

STEM Education and Measurement, Evaluation and Feedback Processes

Zeynel Amac

Kilis 7 Aralık University, Türkiye

Chapter Highlights

This section highlights key principles, challenges, and innovative frameworks shaping assessment practices in STEM education, with a focus on equity, feedback, and 21st-century skills.

- Assessment in STEM education must shift from typical testing to include a broad range of diagnostic, formative, and summative strategies. They must assess high-level, interdisciplinary skills like critical thinking, collaboration, and applied problem-solving.
- Good feedback is a cornerstone of excellent STEM education. Where timely, specific, and growth-oriented, feedback empowers learners, supports self-regulation, and allows the iterative cycle of inquiry and design that underpins learning in STEM subjects.
- Inclusive STEM Assessment Framework (ISAF) is presented as an integrated framework in response to significant challenges. Its four pillars—Multi-Modal Measurement, Culturally Responsive Evaluation, Dynamic Assessment, and Empowering Feedback—work together to facilitate equitable and rigorous assessment practices.
- Some of the major STEM assessment challenges include interdisciplinary complexity, ensuring reliability and validity for non-traditional activities, teacher assessment literacy, and creating inclusive measures that are equitable for diverse student populations.

To Cite This Chapter:

Amac, Z. (2025). STEM education and measurement, evaluation and feedback processes. In M. T. Hebebcı (Ed.), *Rethinking STEM Education: Theoretical and Sociocultural Frameworks* (129-156). ISRES Publishing

Introduction

As countries struggle with technology problems and knowledge economy transformations, an increasing consensus is that we need STEM-literate citizens. STEM education, however, is much more than just teaching the STEM subjects. Our STEM vision is really more of a transdisciplinary and transferable environment for developing the critical, creative, and problem-solving skills needed to propel society forward (Rosenberg et al., 2018). Education systems worldwide are therefore redefining curriculum in order to become comprehensive STEM experiences that integrate theory with practice and allow students to prepare to navigate complex real-world problem spaces (English, 2016).

This does not mean that the revolutionizing potential of STEM education itself is flawed because, in fact, it is, particularly with regard to measurement. The mean by which measuring, evaluation, and giving feedback occur is perhaps the most vital process that will deliver quality and equity in STEM teaching and education. These procedures will provide evidence of what students can do to affect teaching and curriculum planning, and for helping students reflect upon and build their own performance (Black & Wiliam, 2009). Rather than the traditional, summative assessment and testing view of assessment, emergent models have placed emphasis on formative assessments that help support learners developing deeper learning, self-regulation, and continuous improvement.

Measurement in education is the process of quantifying the knowledge, skills or attitudes that students have systematic processes (Brookhart & Nitko, 2019); evaluation is the process of making value judgments about the quality or effectiveness of the teaching and learning based on measured data (Guskey, 2016). Feedback is information provided to students to facilitate future learning (Hattie & Timperley, 2007); performance assessment involves learners demonstrating skills, or creating products that represent true-to-life, real-world tasks (Darling-Hammond, 2014). Alternative measures are constructed to capture a fuller picture of learning, with an emphasis on creativity, collaboration, and problem-solving besides standardized testing (Torrance, 2012). These are the foundations of a diverse and rich assessment culture in STEM education.

On an international scale, STEM measurement and assessment is one of the priorities in OECD and UNESCO agendas committing to, accountability, lifelong learning, and equity (OECD, 2019; UNESCO, 2017). Measurement and assessment does not only imply endpoints are being measured; they

are processes with scaffolds for the measurement of change to build a knowledge economy that fosters innovation, competitiveness, and social inclusion. STEM assessment and evaluation are related to system change, responsibility, and equity in learning. They serve to identify areas of growth beyond absolute measures, enable international comparability, and guide curriculum reform. Measurement is in quantitative forms and as assessment, measure provides context for areas of growth and therefore use attention for formative and summative assessment activity that can lead to instructional improvement quality (Brookhart & Nitko, 2019). STEM measurement and evaluation are more grounded on lesser pure positivist and more on those that emphasize engagement, equity, and identity that call for the importance of knowledge-construction (Abedi, 2010). Computerization of some of the educational accountability measures provides real-time monitoring and quality measure experience for the students. In addition to traditional relative or absolute assessments, large-scale national and international benchmark assessments now use adaptive testing and learning analytics to deliver personalized, multimodal feedback through real-time, authentic digital evaluation (Ifenthaler & Yau, 2020).

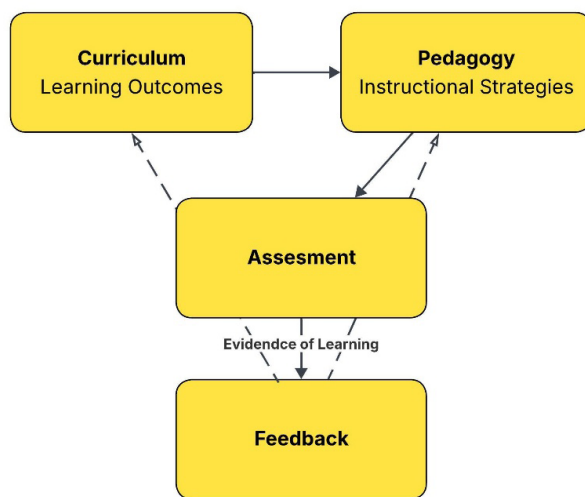


Figure 1. The Curriculum-Pedagogy-Assessment Cycle

Evaluation and assessment offer vital data for shaping policy, allocating resources, and ensuring institutional accountability (Stiggins, 2014). When combined with culturally responsive practices, these processes foster inclusivity by acknowledging and integrating diverse knowledge systems and

forms of representation (Gay, 2018). Curriculum, pedagogy, and assessment are inherently functional and cyclical (Figure 1): The curriculum lays out the intended learning outcomes, pedagogy determines the processes and strategies employed to deliver the lesson, and assessment provides credible evidence used for reflection and evidence to inform/enhance the curriculum and pedagogy. Feedback and the cyclical nature of the three elements ensures that assessment is not simply a phase added onto and is intrinsic to teaching and student success.

Measurement and Evaluation Types, Strategies, Methods, and Techniques

In STEM education, assessment extends far beyond traditional testing to encompass a diverse array of evaluation and measurement methods tailored to interdisciplinary, skills-oriented learning. These approaches are designed not only to measure outcomes but also to actively support the development of critical thinking, collaboration, creativity, and problem-solving skills. Because STEM education emphasizes authentic, inquiry-based experiences, assessment must similarly evolve to capture complex competencies through both formative and summative strategies. The following section and Table 1 (Adapted from (Black & Wiliam, 1998; Brookhart, 2017; Bybee, 2013; Popham, 2009; Stiggins, 2014) present key assessment methods, including their definitions, advantages, limitations, and practical examples, to assist educators in designing meaningful, research-informed evaluations that align with contemporary STEM learning goals.

Diagnostic Assessment

Diagnostic assessment focuses on mapping students' prior knowledge and misconceptions, as delineated in Table 1. This preventative assessment can then better inform the pedagogy of practitioners, as they know precisely what manner of teaching practice can fill the gaps and what prior knowledge can be built on (Brookhart & Nitko, 2019). Pre-tests are a clear example as they give a quantitative picture of where a student is at in specific prerequisite skills - i.e., for a calculus unit, an algebra pre-test identifying the gaps in knowledge. Whereas concept maps develop a qualitative understanding of the student, as to how they have structured their own cognition, and how they have laid out the connection between ideas within a discipline, for instance, producing a concept map for parts of the human nervous system before studying in-depth it (Novak & Cañas, 2008). While being valuable for developing individualized learning pathways, Table 1 also appropriately points out the weaknesses: A pre-test may not predict future performance,

Table 1. Measurement and Evaluation Types, Methods, Advantages, Disadvantages and Examples

Type	Method / Tool	Definition	Advantages	Disadvantages	Example in STEM
Diagnostic	Pre-Test	Pre-assessments to identify prior knowledge or misconceptions before instruction.	Tailors instruction to student needs; prevents learning gaps.	May not predict future performance; time-intensive to design.	A pre-test on algebraic functions in a calculus class to identify gaps before teaching derivatives.
	Concept Map	A visual diagram that illustrates the relationships between concepts and ideas, often represented in a hierarchical structure with connecting lines and linking words.	Effectively evaluate a student's comprehensive knowledge structure, uncover hidden misunderstandings, and directly assess critical analytical and synthesis skills.	The assessment is often labor-intensive and subjective, requires significant time for large groups, and demands prior student training to be effective.	Mapping the human nervous system. Students create a map linking central concepts (Central Nervous System, Peripheral Nervous System) to sub-concepts (neurons, brain, spinal cord, senses) with labelled connections (e.g., "contains", "sends signals to").
	Standardized Testing	Uniform tests to measure STEM proficiency across populations.	Comparable across contexts; aligns with standards.	May not capture interdisciplinary skills; risk of teaching to the test.	An AP Computer Science exam assessing coding and debugging skills.
Summative	Lab Report	Assessing student achievement at the end of a unit or a course to determine mastery of objectives.	Measures overall competency; useful for grading and certification; aligns with standards.	Limited feedback for improvement; may not capture learning process.	A final physics lab report analyzing pendulum motion to evaluate understanding of harmonic motion.
	Project-Based Assessment	Tasks requiring students to integrate STEM knowledge in real-world projects.	Mirrors real-world applications; fosters collaboration and problem-solving.	Resource-intensive; subjective scoring without clear rubrics.	Designing a model bridge in an engineering class, assessed via rubric for design and teamwork.

Summative	Performance-Based Assessment	Hands-on tasks demonstrating practical STEM skills.	Assesses application of knowledge; authentic to STEM practices.	Time-consuming; requires clear criteria to avoid bias.	Conducting a titration experiment in chemistry to demonstrate lab techniques and data analysis.
	Portfolio	Collecting student work over time to show growth and competencies.	Captures long-term progress; interdisciplinary focus.	Time-intensive to compile and assess; requires clear evaluation criteria.	A robotics portfolio including code iterations and prototype designs to show skill development.
	Clicker Questions	Ongoing assessments during instruction to provide feedback and guide teaching adjustments.	Enhances student engagement; allows real-time instructional changes; fosters learning growth.	Time-consuming; requires teacher expertise to implement effectively.	Using clicker questions in a biology class to assess understanding of ecosystems and adjust the lesson.
	Observation	Capturing non-numerical data through observations, reflections, or portfolios to assess processes and attitudes.	Offers deep insights into complex STEM skills; supports interdisciplinary assessment.	Subjective; time-intensive to analyze; less standardized.	Evaluating design journals in an engineering project to assess students' problem-solving process.
Formative	Peer and Self Assessment	Students evaluating their own or peers' work to foster metacognition.	Promotes reflection and accountability; enhances student engagement.	Subjective; requires training for reliability.	Peer review of contributions in a group solar-powered car project.
	Technology-Enhanced Assessment	Using digital tools like simulations for dynamic evaluation.	Engages students; mirrors real-world STEM tools; provides instant feedback.	Requires technology access; potential equity issues.	A PhET simulation assessing physics concepts through virtual pendulum experiments.

Formative	Rubric	A scoring guide that uses specific pre-defined criteria to evaluate student work, with detailed descriptions of performance levels for each criterion.	Enhances grading fairness, provide clear feedback, clarify goals, and streamline the evaluation process.	The development requires significant time, they can stifle creativity, and overly strict criteria may hinder a holistic assessment.	Evaluating a lab report. Criteria include Hypothesis, Methodology, Data Analysis, Conclusions, and Safety. Each has performance levels (e.g., Novice, Proficient, Exemplary) with clear descriptors.
	Rating Scale	An assessment tool that uses a set of categories designed to measure the degree to which a behavior, attitude, or skill is present or performed.	Efficient to use, effectively assess observable behaviors and attitudes, and provide quantifiable data for progress monitoring.	They are susceptible to rater bias, offer little constructive feedback, and can suffer from inconsistent interpretation among evaluators.	Assessing collaboration in a robotics project. A scale from 1 (Rarely) to 5 (Always) is used to rate behaviors like "Shares ideas effectively," "Listens to team members," and "Completes assigned tasks on time."
Program Evaluation	Data Analysis	Judging program or curriculum effectiveness using aggregated data.	Informs systemic improvements; supports accountability.	Broad focus may overlook individual needs; requires robust data collection.	Analyzing student performance in a robotics curriculum to evaluate program effectiveness.

or analyzing concept maps may take too much time and be too subjective that students (and teachers) may not be able to be educated effectively.

Summative Assessment

Summative assessment, which involves measuring student achievement at the end of a period of learning, has a range of methods that certify mastery and include accountability data (Pellegrino et al., 2001). Summative assessment in STEM instruction is very different from simply giving students a final exam. As can be seen in Table 1, summative assessments include a variety of authentic assessments that address the integrated nature of STEM. Similarly, standardized tests (e.g., AP Computer Science exams) afford reliable, comparable data across populations, but also run the risk of missing interdisciplinary application and the promotion of “teaching to the test” (Koretz, 2017). Performance assessments and project-based assessments are the best forms of summative assessment to measure students’ application of knowledge and real-world skills (Darling-Hammond, 2014). Additionally, portfolios offer a more complete and qualitative picture of long-term growth and competency development by including selections of work over time (Barrett, 2007), such as iterations of code and design documents from a robotics class. Although these techniques present a fuller picture of learning, these techniques require resources and rely on well-defined rubrics to support consistent and objective scoring (Panadero & Jonsson, 2013). These methods consider not just the final product, but also the strength of the inquiry and collaboration.

Formative Assessment

Formative assessment consists of a set of continuous assessment strategies as described in Table 1 that are designed to occur during instruction and offer immediate feedback, allowing for modifications in instruction (Black & Wiliam, 2009). In STEM education, they especially help students continuously grow their conceptual understanding. There are a number of ways to gain timely and flexible information about students’ developing ideas. Clicker questions and other technology-based assessments (PhET, etc.) offer quick feedback to check for understanding and make adjustments to teaching in real-time (Shute & Rahimi, 2017). More qualitatively, allowing students to show their work during the processes, and involving students’ peers in self-assessments or peer assessments, introduce incentive toward metacognition, accountability and developing collaborative skills (Topping, 2010) which are critical skills for STEM practices. Some rubrics and rating scales are very formative in nature and almost more structure and less standards to clarify

the goals and expectations of students and simplify the evaluation of work done for students. Nevertheless, while Table 1 addresses the importance of teacher knowledge to carry out these methods effectively, they also can present significant time demands. In inquiry-based STEM classroom settings, these formative assessments are vital to making decisions effectively about the next instructional action, permitting time for self-regulation, and investigating deeper problem-solving (Heritage, 2022).

Program Evaluation

Program evaluation verifies how well a program or curriculum functions by utilizing hyphenated data as the central focus, as indicated in Table 1 by the method of data analysis (Popham 2009). Although evaluative assessment considers the individual learners' outcomes of learning to guide decision-making at a bigger scale, it principally synthesizes outcomes and assists in making intelligent decisions about policies, the use of resources, and upgrades. The analysis of how data occurs is the central method through which evaluative assessment occurs through tallying and observation of aggregated data through numerous sources. The quantitative analysis examines data gathered as student performance measures and examined collectively. The key advantage is that it may bring about change at the whole system level and hold individuals accountable. Validity questions are significant when numerous sources and systems are needed to gather and report credible data (Madaus et al., 1983). Ultimately, evaluative assessment assembles significant evidence to respond to questions such as, "Is our STEM program accomplishing its objectives?"

Alternative Assessment

In the case of STEM, alternate assessment can reveal what students know and how students learn that cannot be revealed by regular testing. It measures real-world skills, how students do what they do, and how they experience something. The issue of whether or not to treat "Alternative Assessment" as a category is a complicated one. In contemporary literature about assessment, it is an umbrella term that includes any assessment that is an alternative to traditional standardized testing and multiple-choice exams. In this sense, most, if not all, in the list of methods from Table 1, such as project-based assessments, portfolios, concept maps, peer review, etc., would fall under the umbrella of alternative assessment. They are considered "alternative" because they provide a fuller, authentic picture of learning instead of focusing primarily on rote memorization, by valuing process, creativity, and application (Darling-Hammond, 2014).

In summary, diagnostic assessments, conducted before or at the start of instruction, reveal students' existing knowledge, misunderstandings, and readiness (Brookhart & Nitko, 2019). A math pretest before an engineering project, for instance, may inform interventions, while algebra diagnostic assessments before robotics operations help group learners based on their skills. These types of assessments not only offer quantitative results but also have qualitative elements, such as teamwork, equivalence problem identification, and developmental benefit based on Vygotsky's Zone of Proximal Development. Their contemporaneous nature makes them downright crucial in preparing students whenever formative or summative assessments are not available.

Mixed methods combined measurement also strengthens measurement, since cognitive, procedural, and affective learning can be best understood by more than one approach. Quantitative measures such as rubric scores and accuracy checks and qualitative measures such as reflective journals and field notes provide richer accounts of learning without overwhelming the assessment design. Interdisciplinary projects benefit the most, since both process dynamics and product quality are measured. In science, for example, laboratory notebooks qualitatively log hypotheses, while experiments are quantitatively measured. In technology, testing automatically verifies coding correctness, while peer reviews gauge creativity, elegance, and efficiency. By blending performance-based and reflective data, mixed methods more equally balance correctness and style.

Innovative Tools and Technological Applications to Support Assessment

Digital tools enhance STEM assessment through immediate data collection, adaptive feedback, and multimodal representation of learning. Typical tools include digital rubrics, electronic portfolios, and learning analytics. Digital rubrics, integrated via learning management systems or designated software, make it possible to conduct targeted assessment, take notes, monitor progress, and report to parents and students (Marzano, 2006). Multimedia-capable e-portfolios showcase STEM competencies in evidence-based form, while learning analytics quantify engagement through tracking variables like simulation use, correctness of problem-solving, and task duration of engagement (Long & Siemens, 2014). These technologies support teachers' decisions on scaffolding and differentiation based on data. Application-based assessments extend these possibilities through mobile apps for fieldwork, coding apps with debugging diagnostics, and simulation software for iterative design (Blikstein, 2013). Resources such as Tinkercad

and Scratch demonstrate this by enabling computation modeling and also providing teachers with an idea of design iteration and problem-solving exercises.

Feedback Mechanisms in STEM Education

Feedback is an essential strategy that not only assesses student work but actively develops and builds learning in education. As STEM learning environments move toward more process than product, creativity than compliance, and inquiry than instruction, feedback emerges as a dynamic process negotiating instruction and metacognition (Hattie & Timperley, 2007). In contrast to summative evaluation, feedback, when prompt, personalized, and constructive, assists learners as they progress through experimental cycles of failing, reflecting, and revising.

A number of interrelated attributes characterizes good feedback: Timeliness, specificity, clarity, and personalization. Feedback is best delivered in real time or as soon as possible so that students can relate the information to the activity being undertaken and correct it in the process (Shute, 2008). In STEM classrooms, feedback allows students to explore their way through iterative refinement (Brookhart, 2017). Clarity and constructive language must also be present. “Correct” or “incorrect” is sometimes all that differentiates the feedback provided on an outcome. However, students do recognize diagnostic feedback as quality assessment that systematically signals strengths, weaknesses, and specific follow-up actions (William, 2011). Finally, feedback is also most helpful for promoting longer-term achievement and motivation when it is personalized to the student at hand, perhaps in regards to variation in their growth or learning, or in regards to an aspect of conceptual misunderstanding (Nicol & Macfarlane-Dick, 2006).

The Influence of Feedback on Learning

Feedback is part of STEM pedagogy for deep learning, if even in regards to the how and why of students’ work, not just the what. Feedback is the number one single school level factor that can influence student outcomes according to Hattie’s (2008) meta-meta-analysis study. In non-negotiable linear teaching and learning characteristics in STEM settings, perhaps more than any other discipline or degree of academia, learning is literally most of the time a linear experience of needing to work through ignorance, and conceptualizing what they thought they knew in terms of evidence-based experimentation.

Feedback also generated in the service of a mastery climate can overcome

the test anxieties and the competitive nature of STEM culture. Feedback facilitates students' self-regulated learning by helping to apply criteria for assessments to the monitoring of one's own performance (Zimmerman & Schunk, 2001). With reflection, goal setting, and discussion that activates a sense of responsibility for learning, feedback can help learners take responsibility for the process of learning. Through role modeling of metacognition and soliciting students' feedback practices, teachers form a routine of self-regulation. Self-regulation supports include reflection diaries and self-assessment checklists that help students identify areas of weakness and work towards developing an improvement plan. These self-regulatory methods aid in the area of STEM, where achievement is usually contingent on planning, monitoring, and revision of procedures (Nicol, 2010).

Placing the Student at the Center of the Assessment Process

The main objective of feedback is to develop student autonomy, which is achieved by making students responsible for their learning (Nicol & Macfarlane-Dick, 2006). For this, students must move from being passive consumers of assessment to active participants in assessments. As well as its value, what matters is how feedback is perceived and acted upon (Carless & Boud, 2018). The efficacy of feedback can also depend on a student's mindset, receptiveness to criticism, and commitment to their goals. In addition, developing student "assessment literacy" has been shown to help students understand criteria, understand where they are and where they want to be, and act on feedback. Teachers can promote assessment literacy by using self-assessment, self-reflection, and peer-assessment and develop feedback as a guide to learning rather than assigning a grade.

Real-Time and Continuous Feedback Approaches in STEM Education

Because of the process-based nature of STEM learning, feedback is often needed to be continuous and embedded within instruction. Real-time feedback can provide immediate cognitive redirection without disrupting the flow of engagement (Ruiz-Primo & Furtak, 2007). Technology-enhanced formative assessment tools like learning management systems, computer simulations, and data dashboards enable educators to offer personalized feedback in a convenient way. For example, web-based applications like Desmos, PhET simulations, or Tinkercad allow teachers to track student work and provide context-specific intervention (Blikstein, 2013). Not only do these sites enable feedback, but they also record rich data that inform

instruction. Audio or video comments can be used to personalize responses in STEM project works, particularly in evaluating multimedia portfolios or explaining nuanced errors in technical work. It can enhance class engagement and provide a more personalized teaching presence, especially for blended or online STEM education (Mahoney et al., 2019).

Examples of Teacher, Peer, and Student Feedback in STEM

Feedback in STEM education occurs in many different formats. Teacher feedback may take the form of a mini-conference, written comments on a lab report, or a project graded with a rubric that reflects expectations and practices within the discipline. Peer feedback applies the principles of social learning. While some feedback processes are informal, e.g., “two stars and a wish,” there are structured ways to provide peer feedback, e.g., “TAG: Tell something you like, Ask a question, Give a suggestion.” Structured feedback opportunities encourage students to develop critical evaluation skills while building a community of learners that responds to feedback (Topping, 2010).

Table 2. Examples of Effective and Ineffective Feedback in STEM

Context	Ineffective Feedback	Effective Feedback
Lab Report	“Your analysis is weak.”	“Accurate data collection. Connect your findings to chemical equilibrium from Chapter 3 to strengthen conclusions.”
Coding Project	“Your code is messy.”	“Program runs correctly. Improve readability with comments and descriptive variable names.”
Engineering Prototype	“This design won’t work.”	“The single beam buckled—consider using triangular supports to distribute load and improve stability.”
Solar System Model	“Your planets are out of order.”	“Your model is very creative! For better accuracy, remember the acronym ‘My Very Educated Mother Just Served Us Noodles’ to recall the planet order.”
Paper Bridge Project	“This isn’t strong enough.”	“Great start on your paper bridge! To hold more weight, try folding the paper into a U-shape or a tube to make a stronger beam.”

Self-feedback encourages metacognition, and allows students to self-reflect using goal-setting worksheets, reflection journals, or analyzing errors. Each of these means requires students to identify strengths and weaknesses, analyze the cause of any weaknesses, and make plans for improvement. Self-feedback develops the metacognitive skills required of students in a cyclical STEM process. Table 2 provides some examples of effective vs. ineffective feedback in STEM.

Evaluation and Assessment Models and Best Practices Globally

STEM education utilizes various ways of assessment, feedback, and measurements based on national contexts, policies, and cultures (OECD, 2020). These distinctions provide adaptable models and generic principles for local reform (Breakspear, 2012). Table 3 portrays how some countries balance centrally mandated examinations with student-centered approaches that include the agency of teachers. An example of a student-centered model is Finland, which has very little national standardized testing, and uses mostly formative assessment (Sahlberg, 2021). South Korea has a similar history of examinations and with various reforms now encourages project-based learning by mandating that schools engage in creative experiential activities (Hong, 2021). In the United States, a federal system promotes high degrees of variability amongst states, especially in light of the broad framework for multidimensional assessment offered by the Next Generation Science Standards (NGSS), which most states have not adopted (NGSS Lead States, 2013).

Global assessments, such as PISA and TIMSS, signify quality in STEM education and provide benchmarks for policy-making and reform based on applied knowledge and curriculum-based achievement (OECD, 2019; Breakspear, 2012). Effective evaluation systems extend well beyond the classroom to professional development and teacher education (OECD, 2020). Finland has included evaluation literacy within their teacher education (Sahlberg, 2021) and South Korea invests in high-quality professional development which may facilitate schools (Hong, 2021). Overall, these examples indicate that for the objectives of equity and quality in STEM evaluation to be realized there must be coherence across policy, practice, and pedagogy (OECD, 2020). International models can be comparatively summed up in a table as follows:

Table 3. Comparative Approaches to STEM Assessment

Country	Dominant Strategies	Key Challenges	Outcomes/ Strengths
United States	NGSS-aligned performance tasks; rubrics	Local variability; equity gaps	Innovation, diverse practices
Finland	Formative, portfolio-based	Limited comparability across systems	Strong teacher autonomy, equity focus
South Korea	Exams + creative experiential activities	Transitioning from exam-heavy culture	High achievement, gradual shift to innovation
Singapore	Inquiry-based, high-stakes integrated	Pressure from exams; teacher workload	Strong international performance
Canada	Competency-based, inclusive assessments	Provincial variability	Equity-oriented and culturally responsive
Australia	National curriculum, digital tools	Balancing standardization vs. innovation	Strong integration of ICT in assessment

At the system level, STEM national assessment policy is incorporated into education reform and teacher development agendas overall. As a case in point, in South Korea, continuous professional development of STEM teachers involves specialized training for designing and deploying assessment instruments. Finland puts emphasis on assessment literacy as part of initial teacher training, encouraging theoretical and practical competencies in non-traditional modes of assessment (OECD, 2020). Private funding initiatives as well as public policies in the United States enable the implementation of school-level STEM testing systems with a particular focus on genuine, inquiry-based assessment practices (Honey et al., 2014).

All these global models collectively highlight the pressing need for aligning macro-level policy decisions with the development of institutional capacity and classroom-level instructional designs. The US example highlights the federal policy provincials' importance in fashioning locally-grounded innovation; the Finnish example offers a powerful vision of teacher-directed, trust-based systems of assessment; and South Korea illustrates how top-down policy reforms can stimulate teaching innovation in the face of systemic constraints. Overall, these examples highlight the importance of policy,

pedagogy, and practice coherence in the pursuit of equitable, high-quality STEM assessment.

Hattie and Timperley's (2007) highly applied model provides an organized method for learning feedback. Their "three questions" approach—Where am I going? How am I going? Where to next?—provides clarity on learning goals, presents evidence of journeying, and plots courses for future development. Translated to STEM contexts, feedback can assist with mathematics iterations of hypothesis construction, prototyping, and refinement.

It also must be distinguished from formative and summative feedback. Formative feedback is provided during learning and is designed to improve processes, while summative feedback is typically backward-looking, presenting evaluative judgments following completion of tasks (Heritage, 2022). Formative feedback is especially critical in STEM because of the trial-and-error nature of scientific discovery and design. Finally, cultural influences shape the perception and perception of feedback. Feedback in East Asian environments is indirect, embedded in shared performance expectations, while Western cultures emphasize direct, individualized comments. Accounting for these kinds of cultural variations is essential while developing feedback systems for more globalized and diverse STEM classrooms.

Challenges in Measurement and Evaluation within STEM Education

STEM education, given its interdisciplinary nature, practice-based learning methods, and focus on higher-order thinking abilities, presents challenges to the field of assessment and evaluation (Honey et al., 2014). This subsection will explain the origins of issues surrounding the issue of assessment and evaluation in STEM education, followed by educational practice implications.

Interdisciplinary Complexities

STEM education links science, mathematics, technology, and engineering as one learning framework, even though the disciplines have different epistemologies, knowledge building practices, and assessments in isolation. Interdisciplinary assessments are challenging to develop, as there are notable differences for assessment such as assessing the engineering design process as opposed to mathematical problem solving. Integrated rubrics are necessary to assess however, both discipline-specific competency and

transdisciplinary competency like problem solving and collaboration. This type of assessment will account for the diverse and complex learning and STEM related outcomes that may develop within interdisciplinary projects by assessing how knowledge has been used across multiple fields. This approach will provide a clear way to assess and measure STEM learning outcomes.

Objectivity, Reliability, Validity Issues

In STEM assessments, open-response items and project-based learning are prone to subjectivity as there is often no clear correlation to the learning objectives (Mohommadi et al., 2022). Rubrics serve to increase reliability but there are few STEM specific rubrics that could threaten validity, especially with respect to emergent outcomes like creativity or collaboration. Collaboratively working with teachers to co-construct rubrics, calibrate against exemplars, or moderation sessions can increase reliability and assist with subjectivity. Collectively, they would improve the objectivity and validity of STEM assessment.

Teacher Competence and the Limitations of Assessment Tools

The pedagogical skill of STEM instructors in assessment and evaluation is crucial to the process's effectiveness. However, the majority of teachers, particularly when handling project or process-based assessments, report insufficient knowledge and training in measurement techniques (Plake & Impara, 1996). This shortage incapacitates them both in the production of quality instruments as well as in justifiably interpreting the results of their tests. Moreover, none of the current tools is holistic enough to encompass interdisciplinary learning, while none provides a system for tracking students' process of learning (National Research Council, 2011). Thus, this indicates the urgency need for constant, proactive professional development courses for educators that can exemplify the ways to build and score complex performances tasks.

Student Diversity and Inclusive Assessment

New cultural, linguistic, and cognitive diversities of the STEM classrooms demand inclusive measurement (Abedi, 2010). Existing assessment methods and techniques fall short in reflecting the knowledge all students have. In the case of special needs students, immigrant, and LGBTQ students for example, if such students are to be provided accessible and flexible assessments, the assessment will be flexible and provide choices, differentiation, assessment

accommodations and continuous monitoring for inclusivity. Schools, staff, and policymakers can collaborate in addressing assessment to serve diverse learners better and to maximize equity in STEM education.

Challenges in Data Analysis and Interpreting Results

Since learning analytics increasingly depend on STEM education, the teacher is not statistical and technological literacy competent enough to play around with data, which may lead to data being interpreted incorrectly and, more importantly, limiting pedagogical potential (Bienkowski et al., 2012). Ethical issues relating to student privacy, student information ownership, and algorithmic bias mean that there need to be sufficient policies in place to ensure accountability, consent, and transparency. High-stakes standardized tests will inevitably focus on surface performances, usually degrading STEM education to accountability rather than authentic creativity. Facilitating some concomitant literacy in assessment will thus require teacher professionalism and social obligation, technology-enabled practice awareness (DeLuca & Klinger, 2010). Artificial intelligence software will enable assessment of high-level STEM products of students, but biased data sets and black box methods--the "black box problem" will ensure that it will render null and void that would impose a commitment to Explainable AI (XAI) for fairness. A normative framework of transparency, justice, autonomy and beneficence will be required for equitable digital STEM assessments

Towards an Inclusive STEM Assessment Framework

The issues explained above reveal that the existing STEM assessment approaches are not productive. Rather than dealing with these matters separately as independent units of concern, we need a framework, which places these concerns altogether in an integrated, adaptable and principle-guided way. This section builds on the idea of system response to such issues and identifies urgency gaps in the profession with an Inclusive STEM Assessment Framework (ISAF). The problems we have outlined can be reframed on the basis of principles for a more equitable and effective assessment system than problems. ISAF, for example, is comprised of a pillar, the "multimodal assessment," which marks a potential to address interdisciplinary complexity and the "Culturally Responsive Assessment" pillar that addresses diversity of students as an asset.

From the gaps and best practice identified in international models (OECD, 2019), the Inclusive STEM Assessment Framework (ISAF) focuses on a commitment to equity and engagement, which is the basis for inclusive

education (Ainscow, 2020) and is consequently changing the framework from an assessment tool to a change tool. The ISAF has four interrelated pillars:

Multi-Mode Measurement: This is about going beyond normed tests and application of full range of evidence (multimodal) (Pellegrino et al., 2001). It includes quantitative data (scores, analytics), qualitative data (journals, observation), and participatory data (self/peer-assessment) to collect the full range of student skill, particularly from students who come from various cultural, language, and cognitive backgrounds (Abedi, 2010).

Culturally Responsive Evaluation: Evaluation criteria and rubrics must be co-constructed with students to reflect diverse ways of knowing and problem-solving (Gay, 2018; Ladson-Billings, 1995). This approach aligns with the principles of inclusive education, which advocates for the removal of barriers to participation and learning for all students (Ainscow, 2020). This involves recognizing that ‘valid’ solutions can be presented in different forms (narrative, graphical, prototype, code) and that context is a critical component of judging quality.

Dynamic Assessment: Assessment is not a singular event but an ongoing, dialogic process integrated into the learning cycle (Black & Wiliam, 2009). This pillar emphasizes diagnostic and formative functions, where assessment is used to scaffold learning in real-time, adapting to the learner’s zone of proximal development (Lantolf & Poehner, 2004).

Empowering Feedback: Feedback mechanisms are designed to be timely, specific, and growth-oriented (Shute, 2008). More importantly, they must be accessible and actionable for all learners (Banks, 2016). This involves leveraging technology for personalized pathways and ensuring language and delivery modes (audio, video, text) are tailored to individual learner needs, fostering self-regulation and a growth mindset (Zimmerman & Schunk, 2001). The interconnected and cyclical nature of these four pillars is visualized in Figure 2:

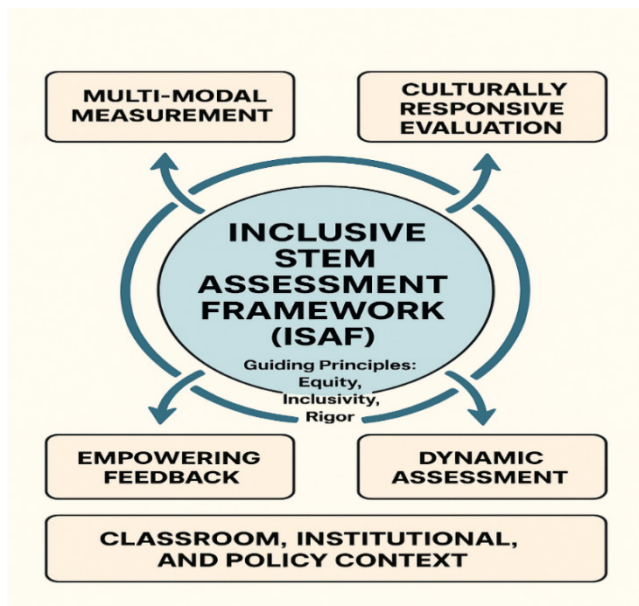


Figure 2. The Inclusive STEM Assessment Framework (ISAF)

The ISAF is a cyclical and dynamic model consisting of four core pillars that interact continuously. The process begins with Multi-Modal Measurement to gather diverse evidence of learning. This evidence is then judged through Culturally Responsive Evaluation based on co-constructed criteria. The evaluation informs Dynamic Assessment, an ongoing process of diagnosis and scaffolding that shapes the learning journey. Throughout this journey, Empowering Feedback is provided to foster growth and self-regulation. The insights from feedback then inform the next cycle of measurement, creating a continuous feedback loop. All pillars are underpinned by the core principles of equity, inclusivity, and rigor, and the framework operates within and is influenced by broader classroom, institutional, and policy contexts.

In this framework, “rigor” refers not only to the difficulty of content knowledge as traditionally understood, but also to the cognitive depth and complexity of skill application (Hess, 2009). ISAF redefines rigor by assessing the extent to which students demonstrate higher-order thinking skills, such as analyzing complex problems, designing creative solutions, synthesizing interdisciplinary knowledge, and constructing evidence-based arguments, rather than simply measuring their capacity to memorize information (Brookhart, 2017). Therefore, this model aims to measure deeper and more authentic learning rigor through multimodal and inclusive methods, as real-world STEM problems are rarely solved with standardized test

items that have a single correct answer (Pellegrino & Hilton, 2012). This framework does not prescribe a single tool but advocates for a principle-based approach. It urges educators and policymakers to design assessment systems that are not only rigorous but also inherently equitable, ensuring that every student has the opportunity to demonstrate their STEM literacy in multiple ways (Rosenberg et al., 2018).

Conclusion

In light of the challenges and opportunities examined in this chapter, following concrete steps are suggested for teachers, administrators, and policy makers to transform the culture of assessment in STEM education.

For Teachers: Incorporate simple tech tools by using free platforms like Socrative or Mentimeter for quick formative checks during lessons. Utilize Padlet, Miro, FigJam, Lucidspark, Microsoft Whiteboard, and Canva for collaborative brainstorming and formative feedback on project ideas. Adopt a feedback protocol by implementing structured methods like “Two Stars and a Wish” (for peer feedback) or “What? So What? Now What?” (for self-reflection journals) to make feedback consistent and developmentally focused. Try a digital portfolio by using applications like Google or Seesaw to have students compile their work. This approach supports a complete evaluation of growth and is an easy way to give feedback on the student’s artifacts (audio/video).

For Administrators: Prioritize professional development on:

1. Rubric Calibration: The procedure to ensure scoring consistency across teachers.
2. Interpret Learning Analytics: Being able to read dashboards from LMS platforms.
3. Culturally Responsive Assessment: The way in which to implement fair tasks and assessments.
4. Audit Assessment Tools/Software: Once a year audit school software and tools, so your school or district is compliant with student data privacy (FERPA in the United States or GDPR in the EU).
5. Encourage Teacher Collaboration: In order for the PLCs (Professional Learning Communities) to analyze student assessment data and moderate the scoring of student work to ensure reliability and share best practices, allow for dedicated time to work together.

For Policymakers: Create grants for schools to pilot new assessment models. For example, model badges to act as micro-credentials for specific STEM skills and produce competency-based progression models. Mandate,

and fund the creation of, clear enforceable ethical guidelines for EdTech procurement to ensure that every tool used in a public school will meet strict standards for data privacy, algorithmic fairness and accessibility. Re-frame accountability metrics by going beyond standardized test scores as the primary measure of school success. Create a balanced dashboard of indicators that include, commonly student engagement in STEM projects, participation in science fairs, and measured growth in portfolios.

Acknowledgements

I would like to extend my sincere gratitude to the editors for their thoughtful guidance and to the anonymous reviewers for their constructive and insightful feedback. Their careful reading and valuable suggestions greatly contributed to strengthening the clarity, coherence, and overall quality of this chapter.

In preparing and refining the manuscript, I also made use of advanced language and writing tools, including ChatGPT and Gemini, to assist with paraphrasing, enhancing readability, and polishing the academic style. All final interpretations, revisions, and substantive content remain entirely my own responsibility.

References

- Abedi, J. (2010). Performance assessments for English language learners. *Stanford Center for Opportunity Policy in Education*. Retrieved August 6, 2025, from <https://edpolicy.stanford.edu/sites/default/files/publications/performance-assessments-english-language-learners.pdf>
- Ainscow, M. (2020). Promoting inclusion and equity in education: Lessons from international experiences. *Nordic Journal of Studies in Educational Policy*, 6(1), 7-16. <https://doi.org/10.1080/20020317.2020.1729587>
- Banks, J. A. (2016). *Cultural diversity and education: Foundations, curriculum, and teaching* (6th ed.). Routledge.
- Barrett, H. C. (2007). Researching electronic portfolios and learner engagement: The REFLECT Initiative. *Journal of Adolescent & Adult Literacy*, 50(6), 436-449.
- Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. U.S. Department of Education, Office of Educational Technology.
- 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>
- Blikstein, P. (2013). Digital fabrication and 'making' in education: The democratization of invention. In J. Walter-Herrmann & C. Büching (Eds.), *FabLab: Of machines, makers and inventors* (pp. 203-222). Transcript Publishers.
- Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance*. OECD Publishing. <https://doi.org/10.1787/5k9fdfqffr28-en>
- Brookhart, S. M. (2017). *How to give effective feedback to your students* (2nd ed.). ASCD.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students*, (8th Ed.). Pearson.
- Bybee, R. W. (2013). *The case for STEM education: Challenges and opportunities*. NSTA Press.
- 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>

Darling-Hammond, L. (2014). Introduction: The rationale and context for performance assessment. In L. Darling-Hammond, & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 1-14). Wiley & Sons.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, 17(4), 419-438. <https://doi.org/10.1080/0969594X.2010.516643>

English, L. D. (2016). STEM education K-12: Perspectives on integration. *International Journal of STEM Education*, 3(3), 1-8. <https://doi.org/10.1186/s40594-016-0036-1>

Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice* (3rd ed.). Teachers College Press.

Guskey, T. R. (2016). How classroom assessments improve learning. In M. Scherer (Ed.), *On formative assessment: Readings from Educational Leadership* (pp. 3-13). ASCD.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

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

Heritage, M. (2022). *Formative assessment: Making it happen in the classroom* (2nd ed.). Corwin.

Hess, D. E. (2009). *Controversy in the classroom: The democratic power of discussion*. Routledge.

Honey, M., Pearson, G., & Schweingruber, H. (Eds.). (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research*. National Academies Press. <https://doi.org/10.17226/18612>

Hong, O. (2021). STEM/STEAM education research in South Korea. In T. W. Teo, A.-L. Tan, & P. Teng (Eds.), *STEM education from Asia trends and perspectives* (pp. 211-227). Routledge. <https://doi.org/10.4324/9781003099888-11>

Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, 68(4), 1961-1990. <https://doi.org/10.1007/s11423-020-09788-z>

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

University of Chicago Press.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, 32(3), 465-491. <https://doi.org/10.3102/00028312032003465>

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, 1(1), 49-72.

Long, P., & Siemens, G. (2014). Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5), 132-137.

Madaus, G. F., & Stufflebeam, D. L., Scriven, M. S. (1983). Program evaluation: A historical overview. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 3-22). Kluwer Academic.

Mahoney, P., Macfarlane, S., & Ajjawi, R. (2019). A qualitative synthesis of video feedback in higher education. *Teaching in Higher Education*, 24(2), 157-179. <https://doi.org/10.1080/13562517.2018.1471457>

Marzano, R. J. (2006). *Classroom assessment and grading that work*. ASCD.

Mohammadi, A., Fagiolini, A., Cirrincione, M., Garone, E., Garone, A., & Varagnolo, D. (2022). Towards an open database of assessment material for STEM subjects: requirements and recommendations from early field trials. *IFAC-PapersOnLine*, 55(17), 7-12. <https://doi.org/10.1016/j.ifacol.2022.09.217>

National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press.

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, 35(5), 501-517. <https://doi.org/10.1080/02602931003786559>

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218. <https://doi.org/10.1080/03075070600572090>

Novak, J. D., & Cañas, A. J. (2008). *The theory underlying concept maps and how to construct and use them*. Institute for Human and Machine Cognition.

<https://cmap.ihmc.us/docs/theory-of-concept-maps>

OECD. (2019). *PISA 2018 results (Volume I): What students know and can do*. OECD Publishing. <https://doi.org/10.1787/5f07c754-en>

OECD. (2020). *TALIS 2018 results (Volume II): Teachers and school leaders as valued professionals*. OECD Publishing, <https://doi.org/10.1787/19cf08df-en>

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>

Pellegrino, J. W. , Chudowsky, N. , & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

Plake, B. S., & Impara, J. C. (1996). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement, and adjustment* (pp. 53-68). Academic Press.

Popham, W. J. (2009). *Assessment literacy for teachers: Faddish or fundamental? Theory into Practice*, 48(1), 4-11. <https://doi.org/10.1080/00405840802577536>

Rosenberg, M. B., Hilton, M. L., & Dibner, K. A. (2018). *Indicators for monitoring undergraduate STEM education*. The National Academies Press. <https://doi.org/10.17226/24943>

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment 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>

Sahlberg, P. (2021). *Finnish lessons 3.0: What can the world learn from educational change in Finland?* (3rd ed.). Teachers College Press.

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

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, 33(1), 1-19. <https://doi.org/10.1111/jcal.12172>

Stiggins, R. J. (2014). *Revolutionize assessment: Empower students, inspire learning*. Corwin.

Topping, K. J. (2010). Peers as a source of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 61-74). Routledge.

Torrance, H. (2012). Formative assessment at the crossroads: Conformative, deformative and transformative assessment. *Oxford Review of Education*, 38(3), 323-342. <https://doi.org/10.1080/03054985.2012.689693>

UNESCO. (2017). *A guide for ensuring inclusion and equity in education*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000248254>

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.

Zimmerman, B. J., & Schunk, D. H. (2001). Reflections on theories of self-regulated learning and academic achievement. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 273-292). Routledge.

About the Author

Zeynel Amaç graduated from Pamukkale University, Faculty of Education, Department of Primary Education in 1994 and began his teaching career the same year in Trabzon. Until 2000, he worked as a primary school teacher under the Ministry of National Education in Türkiye. He completed his master's degree in Curriculum and Instruction at University of Missouri-Columbia in 2002 and earned his Ph.D. in Curriculum and Instruction from Dicle University in 2023. He is currently a faculty member in the Department of Elementary Education at Kilisli Muallim Rıfat Faculty of Education, Kilis 7 Aralık University, Türkiye. His research interests include curriculum development, teacher education, refugee education, and inclusive education. He has published articles in various national and international journals and serves as a reviewer for several international journals and ECER conferences.

✉ zamac@kilis.edu.tr

ORCID [0000-0003-1387-5773](https://orcid.org/0000-0003-1387-5773)