvements in teaching and learning, the quality of the assessment process and information, and assessment literacy requirements. Expectations about the future of assessment to inform teaching and learning suggest that these types are changing and will continue to change.

All told, descriptions of these eight types of assessment work together to demonstrate the movement that Bunderson et al. (1989) anticipated and recent history has borne out. The arc of the history of assessment is moving from institutional purposes and bending toward assessment to inform teaching and learning.

At this time, the argument is not that the formative trend in assessment is largely successful in informing teaching and learning. The chapter has documented both progress and significant challenges. Rather, the argument is that assessments are being pressed to do so. Four of the types of assessment discussed in this chapter—classroom formative assessment, curriculum-based assessment, common formative assessments, and assessments based on learning progressions—were developed specifically to inform teaching and learning. Interest in classroom formative assessment especially is growing, and research suggests that, when done well, it can be highly effective. The other four types—student learning objectives, interim assessment, grading, and state summative assessment—were developed for more institutional purposes and bent toward the purpose of informing teaching and learning as a response to stakeholder and public demand. As the chapter has shown, that bending produced some cracks in validity and reliability, and efficacy research, where available, has shown that the latter four types of assessment are less effective for producing learning gains than are the types of assessment designed for that purpose.

Nevertheless, both educators and the public now believe that wherever information about student achievement is available, it should be harnessed in the service of teaching and learning, and they press for that at every turn. This chapter has discussed some of that pressure, including pressure on tests at all levels to produce formative information, the relatively recent development of adding a layer of district-mandated assessment (e.g., interim and common formative assessment) in order to have more assessment information, and the rise of a grading reform movement that for the first time explicitly states that grading should support learning. The conviction that assessment should inform teaching and learning is expressed on state and district websites and materials, in meetings at all levels of educational administration, in the classroom, and in communication with parents and caregivers.

Related to this movement toward the formative, assessment to inform teaching and learning is moving toward more student involvement. Pressure toward more student involvement comes from the advent of more situated, sociocognitive, and sociocultural views of learning, as well as concerns for students' social-emotional well-being, social justice, and fairness. As it has become clearer that students construct knowledge with the help of teachers, peers, and materials, and that students regulate their learning, it follows that students need to be involved in the generation and interpretation of evidence of that learning. Research has affirmed this general trend to be useful for informing teaching and learning. For example, feedback research has shown that elaborated feedback on which students can build is generally more effective than simple knowledge of results. Formative assessment research also has shown that students do better when they understand the criteria by which they will be evaluated and use them as learning is taking place.

Therefore, informing teaching and learning means designing, administering, and using information that is, to the extent possible, student specific, producing valid and reliable information about individual learners as well as groups; pedagogy and content specific, producing information at instructionally actionable grain sizes; and situation specific, producing information relevant to a learning community, most often a classroom. This is not to say that assessment should not continue to serve institutional, accountability, or administrative purposes or produce valid and reliable data for groups relevant to administrative decision-making—for example, schools, districts, or states and demographic groups within those. The research reviewed in this chapter suggests that those purposes are being—and should be—overtaken, not eliminated, by purposes about supporting teaching and learning. This is happening in both effective and ineffective ways, as the reviews above described. The field of educational measurement needs to understand this phenomenon and continue to work to support quality assessment in a world where the balance of assessment information tips toward assessment to inform teaching and learning.

Points from this review may suggest ways to do this. However, these changes are not guaranteed, or even likely, to occur without concerted and coordinated effort among measurement professionals, content specialists, learning theorists, practicing educators, and others. As this review showed, progress is being made, but it is not easy, quick, or without challenges. With this caution, this review suggests three potential ways forward toward the intelligent assessment Bunderson et al. (1989) and others aim for.

First, the nature of assessment information needs to be understood in broader terms than conventional scales. At least one type of assessment, classroom formative assessment, includes verbal information, in the form of feedback comments, that is important for teaching and learning. In addition, verbal descriptions add meaning to performance levels in rubrics and developmental levels in learning progressions. Learning analytics can add process information, and such multiple measures can contribute to more precise estimates of student learning. This attention to meaning is not a new concept in educational measurement. The importance of inferences, interpretation, and the utility of tests and test scores has been a central tenet of educational measurement for decades. However, the movement toward the formative increases the importance of the underlying meaning of assessment information and the necessity of connecting interpretations of assessment results to understandings of content and pedagogy.

Second, attention to both the quality of the assessment process and the quality of the information will help with understanding and improving assessment to inform teaching and learning. Formative assessment, on the one hand, typically emphasizes the process, and more attention to the quality of the questions or discourse used in the formative assessment process might improve its effectiveness. Interim assessment, on the other hand, typically emphasizes the quality of the information and has not yet really found an effective process for using that information in the service of teaching and learning. More work on this point may lead to diminished interest in interim assessments or more effective processes or designs for their use in teaching and learning. Currently, assessments or assessment systems that show promise for continuous or intelligent assessment are demonstration projects or successful research projects that explore assessment processes and tools that are far from commonplace in schools. In any case, insistence on quality of both process and information for all assessment information can be expected to improve teaching and learning.

Third, the more assessment turns away from institutional purposes and moves toward assessment to inform teaching and learning, the more assessment literacy demands rise for educators and students. In large-scale testing for institutional purposes, educators and students are consumers, and the bulk of assessment literacy requirements reside with test developers. As the trend moves toward assessment to inform teaching and learning, assessment users—educators and students—have a larger role to play and thus greater assessment literacy needs. This is the case even with professionally developed assessments that are used to inform teaching and learning, because the teachers and learners become the owners of the information and its use and consequences. It is even more the case with educator-developed assessments that are used to inform teaching and learning, for example, classroom formative assessment, common formative assessments, and grading. As the connection between assessment and the teaching and learning process expands, assessment literacy expands to include deep content knowledge and an understanding of how learning happens, in general and for specific concepts.

As assessment to inform teaching and learning evolves in these ways, as the chapter sections have suggested, advances will be required in validity theory and methods, measurement models and methods, technology, and assessment literacy. This is a broad list offering a wide array of productive opportunities for scholarship and practice and is not offered lightly—as this chapter shows, bumps in the road may be expected and collaborative work will be needed.

Finally, it is important to note that this chapter about assessment to inform teaching and learning has been written from an assessment point of view. The reviews have described scholarship and practice in eight types of assessment, described a broad assessment trend toward the formative, and derived three assessment-related subthemes. The authors envision a companion chapter that could have been written by experts in teaching and learning, from a teaching and learning point of view. In other words, given the trend toward the formative, assessment to inform teaching and learning will also need to be understood from a learning point of view. Some of the research agendas reviewed have already resulted in partnerships with instructional experts and learning theorists. Advances in this kind of work may truly be the future of assessment to inform teaching and learning.

ACKNOWLEDGMENTS

The authors express sincere gratitude to Drew Gitomer, Dylan Wiliam, and Mark Wilson, whose thoughtful reviews helped us immensely. The reviewers were instrumental, especially in organizing and presenting the vast amount of material covered in this chapter and in contributing suggestions from their deep subject matter expertise. We would also like to thank the editors of this volume, Linda Cook and Mary Pitoniak, for their wise leadership and guidance throughout this enormous undertaking.

REFERENCES

Abrams, L. M., & McMillan, J. H. (2013). The instructional influence of interim assessments: Voices from the field. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment* (pp. 105–133). Information Age Publishing.

Abrams, L. M., McMillan, J. H., & Wetzel, A. P. (2015). Implementing benchmark testing for formative purposes: Teacher voices about what works. *Educational Assessment, Evaluation and Accountability, 27,* 347–375. https://doi.org/10.1007/s11092-015-9214-9

Abrams, L., Varier, D., & Jackson, L. (2016). Unpacking instructional alignment: The influence of teachers' use of assessment data on instruction. *Perspectives in Education, 34*(4), 15–28. http://dx.doi.org/10.18820/2519593X/pie.v34i4.2

Achieve, Inc. (2013). *Proficient vs. prepared: Disparities between state tests and the 2013 National Assessment of Educational Progress (NAEP).* https://issuu.com/achieveinc/docs/naepbrieffinal05141594b2daa338c234/8

Ainsworth, L., & Viegut, D. (2006). *Common formative assessments: How to connect standards-based instruction and assessment*. Corwin.

Allal, L. (2010), Assessment and the regulation of learning. In P. Peterson, E. Baker, & B. McGaw, (Eds.), *International Encyclopedia of Education* (Vol. 3), pp. 348–352. Elsevier.

Allal, L. (2011). Pedagogy, didactics and the co-regulation of learning: A perspective from the French-language world of educational research. *Research Papers in Education, 26*(3), 329–336. http://dx.doi.org/10.1080/02671522.2011.595542

Allen, A., & Smith, R. A. (2022). Curriculum-based measurement. In D. Fisher (Ed.), *Routledge resources online*: *Education*. https://doi.org/10.4324/9781138609877-REE114-1

Alonzo, A. C. (2018). An argument for formative assessment with science learning progressions. *Applied Measurement in Education, 31*(2), 104–112. https://doi.org/10.1080/08957347.2017.1408630

Alonzo, A. C., & Elby, A. (2019). Beyond empirical adequacy: Learning progressions as models and their value for teachers. *Cognition and Instruction, 37*(1), 1–37. https://doi.org/10.1080/07370008.2018.1539735

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261–271. https://doi.org/10.1037/0022-0663.84.3.261

Anderson, L. W. (2018, April). A critique of grading: Policies, practices, and technical matters. *Education Policy Analysis Archives, 26*(49). https://epaa.asu.edu/index.php/epaa/article/view/3814/2053

Andrade, H. L. (2013). Classroom assessment in the context of learning theory and research. In J. H. McMillan (Ed.), *Handbook of research on classroom assessment* (pp. 17–34). Sage.

Andrade, H. L., & Brookhart, S. M. (2020). Classroom assessment as the co-regulation of learning. *Assessment in Education, 27*(4), 350–372. https://doi.org/10.1080/0969594X.2019.1571992

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education, 17*(2), 199–214. http://dex/doi.org/10.1080/09695941003696172

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary students' writing. *Educational Measurement: Issues and Practice, 27*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2008.00118.x

Andrade, H. L., Hefferen, J., & Palma, M. (2019). Formative assessment in the arts. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (pp. 126–145). Routledge Taylor & Francis.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12–19. https://doi.org/10.1080/00405840802577544

Arneson, A., Wihardini, D., & Wilson, M. (2019). Assessing college-ready data-based reasoning. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 93–120). Routledge.

Assessment Reform Group. (2002). *Assessment is for learning: 10 principles*. https://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf

Au, W. (2007). High-stakes testing and curricular control: A qualitative meta-synthesis. *Educational Researcher, 36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Austin Independent School District. (2015). *Student learning objectives (SLOs): Analysis of student growth in 2013–2014, by type and source of assessment*. Department of Research and Evaluation, Austin Independent School District. https://www.austinisd.org/sites/default/files/dre-surveys/DRE_14.85_Analysis_of_Student_Growth_in_2013_2014_by_Type_and_Source_of_Assessment.pdf

Babo, G., Tienken, C. H., & Gencarelli, M. A. (2014) Interim testing, socio-economic status, and the odds of passing Grade 8 state tests in New Jersey. *RMLE Online, 38*(3), 1–9. https://doi.org/10.1080/19404476.2014.11462116

Bailey, A. L., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. *TESOL Quarterly, 48*(3), 480–506. https://doi.org/10.1002/tesq.176

Baird, J-A., Andrich, D., Hopfenbeck, T., & Stobart, G. (2017). Metrology of education. *Assessment in Education, 24*(3), 463–470. https://doi.org/10.1080/0969594X.2017.1337628

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407. https://doi.org/10.1177/0022487108324554

Balloo, K., Evans, C., Hughes, A., Zhu, X., & Winstone, N. (2018). Transparency isn't spoon-feeding: How a transformative approach to the use of explicit assessment criteria can support student self-regulation. *Frontiers in Education, 3*(69). https://doi.org/10.3389/feduc.2018.00069

Bancroft, K. (2010). Implementing the mandate: The limitations of benchmark tests. *Educational Assessment, Evaluation and Accountability, 22,* 53–72. https://doi.org/10.1007/s11092-010-9091-1

Banker, H. J. (1927). The significance of teachers' marks. *The Journal of Educational Research, 16*(3), 159–171. https://doi.org/10.1080/00220671.1927.10879778

Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice, 34*(4), 31–40. https://doi.org/10.1111/emip.12096

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bennett, R. E. (2014). Preparing for the future: What educational assessment must do. *Teachers College Record, 116*(11). https://www.learntechlib.org/p/155766/

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39,* 370–407. https://doi.org/10.3102/0091732X14554179

Bennett, R. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational Measurement: Issues and Practice, 37*(4), 7–15. https://doi.org/10.1111/emip.12231

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). Springer.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Open University Press.

Black, P., McCormick, R., James, M., & Pedder, D. (2006). Learning how to learn and assessment for learning: A theoretical inquiry. *Research Papers in Education, 21*(2), 119–132. https://doi.org/10.1080/02671520600615612

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement*: *Interdisciplinary Research and Perspectives, 9*, 71–123. https://doi.org/10.1080/15366367.2011.591654

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85*(2), 205–225. https://doi.org/10.1080/01619561003685379

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw–Hill.

Boekaerts, M., de Koning, E., & Vedder, P. (2006). Goal-directed behavior and contextual factors in the classroom: An innovative approach to the study of multiple goals. *Educational Psychologist, 41*(1), 33–51. https://doi.org/10.1207/s15326985ep4101_5

Bowers, A. J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration, 47*(5), 609–629. https://doi.org/10.1108/09578230910981080

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation, 17*(3), 141–159. https://dx.doi.org/10.1080.13803611.2011.597112

Box, C., Skoog, G., & Dabbs, J. M. (2015). A case study of teacher personal practice assessment theories and complexities of implementing formative assessment. *American Educational Research Journal, 52*(5), 956–983. https://doi.org.10.3102/0002831215587754

Bradshaw, L., Izsàk, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understanding of rational number: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2–14. https://doi.org/10.1111/emip.12020

Braun, H., Chapman, L., & Vezzu, S. (2010). The Black–White achievement gap revisited. *Education Policy Analysis Archives, 18*(21). http://epaa.asu.edu/ojs/article/view/772

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 293–316). Sense Publishers.

Briggs, D. C., Chattergoon, R., & Burkhardt, A. (2019). Examining the dual purpose use of student learning objectives for classroom assessment and teacher evaluation. *Journal of Educational Measurement, 56*(4), 686–714. https://doi.org/10.1111/jedm.12233

Briggs, D. C., & Furtak, E. M. (2020). Learning progressions and embedded assessment. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 146–169). Routledge.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education,* 7(4), 279–301. https://doi.org/10.1207/s15324818ame0704_2

Brookhart, S. M. (2001). Successful students' formative and summative use of assessment information. *Assessment in Education, 8*(2), 153–169. https://doi.org/10.1080/09695940123775

Brookhart, S. M. (2013). Grading. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 257–271). Sage.

Brookhart, S. M. (2015). Graded achievement, tested achievement, and validity, *Educational Assessment, 20*(4), 268–296. http://dx.doi.org/10.1080/10627197.2015.1093928

Brookhart, S. M. (2017). Formative assessment in teacher education. In D. J. Clandinin & J. Husu (Eds.), *International handbook of research on teacher education* (pp. 927–943). Sage.

Brookhart, S. M. (2018). Summative and formative feedback. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 52–78). Cambridge University Press.

Brookhart, S. M. (2020). Feedback and measurement. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 63–78). Routledge.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (REL 2007–No. 017). U.S. Department of Education. https://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2007017_sum.pdf

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education, 85*(2), 186–204. https://doi.org/10.1080/01619561003685346

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 367–407). Macmillan.

Burch, P. E. (2006). The new educational privatization: Educational contracting and high-stakes accountability. *Teachers College Record, 108*(12), 2582–2610. https://doi.org/10.1111/j.1467-9620.2006.00797.x

Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education, 85*(2), 147–162. https://doi.org/10.1080/01619561003685288

Burkhardt, H., & Schoenfeld, A. (2019). Formative assessment in mathematics. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (pp. 35–67). Routledge.

Butler, D. L., & Schnellert, L. (2015). Success for students with learning disabilities: What does self-regulation have to do with it? In T. Cleary (Ed.), *Self-regulated learning interventions with at-risk youth: Enhancing adaptability, performance, and well-being* (pp. 89–112). APA Press.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Campbell, C. (2013). Research on teacher competency in classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 71–84). Sage.

Campbell, H., Espin, C. A., & McMaster, K. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Reading and Writing, 26*(3), 431–452. https://doi.org/10.1007/s11145-012-9375-6

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International, 44*(1), 57–66. https://doi.org/10.1080/14703290601081332

Carless, D., & Boud, D. (2018) The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*(3), 378–398. https://doi.org/10.3102/0162373711412765

Carter, P. L., & Darling-Hammond, L. (2016). Teaching diverse learners. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 593–638). American Educational Research Association.

Chapman, L. H. (2014). *The marketing of SLOs: 1999–2014*. http://vamboozled.com/laura-chapman-slos-continued/

Chappuis, J. (2022). Student involvement in assessment. In D. Fisher (Ed.), *Routledge resources online*: *Education*. https://doi.org/10.4324/9781138609877-REE8-1

Chojnacki, G., Eno, J., Liu, F., Meyers, C., Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013, September 26–28). *Do interim assessments influence instructional practice in year one? Evidence from Indiana's elementary school teachers* [Paper presentation]. Society for Research on Educational Effectiveness Annual Meeting, Washington, DC, United States. https://files.eric.ed.gov/fulltext/ED563059.pdf

Christman, J. B., Neild, R. C., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessment data: Lessons from Philadelphia*. Research for Action. https://eric.ed.gov/?id=ED505863

Cizek, G. J., Andrade, H. L., & Bennett, R. E. (2019). Formative assessment: History, definition, and progress. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (pp. 3–19). Routledge.

Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence Public Schools* (WCER Working Paper No. 2008–10). Wisconsin Center for Education Research. https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2008_10.pdf

Cochran-Smith, M., & Lytle, S. L. (1999). Relationships of knowledge and practice: Teacher learning in communities. *Review of Research in Education, 24*(1), 249–305. https://doi.org/10.3102/0091732X024001249

Coe, M., Hanita, M. Nishioka, V., & Smiley, R. (2011, December). *An investigation of the impact of the 6+1 Trait Writing Model on grade 5 student writing achievement.* Institute of Education Sciences Report NCEE 2012-4010. https://ies.ed.gov/ncee/edlabs/projects/rct_52.asp?section=ALL

Common Core Standards Writing Team. (2019). *Progressions for the Common Core State Standards for Mathematics (draft February 7, 2019)*. Institute for Mathematics and Education, University of Arizona.

Commonwealth of Australia. (2018, March). *Through growth to achievement: Report of the review to achieve educational excellence in Australian schools*. https://docs.education.gov.au/system/files/doc/other/662684_tgta_accessible_final_0.pdf

Community Training and Assistance Center. (2004). *Catalyst for change: Pay for performance in Denver. Final report*. https://ctacusa.com/wp-content/uploads/2013/11/CatalystForChange.pdf

Community Training and Assistance Center. (2013). *It's more than money: Teacher incentive fund—leadership for educators' advanced performance Charlotte–Mecklenburg Schools*. https://ctacusa.com/wp-content/uploads/2013/11/MoreThanMoney.pdf

Community Training and Assistance Center. (2018). *Student learning objectives (SLOs)*. http://www.ctacusa.com/education/student-learning-objectives-slos/

Confrey, J. (2019, May). *Future of education and skills 2030*: *Curriculum analysis: A synthesis of research on learning trajectories/progressions in mathematics*. Organisation for Economic Co-operation and Development.

Confrey, J., Maloney, A., & Corley, A. K. (2014). Learning trajectories: A framework for connecting standards with curriculum. *ZDM—The International Journal on Mathematics Education, 46*(5), 719–733. https://doi.org/10.1007/s11858-014-0598-7

Confrey, J., & Toutkoushian, E. (2019). A validation approach to middle-grades learning trajectories within a digital learning system applied to the "Measurement of Characteristics of Circles." In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 67–92). Routledge.

Confrey, J., Toutkoushian, E., & Shah, M. (2020). Working at scale to conduct ongoing validation of learning trajectory-based classroom assessments for middle grade mathematics. *Journal of Mathematical Behavior, 60.* https://doi.org/10.1016/j.jmathb.2020.100818

Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report No. RR-63). Consortium for Policy Research in Education. https://files.eric.ed.gov/fulltext/ED506730.pdf

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The impact of the Measures of Academic Progress (MAP) program on student reading achievement* (NCEE 2013–4000). U.S. Department of Education. https://ies.ed.gov/ncee/rel/regions/midwest/pdf/REL_20134000.pdf

Cotton, K. (1988, May). *Close-up #5: Classroom questioning*. Education Northwest. https://educationnorthwest.org/sites/default/files/classroom-questioning.pdf

Council of Chief State School Officers. (2013). *Criteria for procuring and evaluating high-quality assessments*. https://ccsso.org/resource-library/criteria-procuring-and-evaluating-high-quality-assessments

Council of Chief State School Officers. (2018). *Revising the definition of formative assessment*. https://ccsso.org/sites/default/files/2018-06/Revising%20the%20Definition%20of%20Formative%20Assessment.pdf

Covington, M. V. (1992). *Making the grade: A self-worth perspective on school reform*. Cambridge University Press.

Crooks, A. D. (1933). Marks and marking systems: A digest. *Journal of Educational Research, 27*(4), 259–272. https://doi.org/10.1080/00220671.1933.10880402

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481. https://doi.org/10.3102/00346543058004438

Crouse, K., Gitomer, D. H., & Joyce, J. (2016). An analysis of the meaning and use of student learning objectives. In K. K. Hewitt, & A. A. Beardsley (Eds.), *Student growth measures in policy and practice* (pp. 203–222). Springer.

Cushing, E., & Meyer, C. (2014). *Balancing autonomy and comparability: State approaches to assessment selection for student learning objectives. Ask the team*. Center on Great Teachers and Leaders, American Institutes for Research. https://files.eric.ed.gov/fulltext/ED553372.pdf

Dadey, N., & Diggs, C. R. (2019, September). *A rapid review of interim assessment use*. National Center for the Improvement of Educational Assessment. https://www.nciea.org/library/a-rapid-review-of-interim-assessment-use/

Dadey, N., Evans, C. M., & Lorié, W. (2023). Through-year assessment: Ten key considerations. The National Center for the Improvement of Educational Assessment. https://www.nciea.org/wp-content/uploads/2023/03/Ten-Key-Considerations-Through-Year-Assessment-Report-March2023-F.pdf

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. Teachers College Press.

Davidson, K. L., & Frohbieter, G. (2011). *District adoption and implementation of interim and benchmark assessments*. (CRESST Report 806). University of California, National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/R806.pdf

Davis, B. (1997). Listening for differences: An evolving conception of mathematics teaching. *Journal for Research in Mathematics Education, 28*(3), 355–376. https://doi.org/10.5951/jresematheduc.28.3.0355

Deane, P., Sabatini, J., Feng, G., Sparks, J., Song, Y., Fowles, M., O'Reilly, T., Jueds, K., Krovetz, R., & Foley, C. (2015, December). *Key practices in the English Language Arts (ELA): Linking learning theory, assessment, and instruction* (ETS Research Report No. RR-15-17). ETS. http://dx.doi.org/10.1002/ets2.12063

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Bűttner, G., Fauth, B., Hondrich, A. L., Rieser, S., Hertel, S., & Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal, 52*(6), 1133–1159. https://doi.org/10.3102/0002831215596412

DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action in Teacher Education, 34*(5–6), 576–591. https://doi.org/10.1080/01626620.2012.730347

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232. https://doi.org/10.1177/001440298505200303

Deno, S. L. (1993). Curriculum-based measurement. In J. C. Conoley & J. J. Kramer (Eds.), *Curriculum-based measurement—complete work* (pp. 1–23). Buros Institute of Mental Measurements.

Deno, S. L. (2003a). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention, 28*(3–4), 3–12. https://doi.org/10.1177/073724770302800302

Deno, S. L. (2003b). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184–192. https://doi.org/10.1177/00224669030370030801

Deno, S. L. (2013). Problem-solving assessment. In R. Brown-Chidsey & K. J. Andren (Eds.), *Assessment for intervention: A problem-solving approach* (2nd ed., pp. 10–36). Guilford Press.

Deno. S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual.* Minneapolis Leadership Training Institute. University of Minnesota. https://files.eric.ed.gov/fulltext/ED144270.pdf

Denvir, B., & Brown, M. L. (1986). Understanding of number concepts in low-attaining 7–9 year olds: Part 1. Development of descriptive framework and diagnostic instrument. *Educational Studies in Mathematics, 17*(1), 15–36. https://doi.org/10.1007/BF00302376

Diaz-Bilello, E., Burkhardt, A., & Briggs, D. (2016). Findings from case studies of the student learning objectives implementation experiences at four schools. Brief commissioned by the Denver Public Schools. Boulder, CO: Center for Assessment Design Research and Evaluation (CADRE). https://www.colorado.edu/cadre/sites/default/files/attached-files/cadre-brief2.pdf

Diaz-Bilello, E., & Thompson, J. (2019). *Student learning objectives: Where are we now? Synthesis of studies examining the impact of student learning objectives on teaching* *practices* [Unpublished manuscript]. Center for Assessment Design Research and Evaluation (CADRE).

DuFour, R., DuFour, R., Eaker, R., & Karhanek, G. (2010). *Raising the bar and closing the gap.* Solution Tree Press.

Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment, 4*(1), 37–73. https://doi.org/10.1207/s15326977ea0401_2

Duschl, R., Maeng, S. & Sezen, A. (2011) Learning progressions and teaching sequences: a review and analysis, *Studies in Science Education, 47*(2), 123–182. https://doi.org/10.1080/03057267.2011.604476

English, D., Ennis, J., Chamber, D., & Lachlan-Hache, L. (2015). *TIF 4 student learning objective review: Results and recommendations*. Maine Schools for Excellence. http://www.sad44.org/pdf/2015_Maine_SLO_Review_Report.pdf

Espin, C. L., & Wallace, T. (2004). *Descriptive analysis of curriculum-based measurement literature* [Working document]. University of Minnesota Institute for Research on Progress Monitoring.

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & de Rooij, M. (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice, 32*(1), 8–21. https://doi.org/10.1111/ldrp.12123

Espin, C. A., Saab, N., Pat-El, R., Boender, P. D., & van der Veen, J. (2018). Curriculum-based measurement progress data: Effects of graph pattern on ease of interpretation. *Zeitschrift für Erziehungswissenschaft, 21*(4), 767–792. https://doi.org/10.1007/s11618-018-0836-9

Espin, C. A., van den Bosch, R. M., van der Liende, M., Rippe, R. C., Beutick, M., Langa, A., & Mol, S. E. (2021). A systematic review of CBM professional development materials: Are teachers receiving sufficient instruction in data-based decision-making? *Journal of Learning Disabilities, 54*(4), 256–268. https://doi.org/10.1177/0022219421997103

Espin, C., Chung, S., Foegen, A., & Campbell, H. (2018). Curriculum-based measurement for secondary-school students. In P. C. Pullen & M. J. Kennedy (Eds.), *Handbook of response to intervention and multi-tiered systems of support* (pp. 291–315). Routledge.

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat.1177 (2015–2016).

Faltis, C. J., & Valdés, G. (2016). Preparing teachers for teaching in and advocating for linguistically diverse classrooms: A *vade mecum* for teacher educators. In D. H Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 549–592). American Educational Research Association.

Faria, A., Heppen, J., Li, Y., Stachel, S., Jones, W., Sawyer, K., Thomsen, K., Kutner, D., Miser, D., Lewis, S., Casserly, M., Simon, C., Uzzell, R., Corcoran, A., & Palacios, M. (2012, Summer). *Charting success: Data use and student* *achievement in urban schools*. Council of the Great City Schools. https://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Charting_Success.pdf

Farrell, C. C., & Marsh, J. A. (2016). Metrics matter: How properties and perceptions of data shape teachers' instructional responses. *Educational Administration Quarterly, 52*(3), 423–462. https://doi.org/10.1177/0013161X16638429

Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). *New assessments, better instruction? Designing assessment systems to promote instructional improvement*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR354.html

Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. *Assessment and Evaluation in Higher Education, 36*(1), 51–62. https://doi.org/10.1080/02602930903197883

Ferrara, S., Maxey-Moore, K., & Brookhart, S. M. (2019). Guidance in the *Standards* for classroom assessment: Useful or irrelevant? In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and measurement* (pp. 97–119). Routledge.

Foegen, A. (2008). Algebra progress monitoring and interventions for students with learning disabilities. *Learning Disability Quarterly, 31*(2), 65–78. https://doi.org/10.2307/20528818

Forman, E. A., Ramirez-DelToro, V., Brown, L., & Passmore, C. (2017). Discursive strategies that foster an epistemic community for argument in a biology classroom. *Learning and Instruction, 48,* 32–29. https://doi.org/10.1016/j.learninstruc.2016.08.005

Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12*(3), 23–30. https://doi.org/10.1111/j.1745-3992.1993.tb00539.x

Frey, N., & Fisher, D. (2013). Using common formative assessments as a source of professional development in an urban elementary school. *Teaching and Teacher Education, 25*(5), 674–680. https://doi.org/10.1016/j.tate.2008.11.006

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188–192. https://doi.org/10.1080/02796015.2004.12086241

Fuchs, L. (2016). Curriculum-based measurement as the emerging alternative: Three decades later. *Learning Disabilities Research & Practice, 32*(1), 5–7. https://doi.org/10.1111/ldrp.12127

Fuchs, D., & Bradley, R. (2012). A review of Deno and Mirkin's special education resource teacher (SERT) model. In C. A. Espin, K. L. McMaster, & S. Rose (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 27–36). University of Minnesota Press.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*(6), 488–500. https://doi.org/10.1177/001440299105700603

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199–208. https://doi.org/10.1177/001440298605300301

Fuchs, L. S., & Fuchs, D. (2007). A model for implementing responsiveness to intervention. *Teaching Exceptional Children, 39*(5), 14–20. https://doi.org/10.1177/004005990703900503

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children, 73*(3), 311–330. https://doi.org/10.1177/001440290707300303

Fuchs, L. S., Fuchs, D., & Hollenbeck, K. N. (2007). Extending responsiveness to intervention to mathematics at first and third grades. *Learning Disabilities Research & Practice, 22*(1), 13–24. https://doi.org/10.1111/j.1540-5826.2007.00227.x

Fuchs, L. S., Fuchs, D., & Stecker, P. M. (1989). Effects of curriculum-based measurement on teachers' instructional planning. Journal of Learning Disabilities, *22*(1), 51–59. https://doi.org/10.1177/002221948902200110

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching, 49*(9), 1181–1210. https://doi.org/10.1002/tea.21054

Furtak, E. M., Bakeman, R., & Buell, J. Y. (2018). Developing knowledge-in-action with a learning progression: Sequential analysis of teachers' questions and responses to student ideas. *Teaching and Teacher Education, 76*, 267–282. https://doi.org/10.1016/j.tate.2018.06.001

Furtak, E. M., Cartun, A., Chrzanowski, A., Circi, R., Grover, R., Heredia, S. C., Johnson, R., McClelland, A., & White, K. H. O. (2015, April 16–20). *Toward a participation metaphor for formative assessment* [Paper presentation]. American Educational Research Association Annual Meeting, Chicago, IL, United States.

Furtak, E. M., Circi, R., & Heredia, S. C. (2018). Exploring alignment among learning progressions, teacher-designed formative assessment tasks, and student growth: Results of a four-year study. *Applied Measurement in Education, 31*(2), 143–156. https://doi.org/10.1080/08957347.2017.1408624

Furtak, E. M., & Heredia, S. C. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching, 51*(8), 982–1020. https://doi.org/10.1002/tea.21156

Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de Léon, V., Morrison, D., & Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to student learning: Findings from a four-year intervention study. *Instructional Science, 44*, 267–291. https://doi.org/10.1007/s11251-016-9371-3

Furtak, E. M., Morrison, D. L., & Kroog, H. (2014). Investigating the link between learning progressions and classroom assessment. *Science Education, 98*(4), 640–673. https://doi.org/10.1002/sce.21122

Furtak, E. M., & Ruiz-Primo, M. A. (2008). Making students' thinking explicit in writing & discussion: An analysis of formative assessment prompts. *Science Education, 92*(5), 799–824. https://doi.org/10.1002/sce.20270

Furtak, E. Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*(4), 360–389. https://doi.org/10.1080/08957340802347852

Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., D'Mello, S. K., Finn, A. S., & Duckworth, A. L. (2019). Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *American Educational Research Journal, 56*(6), 2077–2115. https://doi.org/10.3102/0002831219843292

Gallimore, R., Ermeling, B. A., Saunders, W. M., & Goldenberg, C. (2009). Moving the learning of teaching closer to practice: Teacher education implications of school-based inquiry teams. *Elementary School Journal, 109*(5), 537–553. https://doi.org/10.1086/597001

Gamlem, S. M., & Smith, K. (2013). Student perceptions of classroom feedback. *Assessment in Education, 20*(2), 150–169. https://doi.org/10.1080/0969594X.2012.749212

Gao, X., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*(4), 323–342. https://doi.org/10.1207/s15324818ame0704_4

Gersten, R., Clarke, B. S., Haymond, K., & Jordan, N. C. (2011). Screening for mathematics difficulties in K–3 students. *WestEd Center on Instruction*. https://files.eric.ed.gov/fulltext/ED524577.pdf

Gesel, S. A., LeJeune, L. M., Chow, J. C., Sinclair, A. C., & Lemons, C. J. (2021). A meta-analysis of the impact of professional development on teachers' knowledge, skill, and self-efficacy in data-based decision-making. *Journal of Learning Disabilities, 54*(4), 269–283. https://doi.org/10.1177/0022219420970196

Gitomer, D. H., & Bell, C. A. (2016). Introduction. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (pp. 1–6). AERA.

Gitomer, D. H., & Duschl, R. A. (1996). Moving toward a portfolio culture in science education. In D. H. Gitomer & R. A. Duschl(Eds.)*, Learning science in the schools* (pp. 299–325). Taylor & Francis.

Gitomer, D. H., & Duschl, R. A. (2007). Indicator systems: establishing multilevel coherence in assessment. *Teachers College Record, 109*(13), 288–320. https://doi.org/10.1177/016146810710901311

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521–563. https://doi.org/10.1080/10508406.2013.837391

Gochyyev, P., & Wilson, M. (2021, September 26–29). *New curriculum efficacy study and Lord's paradox* [Paper presentation]. Society for Research on Educational Effectiveness Annual Meeting, Arlington, VA, United States. https://sree.confex.com/sree/2021/meetingapp.cgi/Paper/2754

Goertz, M. E., Oláh, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction* (CPRE Research Report No. RR-65).

Consortium for Policy Research in Education. https://files.eric.ed.gov/fulltext/ED519791.pdf

Goldin, I., Pavlike, P., & Ritter, S. (2016). Discovering domain models in learning curve data. In R. A. Sottilare, A. C. Graesser, X. Hu, A. Olney, B. D. Nye, & A. M. Sinatra (Eds.), *Design recommendations for intelligent tutoring systems: Vol. 4. Domain modeling* (pp. 115–126). U.S. Army Research Laboratory.

Gong, B. (2021, March 3). *Could two through year assessment designs provide both summative and instructional information: An exploration of feasibility for a through year assessment system*. National Center for the Improvement of Educational Assessment. https://www.nciea.org/blog/state-testing/could-two-through-year-assessment-designs-provide-both-summative-and

Goral, D. P., & Bailey, A. L. (2019). Student self-assessment of oral explanations: Use of language learning progressions. *Language Testing, 36*(3), 391–418. https://doi.org/10.1177/0265532219826330

Graf, E. A., & van Rijn, P. W. (2016). Learning progressions as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 165–189). Taylor & Francis.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523–547. https://www.journals.uchicago.edu/doi/10.1086/681947

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Corwin.

Hadwin, A. F., Järvelä, S., and Miller, M. (2011). Self-regulated, co-regulated, and socially shared regulation of learning." In D. Schunk & B. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (pp. 65–84). Routledge.

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of EducationalResearch,55*(1),23–46.https://doi.org/10.3102/00346543055001023

Hafner,J.C.,&Hafner,P.M.(2003).Quantitativeanalysisoftherubricasanassessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509–1528. https://doi.org/10.1080/0950069022000038268

Hall, E., Gagnon, D., Schneider, C. M., Thompson, J., & Marion, S. (2014). *State practices related to the use of student achievement measures in the evaluation of teachers in non-tested subjects and grades*. National Center for the Improvement of Educational Assessment. https://www.nciea.org/sites/default/files/publications/Gates_NTGS_Hall_082614.pdf

Hamilton, A., Wiliam, D., & Hattie, J. (2023, August). The future of AI in education: 13 things we can do to minimize the damage [Working paper]. https://edarxiv.org/372vr/

Hanushek, E. A. (2011). Valuing teachers: How much is a good teacher worth? *Education Next, 11*(3), 41–45. https://www.educationnext.org/valuing-teachers/

Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2019). The achievement gap fails to close. *Education Next, 19*(3), 8–17. https://www.educationnext.org/achievement-gap-fails-close-half-century-testing-shows-persistent-divide/

Harlen, W. (2005). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (pp. 103–117). Sage.

Harris, L. R., Brown, G. T. L., & Harnett, J. A. (2014). Understanding classroom feedback practices: A study of New Zealand student experiences, perceptions, and emotional responses. *Educational Assessment, Evaluation, and Accountability, 26*, 107–133. https://doi.org/10.1007/s11092-013-9187-5

Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancies effects: 31 meta-analyses. *Psychological Bulletin, 97*, 363–386. https://doi.org/10.1037/0033-2909.97.3.363

Hattie, J. (2009). *Visible learning*. Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (REL 2007–No. 039). U.S. Department of Education. https://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2007039_sum.pdf

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). *A second follow-up year for "Measuring how benchmark assessments affect student achievement"* (REL 2008–No. 002). U.S. Department of Education. https://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/techbrief/tr_00208.pdf

Heppen, J., Faria, A., Thomsen, K., Sawyer, K., Townsend, M., Kutner, M., Stachel, S., Lewis, S., & Casserly, M. (2010, December). *Using data to improve instruction in the Great City Schools: Key dimensions of practice*. Council of the Great City Schools and American Institutes for Research. https://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Strand%202%20Report%20-%20Key%20Dimensions%20of%20Data%20Use_122110.pdf

Heppen, J., Jones, W., Faria, A., Sawyer, K., Lewis, S., Horowitz, A., Simon, C., Uzzell, R., & Casserly, M. (2011, January). *Using data to improve instruction in the Great City Schools: Documenting current practice*. Council of the Great City Schools and American Institutes for Research. Strand 2 Report - Documenting Current Practice.pdf

Heredia, S. C., Furtak, E. M., Morrison, D., & Renga, I. P. (2016). Science teachers' representations of classroom practice in the process of formative assessment design. *Journal of Science Teacher Education, 27*, 697–716. https://doi.org/10.1007/s10972-016-9482-3

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment.* Council of Chief State School Officers. https://csaa.wested.org/resource/learning-progressions-supporting-instruction-and-formative-assessment/

Heritage, M., & Bailey, A. L. (2006). Assessing to teach: An introduction. *Educational Assessment, 11*(3–4), 145–148. https://doi.org/10.1080/10627197.2006.9652987

Heritage, M., & Heritage, J. (2013). Teacher questioning: The epicenter of instruction and assessment. *Applied Measurement in Education, 26*(3), 176–190. https://doi.org/10.1080/08957347.2013.793190

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice, 28*(3), 24–31. https://doi.org/10.1111/j.1745-3992.2009.00151.x

Herman, J., Epstein, S., Leon, S., La Torre Matrundola, D., Reber, S., & Choi, K. (2015). *Implementation and effects of LDC and MDC in Kentucky districts* (CRESST Policy Brief No. 13). University of California, National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/PB_13.pdf

Hess, K. K., & Kearns, J. (2010, December). *Learning progression frameworks designed for use with the Common Core State Standards in Mathematics K–12*. National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment. (updated, v.3) https://www.nciea.org/wp-content/uploads/2022/07/Math_LPF_KH11.pdf

Hess, K. K., & Kearns, J. (2011, December). *Learning progression frameworks designed for use with the Common Core State Standards in English Language Arts & Literacy K–12*. National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment. http://www.naacpartners.org/publications/ELA_LPF_12.2011_final.pdf

Higgins, K. M., Harris, N. A., & Kuehn, L. L. (1994). Placing assessment into the hands of young children: A study of student-generated criteria and self-assessment. *Educational Assessment, 2*(4), 309–324. https://doi.org/10.1207/s15326977ea0204_3

Hosp, J. L., & Hosp, M. (2012). When the emerging alternative became the standard. In C. A. Espin, K. L. McMaster, S. Rose, & M. Miura Wayman (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 49–56). University of Minnesota Press.

Immekus, J. C., & Atitya, B. (2016). The predictive validity of interim assessment scores based on the full-information bi-factor model for the prediction of end-of-grade test performance. *Educational Assessment, 21*(3), 176–195. https://doi.org/10.1080/10627197.2016.1202108

Ing, M., Chinen, S., Jackson, K., & Smith, T. M. (2021). When should I use a measure to support instructional improvement at scale? The importance of considering both intended and actual use in validity arguments. *Educational Measurement: Issues and Practice, 40*(1), 92–100. https://doi.org/10.1111/emip.12393

James, M. (2017). (Re)viewing assessment: Changing lenses to refocus on learning. *Assessment in Education, 24*(3), 404–414. https://doi.org/10.1080/0969594X.2017.1318823

James, M., & Pedder, D. (2006). Beyond method: Assessment and learning practices and values. *The Curriculum Journal, 17*(2), 109–138. https://doi.org/10.1080/09585170600792712

Jenkins, J., & Fuchs, L. S. (2012). Curriculum-based measurement: The paradigm, history, and legacy. In C. A. Espin, K. L. McMaster, S. Rose, & M. Miura Wayman

(Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 7–23). University of Minnesota Press.

Johnson, C. C., Sondergeld, T. A., & Walton, J. B. (2019). A study of the implementation of formative assessment in three large urban districts. *American Educational Research Journal, 56*(6), 2408–2438. https://doi.org/10.3102/0002831219842347

Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education, 14*(1), 63–76. https://doi.org/10.1177/1469787412467125

Jonsson, A., & Panadero, E. (2018). Facilitating students' active engagement with feedback. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 531–553). Cambridge University Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Jung, P. G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-analysis. *Learning Disabilities Research & Practice, 33*(3), 144–155. https://doi.org/10.1111/ldrp.12172

Ker, H. W. (2016). The impacts of student-, teacher-and school-level factors on mathematics achievement: an exploratory comparative investigation of Singaporean students and the USA students. *Educational Psychology, 36*(2), 254–276. https://doi.org/10.1080/01443410.2015.1026801

Kingston, N. M., & Broaddus, A. (2017). The use of learning map systems to support the formative assessment in mathematics. *Education Sciences, 7*(1), 41. https://doi.org/10.3390/educsci7010041

Kingston, N. M., Karvonen, M., Bechard, S., & Erickson, K. A. (2016). The philosophical underpinnings and key feathers of the Dynamic Learning Maps alternate assessment. *Teachers College Record, 118*(14), pp. 1–30. https://doi.org/10.1177/016146811611801410

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37. [Erratum (2015), *34*(2), 55.] https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kirschenbaum, H., Napier, R., & Simon, S. B. (1971). *Wad-ja-get? The grading game in American education*. Hart.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A review of the evidence* (REL 2017–259). U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED572929.pdf

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2019): A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback. *Educational Psychology, 41*(7), 922–947. https://doi.org/10.1080/01443410.2019.1659939

Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2018). *Automated student model improvement*. In Proceedings of the 5th International Conference on Educational Data Mining. EDM. https://files.eric.ed.gov/fulltext/ED537201.pdf

Konstantopoulos, S., Li, W., Miller S. R, & van der Ploeg, A. (2016). Effects of interim assessments across the achievement distribution: Evidence from an experiment. *Educational and Psychological Measurement, 76*(6), 587–608. https://doi.org/10.1177/0013164415606498

Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). The impact of Indiana's system of diagnostic assessments on mathematics achievement. *Educational Evaluation and Policy Analysis, 35*(4), 481–499. https://doi.org/10.3102/0162373713498930

Konstantopoulos, S., Miller, S. R., van der Ploeg, A., & Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness, 9*(Suppl. 1), 188–208. https://doi.org/10.1080/19345747.2015.1116031

Kopriva, R. J., Thurlow, M. L., Perie, M., Lazarus, S. S., & Clark, A. (2016) Test takers and the validity of score interpretations. *Educational Psychologist, 51*(1), 108–128. https://doi.org/10.1080/00461520.2016.1158111

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. University of Chicago Press.

Kroog, H. I., Ruiz-Primo, M. A., & Sands, D. (2014, April 3–7). *Understanding the interplay between the cultural context of classrooms and formative assessment* [Paper presentation]. American Educational Research Association Annual Meeting, Philadelphia, PA, United States.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*(2), 211–232. https://doi.org/10.3102/00346543047002211

Lachlan-Hache, L., Matlach, L., Guiden, A., & Castro, M. (2015). *What we know about SLOs: An annotated bibliography of research on and evaluations of student learning objectives*. American Institutes for Research. https://www.air.org/resource/what-we-know-about-slos-annotated-bibliography-research-and-evaluations-student-learning

Lam, E. A., McMaster, K. L., & Rose, S. (2020). Systematic review of curriculum-based measurement with students who are deaf. *The Journal of Deaf Studies and Deaf Education, 25*(4), 398–410. https://doi.org/10.1093/deafed/enaa020

Lane, S., & DePascale, C. (2016). *Psychometric considerations for performance-based assessments and student learning objectives. Meeting the challenges to measurement in an era of accountability*. Routledge. https://doi.org/10.4324/9780203781302

Lee, E., & Lee, S. (2009). Effects of instructional rubrics on class engagement behaviors and the achievement of lesson objectives by students with mild mental retardation and their typical peers. *Education and Training in Developmental Disabilities, 44*(3), 396–408. https://www.jstor.org/stable/24233483

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K–12 education: A systematic review. *Applied Measurement in Education, 33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Lehrer, R. (2021, August 16–20). *Accountable assessment* [Keynote address]. ACER 2021 Research Conference. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1009&context=rc21-30

Lehrer, R., Kim, M.-J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. Maloney, J. Confrey, & K. Nguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. 31–60). Information Age Publishers.

Leighton, J. P. (2019). Cognitive diagnosis is not enough: The challenge of measuring learning with classroom assessments. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 27–45). Routledge.

Leighton, J. P., Chu, M.-W., & Seitz, P. (2013). Errors in student learning and assessment: The learning errors and formative feedback (LEAFF) model. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment* (pp. 185–208). Information Age.

Lekholm, A. K., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: Effects of gender and family background. *Educational Research and Evaluation, 14*(2), 181–199. https://doi.org/10.1080/13803610801956663

Lipnevich, A. A., & Smith, J. K. (2009a). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied, 15*(4), 319–333. https://doi.org/10.1037/a0017841

Lipnevich, A. A., & Smith, J. K. (2009b). "I really need feedback to learn:" Students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability, 21*, 347–367. https://doi.org/10.1007/s11092-009-9082-2

Liu, L., Hao, J., Andrews, J. J., Zhu, M., Mislevy, R. J., Kyllonen, P., von Davier, A. A., Kerr, D., Ricarte, T., & Graesser, A. (2017). *Collaborative problem solving: Innovating standardized assessment*. International Society of the Learning Sciences. https://doi.org/10.22318/CSCL2017.118

Liu, R., & Koedinger, K. R. (2017). Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining, 9*(1), 25–41. https://doi.org/10.5281/zenodo.3554625

Lohbeck, A. (2019). Social and dimensional comparison effects on academic self-concepts and self-perceptions of effort in elementary school children. *Educational Psychology, 39*(1), 133–150. https://doi.org/10.1080/01443410.2018.1527018

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing*. Information Age Publishing.

Marion, S. F., & Buckley, K. (2012). *Approaches and considerations for incorporating student performance results from "non-tested" grades and subjects into educator effectiveness determinations*. The National Center for the Improvement of Educational

Assessment. https://www.nciea.org/sites/default/files/publications/Marion_Buckley_Considerations_for_non-tested_grades_2011.pdf

Marion, S. F., DePascale, C., Domaleski, C., Gong, B., & Diaz-Bilello, E. (2012). *Considerations for analyzing educators' contributions to student learning in non-tested subjects and grades with a focus on student learning objectives*. The National Center for the Improvement of Educational Assessment. https://www.nciea.org/sites/default/files/publications/Measurement%20Considerations%20for%20NTSG_052212%20v2.pdf

Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record, 114*(11), 1–48. https://doi.org/10.1177/016146811211401106

Marso, R. N., & Pigge, F. L. (1993). Teachers' testing knowledge, skills, and practices. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 129–185). Buros Institute of Mental Measurements, University of Nebraska.

Marston, D. (2012). School- and district-wide implementation of curriculum-based measurement in the Minneapolis public schools. In C. A. Espin, K. L. McMaster, & S. Rose (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 59–78). University of Minnesota Press.

Martone, A., Reagan, D., & Reed, G. (2018). Understanding the use of mathematics interim assessments: A case study. *International Electronic Journal of Elementary Education, 10*(5), 515–523. https://www.iejee.com/index.php/IEJEE/article/view/318

Mason, B. J., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us*. University of Nebraska–Lincoln. http://dwb.unl.edu/Edit/MB/MasonBruning.html

Massachusetts Department of Education. (1999). *1999 MCAS technical report*. https://archive.org/details/ERIC_ED459184/page/n1

McBee, M. M., & Barnes, L. L. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education, 11*(2), 179–194. https://doi.org/10.1207/s15324818ame1102_4

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education, 41*(2), 68–84. https://doi.org/10.1177/00224669070410020301

McMaster, K. L., Ritchey, K. D., & Lembke, E. (2011). Curriculum-based measurement for beginning writers: Recent developments and future directions. *Assessment and Intervention, 24,* 111–148. https://doi.org/10.1108/S0735-004X(2011)0000024008

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20–32. https://doi.org/10.1111/j.1745-3992.2001.tb00055.x

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *Journal of Educational Research, 95*(4), 203–213. https://doi.org/10.1080/00220670209596593

Michaels, S., Shouse, A. W., & Schweingruber, H. A. (2008). *Ready, set, science!* National Academies Press.

Michigan Assessment Consortium. (2017, Fall). *Assessment literacy standards*. Michigan Assessment Consortium. https://www.michiganassessmentconsortium.org/assessment-literacy-standards/

Miner, B. C. (1967). Three factors of school achievement. *The Journal of Educational Research, 60*(8), 370–376. https://doi.org/10.1080/00220671.1967.10883518

Minstrell, J., Anderson, R., & Li, M. (2010). *Assessing teacher competency in formative assessment*: *Annual report 2010*. National Science Foundation.

Mislevy, R. J. (2018a). On integrating psychometrics and learning analytics in complex assessments. In H. Jiao & R. W. Lissitz (Eds.), *Test data analytics and psychometrics*: *Informing assessment practices* (pp. 1–52). Information Age.

Mislevy, R. J. (2018b). *Sociocognitive foundations of educational measurement*. Routledge.

Missouri Department of Education. (2015). *Setting growth targets for student learning objectives: Methods and considerations*. https://dese.mo.gov/sites/default/files/Methods-and-Considerations.pdf

Miura Wayman, M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85–120. https://doi.org/10.1177/00224669070410020401

Moore, C. C. (1939). The elementary school mark. *The Pedagogical Seminary and Journal of Genetic Psychology, 54*, 285–294.

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching, 54*(8), 1024–1048. https://doi.org/10.1002/tea.21397

Morton, C. (2013). *Judging alignment of curriculum-based measures in mathematics and common core standards* [Unpublished doctoral dissertation, University of Oregon]. https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/17879/Morton_oregon_0171A_10841.pdf?sequence=1&isAllowed=y

Mosher, F. A. (2022). Learning progressions. In D. Fisher (Ed.), *Routledge resources online*: *Education*. https://doi.org/10.4324/9781138609877-REE115-1

Myers, C. B. (1996, April 8–12). *Beyond the PDS: Schools as professional learning communities: A proposal based on an analysis of PDS efforts of the 1990s* [Paper presentation]. American Educational Research Association Annual Meeting, New York, NY, United States.

National Center on Response to Intervention. (March 2010). *Essential components of RTI—a closer look at Response to Intervention*. U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED526858.pdf

National Centre for Excellence in the Teaching of Mathematics. (2020). *Mathematics guidance: Key stages 1 and 2: Non-statutory guidance for the national curriculum in England*. UK Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1017683/Maths_guidance_KS_1_and_2.pdf

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K–8*. National Academies Press. https://nap.nationalacademies.org/read/11625/chapter/1

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*(2), 155–175. https://doi.org/10.1207/s15326985ep2202_4

New York State Education Department. (2014, March). *Guidance on the New York State district-wide growth goal-setting process for teachers*: *Student learning objectives*.

Newton, P. E. (2010). The multiple purposes of assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed, pp. 392–396). Elsevier.

No Child Left Behind Act of 2001, P.L. 107–110, 20 U.S.C. § 6319 (2002).

Northwest Evaluation Association. (2019, March). *MAP® Growth™ technical report*.

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education, 85*(2), 226–245. https://doi.org/10.1080/01619561003688688

Olsen, B., & Buchanan, R. (2019). An investigation of teachers encouraged to reform grading practices in secondary schools. *American Educational Research Journal, 56*(5), 2004–2039. https://doi.org/10.3102/0002831219841349

Opper, I. M. (2019). *Teachers matter: Understanding teachers' impact on student achievement*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR4312.html

Otero, V. K. (2006). Moving beyond the "get it or don't" conception of formative assessment. *Journal of Teacher Education, 57*(3), 247–255. https://doi.org/10.1177/0022487105285963

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences, 22*(6), 806–813. https://doi.org/10.1016/j.lindif.2012.04.007

Parsi, A., & Darling-Hammond, L. (2015). *Performance assessments: How state policy can advance assessments for 21st century learning*. National Association of State Boards of Education, Stanford Center for Opportunity Policy in Education. https://pdfs.semanticscholar.org/d479/05d18b23dd7eb367c0627656e3e7f321567b.pdf

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. National Board on Educational Testing and Public Policy. https://www.bc.edu/research/nbetpp/statements/nbr2.pdf

Pellegrino, J. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65–77. https://doi.org/10.1016/j.pse.2014.11.002

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 787–850). American Educational Research Association.

Pereira, M., & Tienken, C. (2012). An evaluation of the influence of interim assessments on Grade 8 student achievement in mathematics and language arts. *International Journal of Educational Leadership Preparation, 7*(3), 1–13. https://files.eric.ed.gov/fulltext/EJ997471.pdf

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5–13. https://doi.org/10.1111/j.1745-3992.2009.00149.x

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes: Towards a wider conceptual field. *Assessment in Education, 5*(1), 85–102. https://doi.org/10.1080/0969595980050105

Peters, R., Kruse, J., Buckmiller, T., & Townsley, M. (2017). "It's just not fair!" Making sense of secondary students' resistance to a standards-based grading. *American Secondary Education, 45*(3), 9–28. https://www.proquest.com/docview/1938071902

Peterson, P. (1979). Direct instruction reconsidered. In P. Peterson & H. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 57–65). McCutchan.

Pham, D. N., Wells, C. S., Bauer, M. I., Wylie, C., & Monroe, S. (2021). Examining three learning progressions in middle-school mathematics for formative assessment. *Applied Measurement in Education, 34*(2), 107–121. https://doi.org/10.1080/08957347.2021.1890744

Pitt, E., & Norton, L. (2017). "Now that's the feedback I want!" Students' reactions to feedback on graded work and what they do with it. *Assessment and Evaluation in Higher Education, 42*(4), 499–516. https://doi.org/10.1080/02602938.2016.1142500

Pokorny, H., & Pickford, P. (2010). Complexity, cues, and relationships: Student perceptions of feedback. *Active Learning in Higher Education, 11*(1), 21–30. https://doi.org/10.1177/1469787409355872

Pollio, M., & Hochbein, C. (2015). The association between standards-based grading and standardized test scores as an element of a high school reform model. *Teachers College Record, 117*(11), 1–28. https://doi.org/10.1177/016146811511701106

Porter, A. C. (1993). School delivery standards. *Educational Researcher, 22*(5), 24–30. https://doi.org/10.3102/0013189X022005024

Powell, D., Lamba, S., Ismail, K., & Marland, J. (2022). *What are through-year assessment? Exploring multiple approaches to through-year design*. Education First. https://www.education-first.com/wp-content/uploads/2023/01/What-are-Through-year-Assessments-1.pdf

Randall, J., & Engelhard, G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement, 46*(1), 1–18. https://doi.org/10.1111/j.1745-3984.2009.01066.x

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

Randel, B., Apthorp, H., Beesley, A. D., Clark, T. F., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *Journal of Educational Research, 109*(5), 491–502. https://doi.org/10.1080/00220671.2014.992581

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435–448. https://doi.org/10.1080/02602930902862859

Resnick, L. B., & Resnick, D. P. (1992) Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Evaluation in education and human services* (Vol. 30, pp. 37–75). Springer. https://doi.org/10.1007/978-94-011-2968-8_3

Rhoades, K., & Madaus, G. (2003). *Errors in standardized tests: A systemic problem*. National Board on Educational Testing and Public Policy. https://files.eric.ed.gov/fulltext/ED479797.pdf

Riggan, M., & Oláh, L. N. (2011). Locating interim assessments within teachers' assessment practice. *Educational Assessment, 16*(1), 1–14. https://doi.org/10.1080/10627197.2011.551085

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249–255. https://doi.org/10.3758/BF03194060

Rosenshine, B. V. (1979). Content, time, and direct instruction. In P. Peterson & H. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 28–56). McCutchan.

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in Grade 5–6 mathematics: Effects on problem-solving achievement. *Educational Assessment, 8*(1), 43–58. https://doi.org/10.1207/S15326977EA0801_03

Ross, J. A., & Starling, M. (2008). Self-assessment in a technology-supported environment: The case of Grade 9 geography. *Assessment in Education, 15*(2), 183–199. https://doi.org/10.1080/09695940802164218

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. Jossey–Bass.

Rowe, K. J., & Hill, P. W. (1996) Assessing, recording and reporting students' educational progress: The case for "subject profiles." *Assessment in Education, 3*(3), 309–352. https://doi.org/10.1080/0969594960030304

Ruiz-Primo, M. A., & Brookhart, S. M. (2018). *Using feedback to improve learning*. Routledge.

Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment, 11*(3–4), 205–235. https://doi.org/10.1080/10627197.2006.9652991

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Informal formative assessment and scientific inquiry: Exploring teachers' practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84. https://doi.org/10.1002/tea.20163

Ruiz-Primo, M. A., Kroog, H. I., & Sands, D. I. (2015, August 25–29). *Teacher judgments on-the-fly: Teachers' response patterns in the context of informal formative assessment* [Paper presentation]. EARLI Biennial Conference, Limassol, Cyprus.

Ruiz-Primo, M. A., & Li, M. (2013). Analyzing teachers' feedback practices in response to students' work in science classrooms. *Applied Measurement in Education, 26*(3), 163–175. https://doi.org/10.1080/08957347.2013.793188

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sadler, D. R. (2014). The futility of attempting to codify academic achievement standards. *Higher Education, 67*(3), 273–288. https://doi.org/10.1007/s10734-013-9649-1

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049–1066. https://doi.org/10.1037/edu0000190

Scalise, K., & Wilson, M. (2011). The nature of assessment systems to support effective use of evidence through technology. *E–Learning and Digital Media, 8*(2), 121–132. https://doi.org/10.2304/elea.2011.8.2.121

Schneider, M. C., & Andrade, H. (2013). Teachers' and administrators' use of evidence of student learning to take action: Conclusions drawn from a special issue on formative assessment. *Applied Measurement in Education, 26*(3), 159–162. https://doi.org/10.1080/08957347.2013.793189

Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning. *Applied Measurement in Education, 26*(3), 191–204. https://doi.org/10.1080/08957347.2013.793185

Schneider, M. C., & Johnson, R. L. (2018). *Using formative assessment to support student learning objectives*. Routledge.

Schwartz, R., Ayers, E., & Wilson, M. (2017). Mapping a data modeling and statistical reasoning learning progression using unidimensional and multidimensional item response models. *Journal of Applied Measurement, 18*(3), 268–298.

Shavelson, R. J. (2009, June 24–26). *Reflections on learning progressions* [Paper presentation]. Learning Progressions in Science Conference, Iowa City, IA, United States.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M. K., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295–314. https://doi.org/10.1080/08957340802347647

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*(7), 2–16. https://doi.org/10.3102/0013189X020007002

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14. https://doi.org/10.3102/0013189X029007004

Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education, 85*(2), 246–257. https://doi.org/10.1080/01619561003708445

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education, 31*(2), 165–174. https://doi.org/10.1080/08957347.2017.1408628

Shepard, L. A., Penuel, W. R., & Davidson, K. L. (2017). Design principles for new systems of assessment. *Phi Delta Kappan, 98*(6), 47–52. https://doi.org/10.1177/0031721717696478

Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology, 73,* 131–149. https://doi.org/10.1016/j.jsp.2019.03.005

Shriner, J. G., & Thurlow, M. L. (2012). Curriculum-based measurement, progress monitoring, and state assessments. In C.A. Espin, K.L. McMaster, & S. Rose (Eds.), *A measure of success: The influence of curriculum-based measurement on education,* (pp. 247–260). University of Minnesota Press.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Simon, M. A. (1995) Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114–145. https://doi.org/10.5951/jresematheduc.26.2.0114

Smarter Balanced Assessment Consortium. (2019). *2020–2021 Interim assessments overview*. https://portal.smarterbalanced.org/library/en/interim-assessments-overview.pdf

Smith, A. Z., & Dobbin, J. E. (1960). Marks and marking systems. In C. W. Harris (Ed.), *Encyclopedia of educational research* (3rd ed., pp. 783–791). Macmillan.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Solberg, L. I., Mosser, G., & McDonald, S. (1997). The three faces of performance measurement: improvement, accountability, and research. *The Joint Commission journal on quality improvement, 23*(3), 135-147. https://doi.org/10.1016/s1070-3241(16)30305-4

Starch, D. (1913). Reliability and distribution of grades. *Science, 38*(983), 630–636. https://doi.org/10.1126/science.38.983.630

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795–819. https://doi.org/10.1002/pits.20113

Stiggins, R. J. (1994). *Student-centered classroom assessment*. Merrill.

Stiggins, R. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324–328. https://doi.org/10.1177/003172170508700414

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5–14. https://doi.org/10.1111/j.1745-3992.1989.tb00315.x

Stoeffler, K., Rosen, Y., Bolsinova, M, & von Davier, A. A. (2020). Gamified performance of collaborative problem solving skills. *Computers in Human Behavior, 104*(106036). https://doi.org/10.1016/j.chb.2019.05.033

Takahashi, S., Jackson, K., Norman, J. R., Ing, M., & Krumm, A. E. (2022). Measurement for improvement. In D. Peurach, J. Russell, L. Cohen-Vogel, & W. R. Penuel (Eds.), *The foundational handbook on improvement research in education* (pp. 423–442). Rowman & Littlefield.

Taylor, J. J., Buckley, K., Hamilton L. S., Stecher, B. M., Read, L., & Schweig, J. (2018). *Choosing and using SEL competency assessments: What schools and districts need to know*. Collaborative for Academic, Social, and Emotional Learning. http://measuringsel.casel.org/pdf/

Templin, J., & Bradshaw, L. (2014). The use and misuse of psychometric models. *Psychometrika, 79*(2), 347–354. https://doi.org/10.1007/S11336-013-9364-Y

Texas Education Agency. (2021, July). *Student learning objectives implementation guide for administrators*. https://texasslo.org/Resource_files/resources/SLO_Administrator_Guide_w_TIA_Tips.pdf

Thomas, S., & Oldfather, P. (1997). Intrinsic motivations, literacy, and assessment practices: "That's my grade. That's me." *Educational Psychologist, 32*(2), 107–123. https://doi.org/10.1207/s15326985ep3202_5

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education*, 2013(958530), 1–29. http://dx.doi.org/10.1155/2013/958530

Topping, K. J. (2013). Peers as a source of formative and summative assessment. In. J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 395–412). Sage.

Tunstall, P., & Gipps, C., (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal, 22*(4), 389–404. https://doi.org/10.1080/0141192960220402

Tyler, J., & McNamara, C. (2011, Fall). *An examination of teacher use of the data dashboard student information system in Cincinnati Public Schools,* Vol. VI. Council of the Great City Schools. https://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/85/SUERF%20VI.pdf

U.S. Department of Education. (2015, October). *Testing action plan*. https://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan

U.S. Department of Education, National Assessment of Educational Progress. (2020). *1990–2019 mathematics and reading assessments*. https://www.nationsreportcard.gov/dashboards/achievement_gaps.aspx

U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2009). *Implementing data-informed decision making in schools: Teacher access, supports and use*. https://www2.ed.gov/about/offices/list/opepd/ppss/reports.html

U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2011). *Teachers' ability to use data to inform instruction*: *Challenges and supports*. https://www2.ed.gov/rschstat/eval/data-to-inform-instruction/report.pdf

van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice, 32*(1), 46–60. https://doi.org/10.1111/ldrp.12122

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Van der Kleij, F. M., & Lipnevich, A. A. (2021). Student perceptions of assessment feedback: A critical scoping review and call for research. *Educational Assessment, Evaluation and Accountability, 33*(2), 345–373. https://doi.org/10.1007/s11092-020-09331-x

Vanlommel, K., & Schildkamp, K. (2019). How do teachers make sense of data in the context of high-stakes decision making? *American Educational Research Journal, 56*(3), 792–821. https://doi.org/10.3102/0002831218803891

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

von Davier, A. A. (2017). Computational psychometrics in support of collaborative assessments. *Journal of Educational Measurement, 54*(1), 3–11. https://doi.org/10.1111/jedm.12129

von Davier, A. A., DiCerbo, K., & Verhagen, J. (2021). Computational psychometrics: A framework for estimating learners' knowledge, skills and abilities from learning and assessments systems. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python* (pp. 25–43). Springer.

von Davier, A. A., Wong, P., Polyak, S. T., & Yudelson, M. (2019). The argument for a "Data Cube" for large-scale psychometric data. *Frontiers in Education,* 4(71). http://doi.org/10.3389/feduc.2019.00071

Waltman, K. K., & Frisbie, D. A. (1994). Parents' understanding of their children's report card grades. *Applied Measurement in Education, 7*(3), 223–240. https://doi.org/10.1207/s15324818ame0703_5

Welsh, M. E., D'Agostino, J. V., & Kaniskan, R. (2013). Grading as a reform effort: Do standards-based grades converge with test scores? *Educational Measurement: Issues and Practice, 32*(2), 26–36. https://doi.org/10.1111/emip.12009

Westrick, P. A. (2017). Reliability estimates for undergraduate grade point average. *Educational Assessment, 22*(4), 231–252. https://doi.org/10.1080/10627197.2017.1381554

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT scores, high school grades, and SES. *Educational Assessment, 20*(1), 23–45. https://doi.org/10.1080/10627197.2015.997614

Wiggins, G. (1998). *Educative assessment: Designing assessment to inform and improve student performance*. Jossey–Bass..

Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 18–40). Routledge.

Wiliam, D. (2018a). *Embedded formative assessment* (2nd ed.). Solution Tree.

Wiliam, D. (2018b). How can assessment support learning? A response to Wilson and Shepard, Penuel, and Pellegrino. *Educational Measurement: Issues and Practice, 37*(1), 42–44. https://doi.org/10.1111/emip.12192

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*(1), 49–65. https://doi.org/10.1080/0969594042000208994

Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Routledge. https://doi.org/10.4324/9781315086545

Wilkerson, S. B., Klute, M., Peery, B., & Liu, J. (2021). *How Nebraska teachers use and perceive summative, interim, and formative data* (REL 2021–054). U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED610179.pdf

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39*(1), 1–37. https://doi.org/10.1111/j.1745-3984.2002.tb01133.x

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730. https://doi.org/10.1002/tea.20318

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. Alonzo & A. Gotwals (Eds.), *Learning progressions in science* (pp. 317–343). Sense Publishers.

Wilson, M. (2021, August 16–20). *Rethinking measurement for accountable assessment* [Keynote address]. ACER 2021 Research Conference. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1015&context=rc21-30

Wilson, M. (2023). Finding the right grain-size for measurement in the classroom. *Journal of Education and Behavioral Statistics, 49*(1), 3–31. https://doi.org/10.3102/10769986231159006

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education, Part II* (pp. 132–154). University of Chicago Press.

Wilson, M., Gochyyev, P., & Scalise, K. (2016). Assessment of learning in digital interactive social networks: A learning analytics approach. *Online Learning, 20*(2), 97–119. https://olj.onlinelearningconsortium.org/index.php/olj/article/view/799/205

Wilson, M., & Lehrer, R. (2021). Improving learning: Using a learning progression to coordinate instruction and assessment. *Frontiers in Education, 6*, 654212. https://doi.org/10.3389/feduc.2021.654212

Wilson, M., & Scalise, K. (2016). Learning analytics: Negotiating the intersection of measurement technology and information technology. In J. M. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1–23). Springer.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208. https://doi.org/10.1207/S15324818AME1302_4

Winstone, N. E., Mathlin, G., & Nash, R. A. (2019). Building feedback literacy: Students' perceptions of the Developing Engagement with Feedback Toolkit. *Frontiers in Education, 4*(39). https://doi.org/10.3389/feduc.2019.00039

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017) Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Wong J., Baars, M., de Koning, B. B., van der Zee, T., Davis, D., Khalil, M., Houben, G.-J., & Paas, F. (2019) Educational theories and learning analytics: From data to knowledge. In D. Ifenthaler, D. K. Mah, &J. K. Yau (Eds.), *Utilizing learning analytics to support study success* (pp. 3–25). Springer.

Wylie, E. D. (2008). *Formative assessment: Examples of practice*. Council of Chief State School Officers. https://www.ccsso.org/sites/default/files/2017-12/Formative_Assessment_Examples_2008.pdf

Wylie, E. C., & Lyon, C. J. (2015). The fidelity of formative assessment implementation: Issues of breadth and quality. *Assessment in Education, 22*(1), 140–160. https://doi.org/10.1080/0969594X.2014.990416

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education, 28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042

Yao, S.-Y., Wilson, M., Henderson, J. B., & Osborne, J. (2015). Investigating the function of content and argumentation items in a science test: A multidimensional approach. *Journal of Applied Measurement, 16*(2), 171–192.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., Hahn, P. R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A. L., & Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature, 573*, 364–369. https://doi.org/10.1038/s41586-019-1466-y

Zeng, W., Huang, F., Yu, L., & Chen, S. (2018). Towards a learning-oriented assessment to improve students' learning—a critical review of literature. *Educational Assessment, Evaluation and Accountability, 30,* 211–250. https://doi.org/10.1007/s11092-018-9281-9

Zheng, G., Fancsali, S. E., Ritter, S., & Berman, S. R. (2019). Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *Journal of Learning Analytics, 6*(2), 153–174. http://dx.doi.org/10.18608/jla.2019.62.11

Zimmerman, B. J. (2011). Motivational sources and outcomes of self-regulated learning and performance. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulated learning and performance* (pp. 49–64). Routledge.