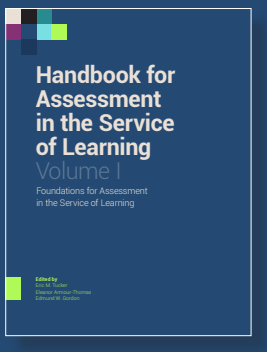
Mind Frames for Improving Educational Assessment

John Hattie, Stephen G. Sireci, and Eva L. Baker

UMassAmherst
University Libraries



Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

ISBN: 978-1-945764-33-2

Suggested Citation:
Hattie, J., Sireci, S. G., & Baker, E. L. (2025). Mind frames for improving educational assessment. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume I: Foundations for assessment in the service of learning.* University of Massachusetts Amherst Libraries.

Mind Frames for Improving Educational Assessment

John Hattie, Stephen G. Sireci, and Eva L. Baker

Abstract

Assessment in education has long prioritized accountability over meaningful interpretation for learning. This chapter calls for a shift toward assessment in the service of learning, emphasizing insights into student progress, learning strategies, emotions, engagement, and self-regulation rather than just achievement. To support this, educators must develop assessment-capable learners who can interpret and act on assessment results. The authors introduce 10 mind frames to enhance assessment, promoting diagnostic and predictive uses, clear success criteria, instructional alignment, and a classroom culture that embraces errors as learning opportunities. They also explore how technology and AI can make assessments more adaptive and personalized. By embedding assessment within teaching and learning, these mind frames transform it from a compliance tool into a driver of student growth and educational improvement.

This chapter calls for changes in how we think about educational assessments. We believe such changes are needed if assessments will truly serve learners. The changes in how we think about assessments, or as we describe *Mind Frames* for improving assessment emphasize the following *Principles of Assessment in the Service of Learning* (Baker et al., this volume):

1. Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.

3. Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.

4. Assessments model the structure of expectations and desired learning over time.

5. Feedback, adaptation, and other relevant instruction should be linked to assessment experiences.

6. Assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences.

7. Assessment quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

Moving Beyond Psychometrics Standards

Too often, assessment is undertaken in the name of accountability of schools, teachers, or students–and there is a place for this purpose. However, the title of this volume highlights *assessment in the service of learning.* A central argument of this chapter is this purpose requires a fundamental shift in the ways of thinking about assessment, changing the focus from "Do we have an assessment with the optimal psychometric properties?" to "Are the interpretations from assessments of value, worth, and significance to improving the teachers' or the student's learning?" Yes, these interpretations need dependable measures, but the stopping and starting points are assessment design and the quality and value of the interpretations.

It is fascinating to watch the focus of the *Standards for Educational and Psychological Testing* change over the seven editions since 1952 (see Sireci, 2020, Table 1). The earlier versions assisted psychometricians in developing and defending tests; even in the latest version (AERA et al., 2014), these Standards include almost nothing to defend the users' rights to optimal interpretations. As Hattie (2014) lamented, "The new edition shows only a grudging move from paying the closest attention to the attributes of items and scores toward, at best, a wink and nudge toward standards for the reporting of test information" (p. 34).

If the AERA et al. Standards had been relabeled as the "*Standards for the Interpretation and Uses of Test Results,*" the field would be in a very different and happier place. If the focus were on interpretations and uses, then teachers and users would more likely attend to the *Standards*, there would be much more discussion about the validity of interpretations and uses, and the field of score reporting would be deep, long, and much further advanced than it is now. Although we are pleased to see a chapter on score reporting finally made it into the soon-to-be-published edition of Educational Measurement (Zenisky, O'Donnell, & Hambleton, in press), we believe 'Interpretation Standards' would assist tests in being seen as worthwhile to users and not as a set of rules for test developers alone. We also believe interpretation standards are supported by the transparent design of assessments to support instruction.

This shift, of course, is not denigrating the importance of creating and validating tests, but if tests STARTED from the nature and depth of valid reporting we wished to make, the liking and development of our profession would be more effective in using assessments to serve learners. Most teachers would know and welcome the *Standards*, most policy makers would not continue to ignore the fundamental premises of the *Standards*, and 'report design' mavericks that flood the internet with glossy reports would have a basis for defending their claims.

Technology has advanced such that so many companies promote glossy reports promising massive improvements in student performance, and it is rare now to expect a user to read a manual or note cautions about the need to triangulate findings. We have seen so many 'reports' placed in folders never to be acted upon; there are so many adverse reactions to 'testing' (e.g., 'we teach students, not data'), and so much data collected that are not used—but schools continue to collect data

because they "have to" and the resulting data and scores provide comfort that "evidence is available, upon call."

We need to do better, and we can. Doing so will require adhering to the standards we set for ourselves, prioritizing using assessment to promote learning over accountability and adopting several other "mind frames for assessment," which are the focus of this chapter.

Shifting the Focus to Improvement

There is value in accountability models of assessment in informing policy makers of the impact of curriculum reforms, investment of resources, and needs for the future. However, classroom assessment's main functions—to inform and improve teaching and learning—are more important. The claim is that assessments can be crucial aids to inform and improve teaching and learning, help educators see "their impact," and accelerate learning when used for diagnostic and predictive purposes.

There needs to be interpretation relating to diagnosis (Where are the students now?), progress (How are they going?), informing next steps (Where should they go next?), and summatively, to ascertain who has attained the success criteria of the lesson(s). Classroom assessment can thus be considered descriptive, ascriptive, and prescriptive. Primarily, the power of assessments comes from feedback to teachers about their impact: What did they teach well, and what not? Who did they teach well, and who they did not? And were the gains made appropriate for the time and resources invested? That is, what, who, and how much. Then, students are the greatest beneficiaries when these three impact questions form a primary assessment function.

Being clear about the diagnosis (including prior performance and current status) and desired success maximizes the power of assessment to know how to accelerate growth best. There is rarely one right way between these two points, and here is where the art of teaching is so critical, especially to evaluate continually the impact of the choice of teaching, the concepts and misconceptions, learning from the errors that are made (they need to be teaching opportunities, not embarrassments, Hattie, 2023). Also, differentiation fundamentally means different ways and times to attain success. Optimal instruction relies on understanding where a student is on a learning progression, the ability to probe for understanding and misunderstandings, and the provision of guidance toward the next steps.

Success involves more than knowing the content, but also includes logical reasoning, understanding cause and effect, organizing prior and new knowledge, problem-solving, and mastering subject matter knowledge and principles. Success involves the 'knowing that" (content and facts), the 'knowing how' (the patterns and deeper conceptual understanding), and 'knowing with' (transfer to near and far situations) (Hattie, 2023; Roland, 1968; Ryle, 1945).

This refocusing of assessment places much emphasis on reliable and dependable interpretations, and these need to be open to critique, moderation, and triangulating (from the teacher noticing, from students' voices about their learning, from others' observing the teacher's impact, and from other measures). It is noted, however, that 'assessment' occurrences can be events (test or assignment) or in-passing, and these need to be planned to inform the teaching status and progress.

This shift in focus aims to develop student assessment capabilities to drive their own learning (Fisher et al., 2023). That is, to teach students how to interpret the results of any assessment such that they can (a) learn to be involved in making decisions about Where to next?, (b) understand more what they know and can do and what they still need to learn to attain success; and (c) fundamentally hear, understand and act upon the interpretations their teachers and they make to improve the rate and depth of learning. The current debates about whether assessments should be graded or marked seem trivial—as the received and understood interpretations matter and not what form they are delivered.

In the remainder of this chapter, we propose 10 *mind frames* to inform thinking about assessment so the power of diagnosis and interpretation can be realized, and assessment can truly be used to accelerate learning.

The Ten Mind Frames for Assessment

Mind frames are ways of thinking related to the cognitive patterns, perspectives, interpretations, evaluations, and mental models we use to interpret our world. They influence how we perceive information, make decisions, solve problems, and navigate with others. What we do matters less than how we think about what we do. In schools, for example, this thinking is the precursor to choosing high-impact strategies, ensuring the fidelity of implementation, and evaluating if there have been important impacts on students from the teacher's delivery of this instruction. It is the teacher's thinking that leads to the choice of interventions, devising and

explaining the learning intentions and success criteria, knowing when a student is successful in attaining those intentions or not, having sufficient understanding of the students' understanding that they bring to the task, and sufficient knowledge about the content to provide meaningful and challenging experiences in various progressive pathways to success in learning.

Mind Frame #1: Assessment in schools needs to consider both progress and achievement, learning emotions, and strategies.

Psychometrics has focused on modeling tests that rank-order students and have little connection to the classroom. We can trace the extraordinary history of item response theory (IRT) back to Thurstone (1925) and Lord's (1952) doctoral dissertations (Luecht & Hambleton, 2021; Thissen & Steinberg, 2020). Most of these models and applications were based on ability testing, and there has been (more in the early days than recently) a robust discussion of whether the assumptions can ever be met when using achievement data (Traub & Wolfe, 1981). There are many problems in using IRT (an essentially norm-referenced model) for achievement testing. Often, there is a major restriction in the range of student proficiency as measured by the test, the assumption of unidimensionality is unlikely to be met (e.g., an achievement test might exhibit unidimensionality when the problems are relatively novel for students, but not as they become more expert; Snow & Lowman, 1984), and there is more often a non-normal distribution of achievement. Moreover, teachers rarely calibrate items, check for even minimal psychometric properties, and most could not even spell IRT and know little about classical test theory. They attempt to compensate for these losses with frequent testing, triangulation with observations, and assessments used to motivate or grade students. In addition, there is a loss of confidence in using many school assessments, especially for reasons of accountability. It is hardly surprising that the major advances in measurement over this past century have rarely crossed the school gate.

When we consider "tests" in schools, the dominant focus is achievement, and all too rarely are there measures of progress, emotions, learning, climate, striving, or engagement. The fundamental claim of this first all-important mind frame is that it is via a high trust, inviting, and safe climate that students are prepared to be challenged to know that which they do not already, engage in error management as opportunities to learn (not embarrassments), be taught to choose optimal learning strategies for the tasks such that there is progress towards higher achievement

relative to their starting position. Instead, the prime focus on achievement often identifies the high and low achievers, and then explanations are sought about why some students can and some cannot.

A fundamental thesis is that educators need to start with high-trust climates, teach learning strategies, and then focus on their students' progress to higher achievement. We should not start with achievement, as this distorts the conversations to privilege those who begin well above average, and enhanced achievement (no matter where the student starts) is supposed to be an outcome of schooling.

The emphasis on high achievement leads to perverse consequences for policy makers, parents, educators, and students. Too many believe a "high achieving school" is necessarily a "great school." If the students start above average, many of these high achieving schools can support 'cruising'—that is, adding no value (e.g., across Australian schools, almost 50% are in this cruising mode; Hattie, 2019). If students start below average and gain more than a year's growth for a year's input, this is stunning and needs esteeming (even if the final achievement is still not above the state or country average). Beliefs in 'high achievement' beliefs damn those who start below average and make substantial progress. We need both high achievement and high progress (See Figure 1). We get high achievement from high progress, and a simple achievement by progress chart could transform the way we make interpretations about what is optimal, how to advance learning and learners, and show that those teachers who impact progress are much more expert and should be esteemed than those who defend cruising. The aim of schooling is to continually move all students from left to right (and by arithmetic, more will increase from lower to higher achievement: every student, no matter where they start, deserves at least a year's progress for a year's input.

Any scores from classroom assessments administered over time can be plotted (and effect size and other measures of progress determined); students thus become their own baselines, and all can see who made progress or not. The interventions to improve will likely differ depending on which quadrant the student is located. Consider two high-achieving students, one in the Cruising and one in the Optimal quadrants. Conventionally, all would be happy that they scored high—but one of these students is cruising, and the other is improving their learning—hence, the teaching strategies need to be different. If only achievement is considered, both seem to be succeeding in the class—but clearly, this is not so for the cruiser.

Similarly, for the two low-achievement students, one is progressing and one is not, and the student progressing should be esteemed similarly to the above-average student who also is progressing. The progression, more than the achievement, is critical for successful schools and learning. Of course, when high progression and high achievement are the outcomes, this is also a highly desirable state.

Measures of learning

When we ask teachers about their 'theory of teaching' this is a prolonged, profound, and plentiful conversation. When we ask about 'their theory of learning' it is too often short, shallow, and subjective. This gap underscores the importance of providing such measures of learning strategies to teachers not to classify them into styles, groups, or hierarchies; but to learn which strategies they use, whether they have fallback strategies when their first choices do not work, and whether their students have the self-regulation skills to choose the optimal learning strategies aligned with the requirements of the task. These are all difficult skills, but essential for successful learning for the specific tasks. There are some learning strategy scales (see Schellings, 2011), but what students self-report may be more what they think they do, not what they do; and may reflect their wishes and beliefs about how they learn. Thus, more sophisticated measures may be needed. Even as adult educators, we struggle with a language about how we learn. Hence, what chance do students have trying to discover these secrets? More powerful measures may include 'thinking aloud' measures, biometric analyses, and it is hoped that the developments in the science of learning can inform the best forms of measuring these skills (Hattie et al., 2024). Other attributes worth developing include measures of cognitive load, retention and forgetting measures, and measures relating to the skills of not only acquiring knowledge and understanding, but skills of consolidating these into chunks or using pattern recognition to better retain for transferring to near or far contexts. All of these efforts need to be grounded in the design of assessments integrally linked to learning goals and processes.

With the push by employers for graduates who can work in groups, translate their knowledge to others, and lead and teach others, schools that do not develop these group skills are not assisting their graduates in being employed (Deming, 2017). There are fascinating measurement problems in measuring an individual's contribution to the group along with group scores and identifying and measuring the 'I' and 'We" skills necessary for successful group functioning (Hattie et al., 2021).

Measures of learning emotions

Students experience many emotions in their experiences in classrooms. Positive emotions, such as happiness, curiosity, and excitement, can enhance learning by increasing motivation, attention, and memory. The most negative influences relate to student anger, procrastination, depression, and anxiety; but the dominant negative emotion is boredom (Blannin et al., 2025; Moeller, Brackett, Ivcevic, & White, 2020). Positive emotions can increase motivation to learn and willingness to take on new challenges, while negative emotions can decrease motivation and lead to lower motivation to engage in learning activities.

Measures of motivation

Many formal and informal measures of motivation involve the observation of students. Early researchers emphasized types of motivation, such as extrinsic and intrinsic rewards. Extrinsic rewards have many limits, including the important idea that the learner must find personal value in the reward. Moreover, there are a range of extrinsic consequences for learning, some short-term and others more delayed. If, for example, good grades in school do not matter to students, they obviously will not inspire improved performance. Intrinsic rewards are thought to engage students by encouraging them to internalize the value of what they are learning. They learn a domain for their own satisfaction. Consider the student who loves history or biology and wants to learn it for its own value. There is an extensive literature on this topic (see Ryan & Deci, 2000) for an excellent review).

More interesting approaches have addressed a specific outcome connected to both archetypal types of motivation. One of these involves the consequence of self-efficacy, whether learning increases the student's capacity to learn or perform (Bandura, 1993). Bandura identified four pillars of self-efficacy development, including 1) mastery experience or performance outcomes (taking on a challenge and being successful;) 2) vicarious experiences (social role models) where the learners see others like themselves succeed: 3) social persuasion, involves positive feedback during learning; and 4) emotional and physiological states; to which we add a fifth, evidence of learning begets more engagement towards more learning. Bandura (1986) noted that being in an environment of wellness may support affective arousal and willingness to learn. It is also clear that a cultural component influences motivation, including the concept of locus of control, the belief that one can change performance by effort compared with the idea that outcomes are predetermined by level of “intelligence” or luck (see Sagone & De Caroli, 2014).

How does one know if a student is "motivated"? Observation has limits, particularly in a teacher or researcher's ability to draw trustworthy inferences from disparate student behaviors. Just because a student is seen to be 'doing' the work is no guarantee of successful learning progress. Short-form self-report measures have been used to determine students' responses to instruction or topics, their willingness to voluntarily seek more of what they have been learning, or their willingness to recommend the topics to a friend. These brief, teacher-made tests, best administered anonymously, can give rapid information about likes and dislikes, interests and self-efficacy regarding lessons. In computer-supported learning, including games, sensors associated with arousal and engagement involve eye-tracking devices, inferences drawn from delays, and other more sophisticated algorithms. Plass and Kalyuga (2019) present a contemporary summary of these efforts. The onset of more artificial intelligence options should allow teachers to make better inferences from students' behaviors, although appropriate applications are on the horizon.

Summary of mind frame #1

This first (and most critical) mindset aims to best inform educators about their impact on students' learning experiences so they can make better diagnostic interpretations leading to more effective instruction and experiences to accelerate students towards the success criteria. The message is to not only look at achievement, but also the causes and correlates of achievement, such as the high trust, high expectations and inviting climate, the choice of learning strategies, the emotions that speed or impede the experience of learning, the consequential levels of engagement, skills in working alone or in groups, and the motivations to want to learn more and deeper about the topics we desire them to learn about.

Mind Frame #2: Develop students' assessment capabilities so they can interpret the feedback and 'where to next' from assessments.

This mind frame emphasizes that we need to ensure students can (a) interpret the assessment results correctly and (b) make some consequential actions, decisions, or thinking that informs their next steps in learning. Sadler (1989, p. 143) noted "it is insufficient for students to rely upon evaluative judgments made by the teacher"; thus, requiring students' critical engagement in discerning the quality of their work and the criteria and standards against which their work is being judged (see also Baird, 2014).

Absolum et al. (2014) recommended all students be educated in ways that develop their capabilities to assess their own learning, and that the success of any national assessment strategy be judged by whether all students are developing the capability and motivation to evaluate, interpret, and use information from quality assessments in ways that affirm or further their own learning. Too often, it is the adults who make the interpretative decisions, and this is as it should be when it informs the impact of and the next decisions by teachers as to what next to teach. However, they claimed that teaching students how to interpret the results of any assessments is necessary to enhance student learning. This permits them to access, interpret, and use information from quality assessments in ways that affirm or further their learning.

Frey et al. (2019) argued that developing assessment capable learners leads to students becoming more aware of their current level of understanding in a learning area, more keen to understand their learning path and have renewed confidence to take on the challenge, better at selecting tools and resources to guide their learning, more ready to seek feedback and recognize that errors are opportunities to learn, more able to monitor their own progress and adjust course as needed, and recognize what they are learning and can teach others (see also Fisher et al., 2015). Thus, making students assessment-capable is perhaps one of the best ways to transform assessments into tools of learning.

To enable students to make these interpretations, teachers must model ways of using assessment information that helps students to meet their learning goals. In this way, students learn about: setting and clarifying challenging learning goals; how to access, interpret, and use evidence; understand the dimensions of engagement that lead to better outcomes; and engage in evaluative thinking about whether the work is good enough, meets the success criteria, and where to make the next learning moves. Wyatt-Smith and Adie (2021) suggested engaging students in discussions and activities to reach a shared understanding of the purpose of the assessment, learning goals, and judgments, applying this understanding to feedback to improve work and develop learning goals, engaging in self- and peer assessment and feedback, co-constructing or deconstructing criteria with teachers, peers or self, build content knowledge and skills to enable decision-making about the quality of one's work, and engage in dialogue with teachers regarding the student's areas of strength/weakness, and learning goals.

These recommendations correspond with the vision outlined by Gordon and Rajagopalan (2016). Specifically, when assessment focuses on improving teaching and learning rather than measuring only what students have learned, we can achieve excellence in education for all students in America. They argued that the challenge for the American education system is not to determine whether or even by how much students have failed to achieve, but to enable them to learn and develop as fully as they are able so they can navigate the world around them, live fuller lives, and contribute as fully as they can to society.

Summary of Mind Frame #2

Students can be empowered and have greater agency in their learning if they understand assessment results and how they lead to next actions. Teachers can help students acquire this understanding by modeling assessment information can help students meet their learning goals

Mind Frame #3: Formative, summative, and ascriptive evaluation refers to the timing and nature of interpretations (and NOT anything to do with kinds of assessment).

Scriven (1967) introduced the concepts of formative and summative evaluation, which are not intrinsically different types of evaluation but have different purposes. Formative evaluation is designed, done, and delivered to make improvements to the evaluand, and summative evaluation is done for, or by any decision-makers who need evaluative conclusions for any reason *other than* conceptual development. The key concept is that in "light of the (formative) processes, or some of them, the product is (or is not) finally revised and released, and summative evaluation begins" (Scriven, 1993, p. 3). So, the distinction is that the purpose is formative during and summative at a key milestone: when the cook tastes the soup it is formative, when the guests taste the soup it is summative (Stake, cited in Miller et al., 2016). Both are important, both need to be appropriately rigorous, neither is more worthwhile than the others, both depend on the quality of interpretations, both require judgments.

One of the major fallacies is that there are concepts such as formative and summative **testing**. Bloom et al. (1971) applied Scriven's terms to education and learning with the release of their book *Handbook on Formative and Summative Evaluation of Student Learning*, where the terms were intertwined with assessment. Give us any test, and we can make formative or summative interpretations. A test

is neither formative nor summative; it depends on when an interpretation is made for improvement or the end of an intervention. It may be that tests that lead to more optimal formative interpretations more closely track instruction and a level of detail to allow sub-tasks to be addressed. In contrast, summative evaluations are more about whether the intentions of the lesson are known and understood. Still, both require defensive psychometric properties, interpretations (to improve or to ascertain status), and the difference is more in the timing and purpose.

This yoking or formative and summative to 'tests' has led to a focus of many professional development programs. For example, Black and Wiliam (2010) noted "formative assessment is an essential component of classroom work" and that "they know of no other way of raising standards for which such a strong prima facie case can be made" (p. 14). Perhaps it is unsurprising that Wiliam later argued, "the biggest mistake that Paul and I made was calling this stuff 'assessment'… because when you use the word assessment, people think about tests and exams" (Booth, 2017, p. 2). He argued that the program may have been more accepted and successful if they had used "Responsive teaching" and not tied it to 'tests.'

It is time to revert to Scriven's original claim: formative and summative evaluation–refers to evaluative thinking and evaluative decision-making when reviewing the outcomes of teaching, learning, and assessment opportunities (Clinton & Hattie, 2024). The quality and defense of the interpretations become critical (and not the test, sui generis). The tie-in with assessment has done a major disservice to the original Scriven notions. It has led to too much emphasis on the assessments and too little on the timing and quality of interpretative information.

Summary of Mind Frame #3

We need to abandon notions of formative and summative tests because those terms are misnomers. Educational assessments can give us formative and summative information, and that it is the timing and nature of the interpretations that make them valuable.

Mind Frame #4: There are at least three levels of knowing: Knowing that (concepts, ideas, facts, surface), knowing how (relations, deeper conceptual), and knowing with (transfer, pattern recognition)—and assessment may need to measure each separately.

Contrary to many test developers' claims, a majority of the items on standardized assessments in the USA can be answered by simply knowing lots, especially in knowledge-rich subjects like science and history (Koretz, 2017). In fact-dominant classrooms, teachers ask questions primarily about the facts, and students soon realize that 'knowing lots' is the sign of a good learner in many classes. However, most theories of schooling ask for more than knowledge-rich graduates. There is also a clamor for students who can see patterns across ideas, have deeper conceptual knowing, and can transfer ideas from one to another situation or problem. The need is not either facts or deep, but facts *and* deep—depending on the nature of the problem. Much deeper and readily available analyses are needed as to how users actually answer items, the knowledge and pattern recognition they use, and the error management they use to solve a problem and decide on their optimal answer.

One of the greatest travesties has been an over-reliance on Bloom's (1956) taxonomy. It mixes knowing (knowledge, comprehension), ways of knowing (analysis and synthesis), outcomes (applications), and evaluating and creating. There is no hierarchy; everything fits somewhere (which means everything is ok), and there is limited to almost no research on the value of the taxonomy (Hattie & Purdie, 1998). In 2001, Anderson and Krathwohls (2001) revised edition was published, acknowledging many of these deficiencies and adding another more powerful dimension: learning objectives for each of the six' levels': factual (what needs to be known, conceptual (interrelationships among the basic elements); procedural (how to do something, methods of inquiry, and criteria for using skills, algorithms, techniques, and methods), and metacognitive (awareness and knowledge of one's cognition). These latter three can be termed degrees of cognitive complexity, and they have been developed using many learning and knowledge models.

Webb (1997, 2002, 2007) developed the 'Depth of Knowledge' model, and it has four levels quite similar to the revised Bloom new dimensions: recall, skill or concept, strategic thinking, and extended thinking (also see Hess, 2006; Francis, 2022). Recall relates to reproducing a fact, principle, or routine procedure. Skills or concepts relate to using information, selecting appropriate procedures for a task, or organizing and displaying interpreted information. Strategic thinking involves reasoning or developing plans to approach a problem, employing decision-making and justification, and solving abstract, complex, or non-routine problems. Extended thinking involves performing investigations or applying concepts and skills to the real world that require time to research, problem-solve, and process multiple conditions of the problem or task. Webb's Depth of Knowledge model relates to the depth of content understanding, the scope of a learning activity, and the skills required to complete tasks.

A powerful taxonomy is SOLO (Structure of Observed Learning Outcome) taxonomy. Developed by Biggs and Collis (1982), it has been critical in developing assessments, scoring, rubrics, and teaching and learning. The SOLO taxonomy consists of five levels of increasing complexity: pre-structural, unistructural, multistructural, relational, and extended abstract (simply referred to as no idea, one idea, many ideas, relating ideas, and extending ideas). Each level represents a different level of understanding and the ability to think about and use information in increasingly sophisticated ways. Thus, it is possible to conceive of 'difficulty' as an increasing challenge within each of the four levels, and 'complexity' as an increasing challenge when moving from instructional to extended abstract.

The SOLO taxonomy provides a valuable framework for understanding and assessing learning outcomes as it is based on a model of cognitive complexity that allows educators to identify the level of understanding that students have reached and to design appropriate learning activities and assessments that challenge students to progress to higher levels of understanding (know where a student is performing and aim one step higher). Students can also use it to reflect on their learning and identify areas where they need to improve their understanding.

Clinton et al. (2021) reviewed Bloom, Depth of Knowledge, and SOLO and developed a model that brought all three together. Their four levels are categorized into two attributes: 'knowing that'/surface and 'knowing how'/deep thinking. The first major attribute of cognitive complexity relates to *knowing that* or surface thinking—the ideas, the factual, and content knowledge. This includes:

1. Factual knowledge recall and reproduction: for example, 'I know and can distinguish various parts of human hand anatomy.'
2. Conceptual knowledge involving basic application and skill: for example, 'I am able to apply Van Gogh's painting techniques to my drawing of the Sydney Opera House.'

The second major attribute is *knowing how* to think deeply or develop relations between ideas, extending to near or far new situations.

3. Strategic or relational thinking: For example, 'I understand that empowerment evaluation design principles have been utilized to design this evaluation compared to other evaluation models.'
4. Transfer: for example, 'I will be able to apply and adapt my methodological learning from my previous understanding of jazz principles to this new piece of music.' This can also be termed *knowing with.*

A summary of the three cognitive taxonomies using Clinton's (2021) model is presented in Table 1.

This powerful way to distinguish the levels of cognitive complexity was suggested by Ryle (1945), who distinguished between 'knowing that' and 'knowing how.' 'Knowing that' is the knowledge that something is the case (e.g., knowing an evaluation theory), and 'knowing how' is the knowledge you have when you know how to do something, such as how to ride a bike, bake a cake, or make an evaluation interpretation—"how to make and appreciate jokes, to talk grammatically, to play chess, to fish, or to argue" (Ryle, 1949, p. 28). 'Knowing how' cannot be defined in terms of 'knowing that,' nor is 'knowing how' necessarily logically before 'knowing that' (Kapur, 2008, 2012, 2016; Oberauer, 2010). For example, you cannot teach a novice chess player how to play to the same standard as the grandmaster just by feeding the novice facts about the game. Thus, 'knowing that' entails 'knowing how' to put the 'knowing that' into practice, but 'knowing how' cannot be built up from pieces of 'knowledge that.'

Clinton et al. (2021) argued that underlying these four levels is the notion of "evaluative thinking" or self-regulation–knowing when to be surface and when to be deep, and this skill invokes a particular kind of critical thinking and problem-solving. Evaluative thinking is the process by which one marshals evaluative data and evidence to construct arguments that allow one to arrive at contextualized value judgments in a transparent fashion (see also Buckley et al., 2015; Lee, Wallace, & Alkin, 2007; Vo et al., 2018).

Summary of mind frame #4

Assessments need to be developed, scored, and reported at these four levels, or at least at the "Knowing that" compared to the "Knowing how and with" levels. This is not to be confused with difficulty at each level (factual, strategic, etc.). There can be increasing difficulty levels, but cognitive complexity moves down the depths from surface to deep to transfer. These levels were the basis for developing the NZ elementary and high school e-asTTle (https://e-asttle.tki.org.nz/; Hattie et al., 2006), where assessments were automatically created using linear programming methods (van der Linden, 2005) so teachers and students could understand how they performed on both easy to more difficult items and on surface to deeper cognitive complexity.

Mind Frame # 5: We care about alignment between standards, success criteria, lessons, tasks, and assessments, and recognize the power of backward design (e.g., reports to tests, summative goals to influence formative directions, etc.).

A fundamental assumption of unidimensionality is that the attribute being measured can be ordered from high to low, more to less, effective to ineffective, etc. But the world of classrooms is multidisciplinary with various teaching methods, learning strategies, safety to learn and err, curriculum progressions, tasks, and assignments –oftentimes, the outcome is also multidisciplinary. This often leads to simplistic dichotomies or decisions that simplify but distort the complex nature of reality. It is not necessarily either-or, but when, how, and with what impact. Surma et al. (2024) make a strong case for both rather than either-or claims about typical dichotomies such as knowledge-rich vs. deeper learning, direct instruction vs inquiry teaching, multiple choice- vs. open-ended, grading vs. comments.

A fundamental mission of education is to influence academic outcomes (alongside many other attributes noted in Mind Frame 1). When trying to understand the underlying reasons for the very discrepant effect sizes of many teaching methods, Hattie (2023) developed the 'Intentional Alignment model'. This model relates to the proportion of knowing that, how, and with (surface, deep, transfer) in the lessons, and then aligning these the notions of success (e.g., one success criteria for knowing that, and one for knowing how and with), the methods of teaching, the optimal strategies of learning, the cognitive requirements in the activities, and the assessment methods.

There are seven major parts of the Intentional alignment model:

1. Determining the learning intentions and success criteria, typically based on a curriculum, understanding progressions (where the students have been progressing, where they are now, and where they need to be), and the motivations and dispositions students bring to the class.
2. Cognitive task analysis of the knowing that, knowing how, and knowing with foci of the lessons (i.e., content, deeper understanding, and transfer skills and knowing). Understanding the cognitive complexity involved in learning ensures students have the strategies for learning, understand their progress, and have a concept of what success looks like.
3. Creating a climate and culture of the class to ensure students and teachers see errors or misconceptions as opportunities to learn. There needs to be safety in working, discovering, and exploring with peers, an inviting climate for all, and acknowledging the diversity of what each and all students bring to the class environment.
4. Teaching methods that align with the cognitive complexity of the various components of the tasks, the required confidence to take on the challenges of this complexity, and efficiently and effectively improve students' progress towards the success criteria.
5. Ensuring students have appropriate and effective learning strategies to engage in the complexity of the tasks.

6. Choosing activities that align with the content's complexity levels, the deeper conceptual and relational thinking.
7. Using assessment methods that focus on the content ('knowing that'), the relational ('knowing how') and transfer ('knowing with') that inform teachers and students of their progress, success, and gaps to be then addressed.

These aspects of alignment are not only critical for using assessment to promote learning, but they are also essential for providing (a) evidence of the validity of these classroom assessments, and (b) professional development for teachers as they improve their teaching and formative evaluation processes. As the AERA et al. (2014) *Standards* stated, "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). The curriculum and goals of instruction provide the theory to support test score interpretation and use, and confirmation of alignment provides the evidence to support such interpretation and use. Moreover, teachers develop and strengthen valuable skills for helping students learn by developing and sequencing assessments aligned to instructional goals.

It would be worthwhile to at least have separate tasks, assignments, and assessments for the 'knowing that' and for the 'knowing how/with.' This would make transparent to students what is being valued, make the feedback from any assessments more easily understood, and encourage the use of differentiated teaching, learning, and activities depending on the focus on knowing that or how/with. Hence, the validity of interpretations from assessments is more robust.

Summary of Mind Frame #5

Alignment is needed across all stages of the education process, from goal and standard formulation to assessment, interpretation, and feedback. The Intentional alignment model provides a comprehensive way for building and evaluating such alignment.

Mind Frame #6: Provide scoring rubrics and success criteria near the beginning and not the end of lessons/courses to provide verisimilitude to any assessment.

Imagine asking students to play their video games and NOT telling them what it means to be successful or what it means to get to the next level. They would say it is pointless, so why would I engage in activities without a purpose or goal? But this is often the case in schools when students are told to "do the work" with little to no understanding of what is 'good enough,' what the criteria of success look like, or now know the evaluative decisions to be made to claim success. Blannin et al. (2025) analyzed the results from an App that asks students about their experience in class, and the dominant emotion is boredom. However, when students can explain what they are learning, there is a 73% (from 37% to 64%) improvement in enjoyment and engagement in their learning.

There must be appropriately challenging expectations embedded in the success criteria. That is, challenging relative to what each student cannot do (yet), and then the teaching focuses on enabling students to engage in the challenge. The Goldilocks principle of challenge is optimal: not too hard, not too easy, not too boring. Unfortunately, some students are reluctant to take on challenges (for fear of failure, becoming an embarrassment in front of peers and the teacher, or not having the skills to address the challenge). Hence, it is important to specifically know, plan for, and develop confidence to take on appropriate challenges (self-efficacy), and have high levels of trust that failure is the best friend of learning in this class.

As noted earlier, most lessons may need two success criteria: one for the 'knowing that' and one for knowing how and with' goals): similarly with assessments. It is a balance of proportions, not an either/or. The effectiveness of assessment can increase when students are provided with scoring rubrics. Scoring rubrics are more effective when they are provided at the outset rather than at the end of activities, and even better when students are involved in creating them (Becker, 2016), when they relate to lessons over time (and not short term(Andrade & Valtcheva, 2009).

Summary of mind frame #6

Teachers know what is coming next in an instructional sequence and how they will evaluate student work. Students should know these things also; so they and their teachers have the same target in sight. By providing students with the rubrics with which their work will be evaluated, they will approach the tasks more confidently, and thus enhancing their learning.

Mind Frame #7: We create a climate of welcoming errors and not knowing; thus, a major role of assessment is diagnosis and discovering where students are currently knowing and not understanding.

Mastery is often claimed to be 80%+; when a student gets 100%, perhaps the test was too easy; when they get <40% they are usually despondent. Testing would have little use if it did not detect what the student does not know and understand, what the focus of the next phase of teaching and learning should be, and if the climate is such that there is negativity from learning from errors and not knowing (Law et al., 2004). Indeed, we do not come to class to learn what we know and understand, so the starting point is 'not perfection.' Sadly, so many learn quickly that classrooms privilege those who know and can do, eagerly respond to teacher questions, and know how to do the work. A major role of assessments needs to be diagnostic and discovery of where students are currently, and then make predictive recommendations for further learning. Too much assessment is making a status report on where 'they are now,' which is valuable, but the starting point to 'where to next' is the essence of learning.

Feedback thrives on errors, and undertaking challenges leads to errors. Therefore, the climate of the class must welcome errors, teach students how to engage in error detection, have teachers positively engage in error repair, and failure needs to be a learner's best friend. Errors can serve as important feedback information, indicating where the student's thinking and knowledge are not yet developed. However, students and teachers often shun errors to avoid negatively impacting a student's self-esteem; peers can be nasty, brutish, and short to those showing they do not know. When students make errors in classroom discussions, they are usually quickly corrected (by the teacher or by the teacher asking a peer), and many students soon learn that it is better to look like they know and hope they do not get asked. Thus, avoiding and withdrawing are successful tactics to maintain their sense of themselves as learners.

Many learning models explicitly include attention to errors and failure. For example, Piaget's (1952) discussed cognitive disequilibrium, which occurs when learners encounter a situation contrary to their current mental model (such misconceptions have been studied mainly in science learning). Then students are challenged or instructed until they either assimilate those differences into their mental model or modify it according to the new information. Productive failure invites students to solve, usually ill-structured or complex, problems before instruction to evaluate

their disequilibrium, try and invent multiple solutions, and realize what they need to learn from the subsequent instruction (Kapur, 2024). Thus, there needs to be a climate for receiving, welcoming, and learning from test information.

Gordon (2020) has long promoted assessment in the service of learning. This involves interrelating assessment, teaching, and learning such that they are reciprocally employed each in the service of the other. Measurement is no longer primarily for testing but to inform "teaching and learning transactions, their outcomes and their continuing assessment" and thus can a) measure the status of developed ability; analyze processes of teaching and learning; understand intentions, appreciations, needs meanings, performances; and cultivate learning and development of abilities, appreciations, competences and skills (p. 73). Assessment can not only measure but also cultivate learning.

Summary of mind frame #7

Classrooms should embrace a culture where mistakes are welcomed and discussed. In such environments, the diagnostic role of assessments can help students discover where they are and what to do next improve their learning

Mind Frame #8: We act on the belief that a major purpose of assessment in schools is to inform teachers of their impact.

This is the major message of the Visible Learning work—we care less about how teachers teach and assess but care more about the impact of their teaching and assessment interpretations. Switching from the act of teaching and testing to the *impact* changes the debate, makes it easier to recognize expertise, and puts the focus on the dependability of the interpretations and consequential actions. Excellence, for example, is not tied to the use of any specific set of teaching strategies or testing methods but to the optimal alignment and fidelity of implementation of strategies and methods that have an impact on student learning. This focus on impact means interpretation of multiple sources of evidence, including test scores, is critical. Other sources include teacher-set assignments and assessments, their noticing, student voice about their learning, the evidence of the safety to not know and make errors, and classroom management to include *all* in the learning. The key is triangulating this evidence from tests, teacher noticing, student voices about their learning, and artifacts of student work. Every school has pockets of high-impact teachers with high levels of evaluative thinking. The

core question is how to understand this evaluative thinking such that it can scale up. In education, we are good at finding and fixing problems, but not so good at identifying excellence and scaling it up (Baker, 2004).

Summary of mind frame #8

Educators can use assessment to assess their impact and the impact of the different teaching tools and instructional practices they use.

Mind Frame #9: Many technologies (especially large language models) can make developing and scoring assessments more efficient and lead to making assessment interpretations more defensible.

The most remarkable change in our lives is the advent of large language models such as ChatGPT, Gemini, Claude, etc. These tools offer the most remarkable transformation of assessment since the development of IRT. While the usual claims about evaluating the dependability and validity remain (and become perhaps even more important), these AI models can more readily 'write' items to specifications, offer more opportunities to adapt tests on the fly, score open-ended assessments, and create more tailored interpretations from the assessments. It is early days, but the possibilities abound. Maybe the over-dependence on multiple-choice and closed items that have dominated many assessment systems (as they are cheaper and easier to score) will be reduced, and items that map the processes of learning, the deeper ideas, and the construction (rather than recognition) of answers will take their rightful place within tests.

As students gain more insight into how they can create, score, and interpret assessments via AI, it will be important to understand the skills we need to teach them to do this in a worthwhile way to enhance their learning. Such skills could include how to ask probative questions (if the wrong prompts are asked AI still gives answers), assessment credibility (are the AI comments and recommendations right or wrong), evaluative thinking (are they 'good enough'), making wise choices (these tools can make recommendations for next learning steps and they may or may not be optimal), oral fluency (you can speak to the AI engines), and collaborative critique (how to engage others in critiquing the use and outputs from these tools).

There have been significant changes in assessment over the last decades due to technology. These changes include automated scoring of constructed responses, digital assessments that provide log (process) data that can be used to understand

better test takers' cognitive processes, advances in computer adaptive testing, personalized or designed-in-real-time methods of administration, and improved score reporting that maximizes the interpretations from the assessments.

Sireci et al. (2024) proposed a new model using many of the advances from AI tools called Design-In-Real-Time (DIRTy) assessment, which reflects the progressive evolution in testing from a single test to an adaptive test, to an adaptive assessment system. It involves: (a) assessment building blocks (individual items or "assessment task modules" (ATMs) that are linked to multiple content standards and skill domains), (b) gathering information on test takers' characteristics and preferences and using this information to improve their testing experience, and (c) selecting, modifying, and compiling items or ATMs to create a personalized test that best meets the needs of the testing purpose and the individual test taker.

What is new in DIRTy assessment is tailoring the test to multiple, personal factors, and delaying test specifications until the interaction of a test taker and a testing purpose occurs. DIRTy assessment, can use individual items or sets of items aligned with both content standards and job tasks that represent the building blocks of a unique instantiation of an assessment, and a system to search and assemble those sets of items (ATMs) to meet the additional configuration goals of personalized assessment (a topic further elaborated by Buzick et al., 2023). A major advantage is that by using the assessments to enable test takers to solve problems, the assessment becomes a vehicle on route to solving the problem, and hence has the potential to promote learning (Gordon, 2020). DIRTy assessments can also be more culturally responsive. For example, students can have choice in which reading passages or other stimuli they respond to, and the choices can reflect the communities in which test takers live. Furthermore, the test delivery system can allow test takers to access translations of test material while taking the assessment, or even to assemble a different language version of the assessment.

Maybe it is time to learn from the fast-growing world of technology. We need to swallow hard and start with evidence-based interventions at the outset to provide models for teachers and then fix them in the wild (the much-reviled idea of flying a plane while repairing or improving them). This would lead, like in technologies, to numerous software updates (as we receive with our cars and phones) using the evidence from users to improve the impact of the software, track how prior users

have used and progressed when using the software, and de-implementing those features which hinder or have little evidence of impact.

Summary of mind frame #9

Technology has much to offer education, and with respect to assessments that serve learners, we can use technology to develop interactive assessments that are optimal for *each* learner, rather than a single assessment optimized for *most* learners. In addition to developing and delivering tests, technology can provide real-time feedback from assessments, and engage learners in understanding their assessment performance and what to do next. Technology has the potential to allow assessments to foster engagement in the testing process, and to be fully aligned with and integrated into instruction.

Mind Frame #10: Transform the purposes of assessment to move away from an overreliance on accountability to more continuous assessments used in learners' acquisition of understanding, motivation for learning, collaboration, and problem-posing settings.

A key theme of this chapter is assessments can be used more effectively to support student learning when we focus on the ways of thinking or the mind frames of the teachers and students. The advent of AI opens new opportunities, and there is not much confidence among teachers and many students about the benefits of the current assessment and accountability regimes.

Sireci (2021) identified four reasons why there has been a loss of confidence in educational assessments: (a) measurement professionals are often hypocritical (i.e., we impose standards, but don't follow them); (b) we present a censured history of educational and psychological testing to ourselves and the public, but the public knows better; (c) we focus on what was important 100 years ago, rather than what is important today; and (d) we are entrenched in a culture of distrust (see also Baker & O'Neil, 2020). Such "psychometric paralysis" requires a new way of thinking about assessments and asks for a de-emphasis of norm-referenced competitiveness in educational testing, except in those rare instances where examinees actually are competing for a benefit. To regain trust, we need to: (a) engage with teachers and other educators to collaboratively develop tests and interpret test scores; (e) reconceptualize our notions of standardization to make tests more flexible to students' needs and funds of knowledge; and (f) design test score reports for

students that emphasize their strengths, rather than their weaknesses. Essentially, we need to reorient our practices to value students more than a score scale.

Most classroom assessments rarely move past estimates of achievement, and few consider the students' strategies of learning, their emotions before, during, and after a lesson(s), and so often, the skills to work alone or in groups are considered. There needs to be a transformation of classroom assessment purpose from annual, time-controlled accountability assessments to more continuous assessments used in learners' acquisition of understanding, motivation for learning, collaboration, and deep application of knowledge in problem-solving, communication, and authentic settings.

Traditional assessment has focused on accountability, promotion, certification, and competition (e.g., admissions). Focusing the assessment purpose more directly on student learning aims to improve student learning greatly. Indeed, suppose we can frame the assessment problem as providing the most appropriate assessment for a particular learner at a particular point in time to provide specific information to support their learning. In that case, we will have solved a much more important problem than accountability. Such student-targeted assessments will have more validity to understand what students know and where they need to go next. We do not need to stop instruction to assess students for accountability purposes; we need assessments as part of instruction to guide it. If we solve the problems of developing, administering, and interpreting valid assessments supporting student learning, aggregating the results from those assessments for accountability will be relatively easy. It is time for assessments to be supportive, not disruptive. It is time to make learners full partners in the assessment process so the results benefit them and place their needs above those of the policy makers.

Gordon and Rajagopalan (2016) noted that too much testing overemphasizes the status of a narrow range of cognitive functions in learners and neglects the affective and situative domains of human performance and the processes by which these functions and domains are engaged. They neglect the diverse contexts and perspectives born of different cultural experiences and cultural identities and the influence of these contexts, perspectives, and identities on human performance. They have privileged accountability, relative positioning and competition to the neglect of criterion-based judgments of competence. They have overly focused on knowing, knowing how to, and mastery of knowledge that is held to be objectively

true. But they have confidence that assessment "can be a powerful and dynamic tool for effecting real transformation in how we view and deliver education in our society today and in the future."

Concluding Comments

In this chapter, we presented 10 mind frames for teachers, students, test developers, and all others involved in educational assessment that will move educational assessments from the measurement "of learning" to "supporting learning." Assessment best supports student learning when it is fully integrated with it. Baker (2018) has argued that it is well nigh that we "connect assessments systematically to instruction. Most approaches to establish instructional sensitivity of assessments to elements in learning programs are primitive at best, but needed for 'personalized' learning approaches" (p. 140).

Assessment should be organized as part of the instructional process rather than be disruptive to teaching and learning. Coordinating curriculum, instruction, and assessment is often described as alignment, which Webb (1997) described as "the degree to which expectations [i.e., standards] and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (p. 4). Including assessment in instruction requires considering when student evidence of learning is needed, what form that evidence should take, and how the evidence will inform future learning and instruction. Such considerations require aligning the content of assessments with the instruction objectives and timing the assessment at key points to inform instruction (Martone & Sireci, 2009). Without a doubt, the focus on improvement through design and teacher and student interpretations is essential, as are the moral and desired modifications to practice that support varied learners where they are. Policy changes that address tools and models for teachers, useful repositories of assessment, and evidence-based interventions should be integrated. We have been extolling the virtues of systematic codesign of teaching, learning, curriculum, and assessment since before Ralph Tyler's seminal efforts (1958). Perhaps creating usable tools that support infrastructure needed in curriculum design, teaching, and learning assessment may be an approach worth exploring once again.

Challenges remain. One is discovering how evaluative thinkers interpret and make decisions based on test scores and how we can best scale up these ways of thinking. Such scaling up remains one of the greatest unsolved issues

in educational research and practice. We see the opposite in developing contemporary technology, emphasizing scaling in a "first-to-market" mentality. While initial tests and evaluations might be undertaken, many technologies are released and then iteratively improved by collecting evidence of how users interact with the app, continually releasing updates.

In the domain of teaching, scaling of innovation or research-based options is hampered in at least three ways. First, many interventions need to be evaluated over a long period, not a day, a week, or in a single grade level, or on only one topic. Such trials are not practicable in a fast-moving market. Second, we continually argue that "our context is unique" and introduce adaptations that can take the innovation out of an implementation so that it becomes similar to what we had been doing anyway (hence, the innovation is too rarely actually implemented). Third, we so often scale without evidence of value, but based on teacher word-of-mouth or top-down mandates.

Scaling-up issues aside, we continue to think that the mind frame shift is likely the most important catalyst for accelerating student learning at a broad level for all students. As the mind frames in this chapter illustrate, by engaging and empowering students in their learning, by making them feel comfortable with their mistakes and allowing them the freedom to explore them, and by thinking about how assessments can be developed and used differently, we will have progressed beyond 19th and 20th-century ways of thinking about tests, to using them to support student learning.

References

Absolum, M., Flockton, L., Hattie, J. A. C., Hipkins, R., Reid, I (2009). *Directions for Assessment in New Zealand: Developing students' assessment capabilities.* Ministry of Education, Wellington, NZ. http://assessment.tki.org.nz/Assessment-in-the-classroom/Directions-for-assessment-in-New-Zealand-DANZ-report

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

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman.

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into practice*, *48*(1), 12–19.

Baird, J. A. (2014). Teachers' views on assessment practices. *Assessment in Education: Principles, Policy & Practice*, *21*(4), 361–364.

Baker, E. L. (2004). *Principles for Scaling Up: Choosing, Measuring Effects, and Promoting the Widespread Use of Educational Innovation.* CSE Report 634. Center for Research on Evaluation Standards and Student Testing CRESST.

Baker, E. (2018). Design for assessment change. *European Journal of Education, 53*(2), 138–140.

Baker, E. L., & O'Neil, H. F. (2020). The assessment landscape in the United States: From then to the future. *Monitoring student achievement in the 21st-century: European policy perspectives and assessment strategies*, 51–61.

Bandura, A. (1986). *Social foundations of thought and action.* Englewood Cliffs.

Bandura, A. (1993). *Perceived self-efficacy in cognitive development and functioning. Educational Psychologist, 28*(2), 117–148.

Becker, A. (2016). Student-generated scoring rubrics: Examining their formative value for improving ESL students' writing performance. *Assessing Writing*, *29*, 15–24.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome).* Academic Press.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90.

Blannin, J., Hattie, J. A. C., Wood, C., & Stubbs, P. (in review). Informing Professional Learning Interventions with Evidence-Based Analysis of Student Feedback: Implications for Software Use and Learning Clarity

Bloom, B. S. (1971). *Handbook on formative and summative evaluation of student learning.* McGraw-Hill.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain.* David McKay Co.

Booth, N. (2017, July 9). *What is formative assessment, why hasn't it worked in schools, and how can we make it better in the classroom?* Impact. https://my.chartered.college/impact_article/what-is-formative-assessment-why-hasnt-it-worked-in-schools-and-how-can-we-make-it-better-in-the-classroom/

Buckley, J., Archibald, T., Hargraves, M., & Trochim, W. M. (2015). Defining and teaching evaluative thinking: Insights from research on critical thinking. *American Journal of Evaluation*, *36*(3), 375–388.

Buzick, H. M., Casabianca, J., & Gholson, M. L. (2023). Personalizing large-scale assessment in practice. *Educational Measurement: Issues and Practice*, https://doi.org/10.1111/emip.12551

Clinton, J. M., & Hattie, J. A. C. (2021). *Cognitive complexity of evaluator competencies.* Evaluation and Program Planning, 89, 1–8.

Clinton, J., & Hattie, J. (2024). Revisiting and Expanding Scriven's Fallacies About Formative and Summative Evaluation. *Journal of MultiDisciplinary Evaluation*, *20*(47), 13–23.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics, 132*(4), 1593–1640.

Fisher, D., Frey, N., Ortega, S., & Hattie, J. A. C. (2023). *Teaching students to drive their learning: A playbook on engagement and self-regulation.* Corwin.

Francis, E. M. (2022). *Deconstructing depth of knowledge*. Solution Tree.

Frey, N., Hattie, J. A. C., &, Fisher, D. (2018). *Developing Assessment Capable Learners*. Thousand Oaks, Corwin.

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, *39*(3), 72–78.

Gordon, E. W., & Rajagopalan, K. (2015). *The testing and learning revolution.* Jossey-Bass.

Hattie, J. A. C. (2014). The Last of the 20th-Century Test Standards. *Educational Measurement: Issues & Practice*, *33*(4), 34–35.

Hattie, J. A. C. (2018). *Implementing, scaling up, and valuing expertise to develop worthwhile outcomes in schools*. William Walker Oration, Presented at the Annual Conference of the Australian Council for Educational Leaders, Sydney. ACEL Monograph #58. http://www.acel.org.au/acel/ACELWEB/Publications/ACEL_Monograph.aspx

Hattie, J. A. C. (2023). *Visible learning: The sequel: A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

Hattie, J. A. C., Brown, G., Ward, L., Irving, E., & Keegan, P. (2006). Formative evaluation of an educational technology innovation: Developer's insights into assessment tools for teaching and learning. *Journal of Multidisciplinary Evaluation, 5*(3), 1–54.

Hattie, J. A. C., Clarke, S., Fisher, D., & Frey, N. (2021). *Collective student efficacy.* Corwin.

Hattie, J. A. C., O'Leary, T., Hattie, K., & Donoghue, G. (2025). *Great Learners by Design.* Corwin.

Hattie, J. A. C., & Purdie, N. (1998). The Solo model: Addressing fundamental measurement issues. In Dart, B., & Boulton-Lewis, G. M. (Eds.), *Teaching and learning in higher education*. Camberwell, Vic, Australian Council of Educational Research.

Hess, K. (2006). *Exploring cognitive demand in instruction and assessment.* National Center for the Improvement of Educational Assessment, Dover NH.

Kapur, M. (2008). Productive failure. *Cognition and instruction*, *26*(3), 379–424.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*, 651–672.

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299.

Kapur, M. (2024). *Productive Failure: Unlocking Deeper Learning Through the Science of Failing.* Jossey-Bass

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

Law, N., Alexander-Hollins, Smith, D., & Hattie, J. A. C. (2024). *10 Mindframes for the culture and climate of schools: Equity, Identities, and Belonging.* Corwin.

Lee, J., LeBaron Wallace, T., & Alkin, M. (2007). Using problem-based learning to train evaluators. *American Journal of Evaluation*, *28*(4), 536–545.

Linden, W. J. (2005). *Linear models for optimal test design.* Springer Science+ Business Media, Incorporated.

Lord, A. Theory of test scores.-, 1952. *Psychometric Monograph*, (7).

Luecht, R. M., & Hambleton, R. K. (2021). Item response theory: A historical perspective and brief introduction to applications. In *The History of Educational Measurement* (pp. 232–262). Routledge.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction, *Review of Educational Research 4*, 1332–1361.

Miller, R. L., King, J. A., Mark, M. M., & Caracelli, V. (2016). The oral history of evaluation: The professional development of Robert Stake. *American Journal of Evaluation*, *37*(2), 287–294.

Moeller, J., Brackett, M. A., Ivcevic, Z., & White, A. E. (2020). High school students' feelings: Discoveries from a large national survey and an experience sampling study. *Learning and Instruction*, *66*, 101301.

Oberauer, K. (2010). Declarative and procedural working memory: Common principles, common capacity limits?. *Psychologica Belgica*, *50*(3–4), 277–308.

Piaget, J. (1952). *The origins of intelligence*. New York: International University Press.

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review*, *31*, 339–359.

Roland, J. (1958). On "knowing how" and "knowing that". *The Philosophical Review, 67*(3), 379–388.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67.

Ryle, G. (1945, January). Knowing how and knowing that: The presidential address. In *Proceedings of the Aristotelian society* (Vol. 46, pp. 1–16). Aristotelian Society, Wiley.

Ryle, G. (1949). *The concept of mind*. Penguin Books.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

Sagone, E., & De Caroli, M. E. (2014). Relationships between psychological well-being and resilience in middle and late adolescents. *Procedia-social and behavioral sciences*, *141*, 881–887.

Schellings, G. (2011). Applying learning strategy questionnaires: Problems and possibilities. *Metacognition and Learning*, *6*, 91–109.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler (ed.), *Perspective of Curriculum Evaluation*, American Educational Research Association (AERA), Monograph of Curriculum Evaluation, No. 1., Rand McNally.

Scriven, M. (1993). The nature of evaluation. *New Directions for Program Evaluation*, *58*, 5.

Sireci, S. G. (2020). "De-"Constructing" Test Validation," *Chinese/English Journal of Educational Measurement and Evaluation*, 1. 教育测量与评估双语季刊: https://www.ce-jeme.org/journal/vol1/iss1/3 https://doi.org/10.59863/CKHH8837

Sireci, S. G. (2020). Standardization and understandardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105. https://doi.org/10.1111/emip.12377

Sireci, S. G. (2021). Valuing educational measurement. *Educational Measurement: Issues and Practice, 40*(1), 7-16. https://doi.org/10.1111/emip.1241

Sireci, S. G., Suárez-Alvárez, J., Zenisky, A. L., & Oliveri, M. E. (2024). Evolving educational testing to meet students' needs: Design-in-real-time assessment. *Educational Measurement: Issues and Practice*, *43*(4), 112–118.

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 26,* 347–376.

Surma, T., Vanhees, C., Wils, M., Nijlunsing, J., Crato, N., Hattie, J., Muijs, D., Rata, E., Wiliam, D., & Kirschner, P. A. (2025). *Developing Curriculum for Deep Thinking: The Knowledge Revival.* Springer

Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation| 教育测量与评估双语期刊*, *1*(1), 5.

Thurstone, L. L. (1925). *The fundamentals of statistics* (Vol. 4). Macmillan.

Traub, R. E., & Wolfe, R. G. (1981). Chapter 8: Latent Trait Theories and the Assessment of Educational Achievement. *Review of Research in Education*, *9*(1), 377–435.

Tyler, R. W. (1958). Curriculum organization. *Teachers College Record*, *59*(11), 105–125.

Vo, A. T., Schreiber, J. S., & Martin, A. (2018). Toward a conceptual understanding of evaluative thinking. *New Directions for Evaluation*, *2018*(158), 29–47.

Webb, N. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education*. Research Monograph Number 6. CCSSO.

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*, *28*(March), 1–9.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7–25.

Wyatt-Smith, C., & Adie, L. (2021). The development of students' evaluative expertise: Enabling conditions for integrating criteria into pedagogic practice. *Journal of Curriculum Studies*, *53*(4), 399–419.

Zenisky, A. L., O'Donnell, F., & Hambleton, R. K. (in press). Reporting scores and other results. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th edition). Oxford University Press.