JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12234 *Journal of Educational Measurement*

A General Framework for the Validation of Embedded Formative Assessment

Dorien Hopster-den Otter
University of Twente
Saskia Wools
Cito Lab
Theo J. H. M. Eggen and Bernard P. Veldkamp
University of Twente

In educational practice, test results are used for several purposes. However, validity research is especially focused on the validity of summative assessment. This article aimed to provide a general framework for validating formative assessment. The authors applied the argument-based approach to validation to the context of formative assessment. This resulted in a proposed interpretation and use argument consisting of a score interpretation and a score use. The former involves inferences linking specific task performance to an interpretation of a student's general performance. The latter involves inferences regarding decisions about actions and educational consequences. The validity argument should focus on critical claims regarding score interpretation and score use, since both are critical to the effectiveness of formative assessment. The proposed framework is illustrated by an operational example including a presentation of evidence that can be collected on the basis of the framework.

There has been increasing attention around formative assessment in education (e.g., Herman, 2013; Torrance & Pryor, 2001; Wiliam, 2011a). Formative assessment is intended to support student learning by providing evidence about this learning. This evidence needs to be used by teachers, students, or their peers for decisions and actions such as determining the next steps in learning and instruction or providing feedback to (peer)students (e.g., Falk, 2012; Schneider & Andrade, 2013).

Since poor quality formative assessment may lead to less effective and less efficient teaching and learning, good quality in formative assessment is necessary. Validity is one of the most important criteria for the evaluation of assessments (AERA, APA, & NCME, 2014) and is often defined as the extent to which an assessment result is appropriate for its intended interpretation and use (e.g., Kane, 2013). The process of purposefully collecting and evaluating evidence regarding the appropriateness of assessment results is called validation.

To validate the proposed interpretation and use of formative assessment, an explicit validation framework can be quite useful. A framework enhances the standardization of the validation process and supports validation practice (Wools, Eggen, & Sanders, 2010). However, a framework aimed at facilitating the validation of formative assessment remains wanting.

This article aims to provide such a framework. As there are many types of formative assessment, we focus on embedded formative assessment, the most formal type.

In the next section, we will explain the concept of (embedded) formative assessment and the characteristics that distinguish it from summative assessment. Subsequently, the concepts of validity and validation will be discussed, and the argument-based approach to validation will be introduced as a general validation framework. We will then present the proposed validation framework for formative assessment. To clarify the proposed framework, we will describe a formative assessment example, to which we will apply the framework. Finally, we will address some implications and recommendations.

Definition and Characteristics of Formative Assessment

Formative assessment is conceptualized in different ways and is used interchangeably with several other concepts in the literature such as assessment for learning, diagnostic assessment, and data-based decision making (Antoniou & James, 2014; Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). The lack of a clear definition makes it difficult to implement formative assessment and evaluate its effectiveness (Bennett, 2011). Therefore, numerous review studies have been conducted to get a better grasp of the concept (e.g., Bennett, 2011; Dunn & Mulvenon, 2009; Gulikers & Baartman, 2017; Heitink, Van der Kleij, Veldkamp, Schildkamp, & Kippers, 2016; Sluijsmans, Joosten-ten Brinke, & Van der Vleuten, 2013; Wiliam, 2011b).

In particular, some authors perceive formative assessment as an instrument that provides feedback (e.g., Dunn & Mulvenon, 2009; Kahl, 2005), while others emphasize the process of using this feedback (e.g., Clark, 2012; Popham, 2008). Bennett (2011) has perceived each position as an oversimplification. Even the most carefully designed instrument is unlikely to be effective if the process surrounding its use is flawed. Similarly, the process is unlikely to work if the instrumentation does not fit its intended purpose. This article follows Bennett's reasoning that formative assessment should be conceptualized as a thoughtful integration of both.

Formative assessment varies on a continuum from "on-the-fly" to "planned-for-interaction" to "curriculum-embedded" assessment (e.g., Forbes, Sabel, & Biggers, 2015; Furtak, 2006; Shavelson, 2003). On-the-fly assessment is the most informal. It does not involve a planned activity and occur as part of instructional activities. Planned-for-interaction assessment occurs, for example, when a teacher deliberately interrupts a lesson to ascertain students' understanding and alters instruction as necessary. Curriculum-embedded assessment is the most formal type. It consists of predefined tasks built into the school's educational program, that provide insights into students' current learning, and that is used to adapt teaching and learning to students' problem areas.

For the purpose of this article, we focus on this latter category of formative assessment, because it most closely relates to summative assessment for which several validation frameworks have already been developed. We define embedded formative assessment (hereafter referred to as formative assessment) as both an instrument and

a process, whereby evidence is purposefully gathered, judged, and used by teachers, students, or their peers for decisions about actions to support student learning. This definition excludes informal formative assessment in which evidence is elicited in an improvised and unscheduled manner (Ruiz-Primo & Furtak, 2007).

This conceptualization of formative assessment differs from that of summative assessment in several ways. Formative assessment is characterized by its purpose in supporting student learning, while summative assessment is intended to provide a final decision about students' learning, for example, for selection, certification, or accountability purposes (Shavelson, 2003; Trumbull & Lash, 2013). This difference has implications for the design and practice of formative assessment (Wiliam, 2011b). In order to make these implications clear, we will discuss the distinctive characteristics of formative assessment.

First, formative assessment is aligned directly with the teaching and learning process, because the evidence obtained is used for actions like adjusting instruction, changing learning strategies, or providing feedback (Harlen & James, 1997; Schneider & Andrade, 2013; Trumbull & Lash, 2013; Wiliam, 2011b). The uses may vary from teachers adjusting their instruction to students and peers changing their learning strategies. Nevertheless, as actions are necessary to support student learning, they make the actual process a distinctive feature of formative assessment (Bennett, 2011; Black & Wiliam, 2009).

Second, alignment with the teaching and learning process implies an assessment instrument that provides fine-grained information rather than a global reflection of students' capability (Goertz, Olah, & Riggan, 2009; Timperley, 2009). This means that a simple correct or incorrect score will usually not be sufficient. Student responses need to be scored in such a way that fine-grained information about the depth of student learning is elicited. The availability of instructionally tractable information built into the curriculum is fundamental for deciding where students are in their learning, where they need to go, and how best to get there (Broadfoot et al., 2002; Herman, 2013; Timperley, 2009; Wiliam, 2011b). Without this kind of information, it would be very difficult to use the assessment information for actions that support learning.

To conclude, formative assessment differs from summative assessment in terms of their explicit purpose in supporting learning. This purpose results in the need for alignment with the teaching and learning process, emphasizing its use by teachers and students and the need for fine-grained information from the assessment instrument. In the next section, the concepts of validity and validation will be discussed, and the argument-based approach to validation will be introduced as a general framework. This framework has been widely adopted in the validation of several summative assessments such as certification testing (Kane, 2004) and admission testing (Chapelle, Enright, & Jamieson, 2010). Furthermore, Nichols, Meyers, and Burling (2009) attempted to use the approach for formative assessment. They especially focused on the proposed use of assessment information, without making demands on the instrument or methodology from which the information was collected. However, we argue that there is a need for a well-designed instrument that fits the proposed use.

Argument-Based Approach to Validation

Since the early 1950s, Cronbach and Meehl's (1955) model of construct validity has been widely accepted and has been developed into a general framework for validation. The most general version of this model is based on three basic principles for validation: (1) the need for an explicit specification of the proposed interpretation; (2) the need for conceptual and empirical evaluation of the proposed interpretation; and (3) the need to consider alternate interpretations (Kane, 2013). These principles continue to be reflected in theories on validity and approaches to validation. For example, in Messick's (1989, p. 13) definition of validity: "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" [italics in original].

While construct validity as a unifying framework has been useful on a theoretical level, it has not been an effective unifying framework for validation in practice (Cronbach, 1989). For example, Messick's conceptualization of validity was translated into a validation practice with the aim of presenting as much validity evidence as possible. This resulted in an overly lengthy process that was difficult to implement. To make the validation process more pragmatic while still being faithful to basic scientific principles of construct validity, Kane (1992, 2004, 2006, 2013) proposed an argument-based approach to validation.

The argument-based approach consists of two stages: a developmental stage and an appraisal stage. In the developmental stage, an interpretation and use argument (IUA) is developed by specifying the proposed interpretation and use of assessment results. In the appraisal stage, the IUA is evaluated by critically examining its clarity, coherence, and plausibility.

The IUA consists of inferences regarding a score interpretation and a score use (Kane, 2013, 2016). A score interpretation involves claims about test takers or other units of analysis (e.g., teachers, schools). Claims about a score use involve decisions and possible consequences about these units of analysis. During the development of the IUA, the proposed interpretation and use are made explicit by incorporating their inherent inferences and assumptions.

Figure 1 shows an example of an IUA for a placement testing system (Kane, 2006). The first inference, named the scoring inference, is the evaluation of the observed performance leading to an observed score. Subsequently, the observed score is generalized to a universe score on a broader test domain. Within the next inference, the universe score is extrapolated toward a claim regarding the construct of interest in the practice domain. The last inference results in a decision on a student's skill level in relation to the construct of interest and placement in a specific course. These four inferences are likely to occur in most, if not all, IUAs for summative assessment (Kane, 2013).

Upon completion of the IUA, a critical evaluation of the inferences and assumptions is made in the appraisal stage, in which a validity argument can validate the proposed interpretation and use. The validity argument examines the coherence and completeness of the IUA and the plausibility of its inferences with respect to the purpose of the test (Crooks, 2004; Crooks, Kane, & Cohen, 1996; Dorans, 2012; Kane,

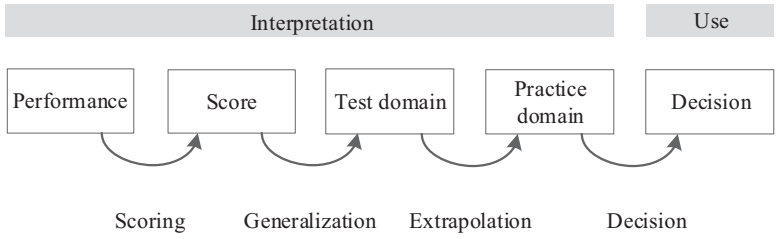

Figure 1. Example of an IUA.

2013). Although the proposed interpretation and use are evaluated together, a given validity argument is not necessarily adequate for both (Cizek, 2016; Sireci, 2016). A valid score interpretation is a prerequisite for a valid score use, but it does not automatically justify it. Similarly, the rejection of a score use does not necessarily invalidate a prior underlying score interpretation.

To sum up, the central idea of the argument-based approach is to build and evaluate an argument that helps test developers demonstrate that assessment scores are sufficiently useful for their intended purpose. To the extent that the assessment results are intended to be used for certain decisions that affect students or institutions, Kane (2013, 2016) emphasized the incorporation of inferences that are inherent in the proposed use, the evaluation of this proposed use, as well as the proposed interpretation. This also implies the inclusion of the consequences of these decisions in the validation process (Kane, 2016; Lane, 2014). If the proposed interpretation and use are supported by evidence and alternative explanations are rejected, it is appropriate to interpret and use assessment results in the proposed way (Kane, 2006). In the next section, the argument-based approach is extended to a validation framework for formative assessment.

The Proposed Validation Framework for Formative Assessment

The procedure of the argument-based approach would be similar for the validation of formative assessment as for the validation of summative assessment. Validation efforts would continue to be structured into a developmental stage to build the IUA, as well as an appraisal stage to critically evaluate the IUA on the basis of a validity argument (Kane, 2004, 2006, 2013). We will begin the current section by describing the proposed inferences in the IUA, after which we will address the validity argument.

IUA for Formative Assessment

The IUA for formative assessment consists of inferences regarding a score interpretation as well as inferences regarding a score use. Score-interpretation inferences cover claims about students' performance from the instrument, while score-use inferences involve decisions on this performance and possible consequences in the learning process.

With regard to the score-interpretation inferences, we propose a structure that is identical to the existing validation framework for summative assessment. This starts with (1) a *scoring inference*, whereby students' performance is converted into

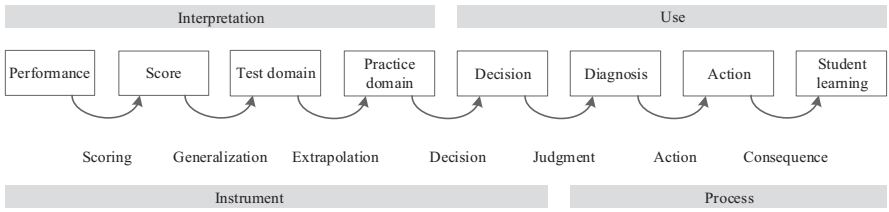

Figure 2. Proposed IUA for formative assessment.

interpretable information about their thinking. In addition, only a limited sample of all possible items is administered to students. This then leads to (2) a *generalization inference*, in which we draw upon the scoring of a limited sample to make inferences about the generalization of this score to all possible items in a so-called test domain. Furthermore, there is (3) an *extrapolation inference*, in which the interpretation of all possible items is extrapolated to a more general claim about students' performance in a so-called practice domain. The practice domain is defined as the domain about which we would like to make a decision.

With regard to the score-use inferences, we propose a different structure from the validation framework for summative assessment. The existing (4) *decision inference* links students' performance regarding the construct in the practice domain to a decision about their performance. In addition, we propose three additional inferences, since the actual use of the decision by teachers and students is an essential part of formative assessment (Bennett, 2011; Kane, 2016). We propose (5) a *judgment inference* because inaccurate understanding of the decision could lead to inappropriate actions (Gearhart et al., 2006; Maciver, Anderson, Costa, & Evers, 2014; C. M. Moss, Brookhart, & Long, 2013). The judgment inference links the decision to a diagnosis by the teacher or student. Moreover, as teachers and students are assumed to use this diagnosis for the selection of appropriate actions (Bennett, 2011; Black & Wiliam, 2009), we propose (6) an *action inference*, which links the diagnosis to an action. Finally, the implementation of these actions is expected to support student learning. We, therefore, propose (7) a *consequence inference*, which links the action to student learning. The proposed IUA for formative assessment is presented in Figure 2. We will describe the assumptions within the inferences of the proposed IUA in the remaining part of this section.

Assumptions within inferences.

Scoring inference (performance-score). It is proposed that students' performance on formative assessment tasks ought to be converted into interpretable information such as a score, rubric, qualitative description, or a score profile with subscores. For this inference, we assume that a set of scoring rules or algorithms provides insights into student learning strategies and mistakes. For example, multiple-choice item distractors are used to score common errors in a student's understanding (Goertz et al., 2009). In the case of manual scoring, we assume that raters are able to observe students' performance and describe their thinking.

Generalization inference (score-test domain). To allow generalization, the tasks needs to be a representative sample of the test domain in terms of content, difficulty, and the kind of answers that provide insights into students' learning strategies and mistakes. Therefore, we assume that the sample of tasks reflects the depth of student learning. Furthermore, we assume that the sample of tasks is sufficiently large to control sampling error (Kane, 2013). A sufficiently large sample is needed to support generalization because the more confident teachers and students are about students' level, the more effectively they can adjust instruction. To illustrate, an error could be a careless mistake, a persistent misconception, or a lack of understanding caused by inadequate knowledge (Bennett, 2011). Depending on the cause, the action will range from minimal feedback to reteaching and significant investment in eliminating misconceptions. With a representative and sufficiently large sample of items, teachers and students can select appropriate action.

Extrapolation inference (test domain–practice domain). For extrapolation, we assume that the tasks in the test domain reflect the particular learning objective, learning goal, or attainment goal in the practice domain. This means that the tasks include all aspects of the learning objective that are relevant for making a distinction between different student performances. None of the important aspects of the learning objective are overlooked (construct underrepresentation) and neither are other aspects confounded (construct-irrelevant variance). Furthermore, it is assumed that the tasks result in the students performing the expected thinking processes we are interested in.

Decision inference (practice domain–decision). The decision inference is drawn from a decision rule that specifies how the decision will be made. It is assumed that the cut-off score is in line with students' mastery of a learning objective. In addition, it is assumed that misclassifications with regard to misconceptions and learning strategies are minimized.

Judgment inference (decision–diagnosis). For the judgment inference, we assume that teachers and students are able to correctly understand the decision derived from the assessment instrument. This means that the presentation of the decision fits teachers' and students' level of assessment literacy (e.g., Popham, 2011). Furthermore, we assume that teachers and students are able to link the decision to students' individual circumstances such as the amount of effort invested, progress over time, and the particular context (Bennett, 2011). This suggests that formative assessment is student-referenced (Harlen & James, 1997), with the possibility of tailoring the actions to individual students' needs and motivating them. For example, a teacher or student can conclude that a nonmastery decision was based on a careless mistake, a persistent misconception, or a lack of understanding that has nothing to do with persistent misconception. It is also possible that the student actually mastered the learning objective but he or she was not focused or motivated, did not read the assignment correctly, or that the program might have crashed.

Action inference (diagnosis–action). To select appropriate actions, we assume that the assessment information is tied to the curriculum and fits teachers' and students' knowledge base including subject-matter knowledge and pedagogical content

knowledge (Falk, 2012; Forbes et al., 2015; Furtak & Heredia, 2014; Goertz et al., 2009; Heritage, Kim, Vendlinski, & Herman, 2009; Herman, Osmundson, Ayala, Schneider, & Timss, 2006; Sabel, Forbes, & Zangori, 2015). This would allow a teacher or student to select a new learning objective if they diagnose that the learning objective has been mastered. If they diagnose that the learning objective has not been mastered, then the student could decide on further practice, or the teacher could choose to provide minimal feedback, reteach the learning objective, or seek to eliminate the misconception.

Consequence inference (action–student learning). To allow the consequences, we assume that the approach to formative assessment results in student learning. However, the impact on learning also depends on the educational context (Bennett, 2011). Even if teachers and/or students act appropriately, the educational context could minimize the effect on students' learning (Bennett, 2011; Goertz et al., 2009). Therefore, this claim also assumes that the context is sufficiently supportive, including tools for data access, school leaders stimulating the use of formative assessment, teachers sharing the learning objectives, and students actively involved and motivated (Herman, 2013; C. M. Moss et al., 2013; Stobart, 2012; Torrance & Pryor, 2001).

Validity Argument for Formative Assessment

The validity argument for formative assessment would focus on both the score interpretation and the score use, because a failure in either part can reduce its effectiveness (Bennett, 2011). If the score interpretation is wrong, the basis of the actions is weakened. Similarly, if the score interpretation is correct and is presented in an understandable and meaningful way, but the action is inappropriate, learning is also less likely to occur. Within the IUA, the underlying inferences that seem to be questionable or critical should receive the most attention because they address the weakest links in the IUA (Kane, 2006; Wools, Eggen, & Béguin, 2016).

To the extent that the inferences are supported with evidence and alternative explanations are rejected, the validity argument is concluded by stating whether it is valid to interpret and use the assessment results. It is important to note that the analytical or empirical evidence will focus on making the claims plausible for a significant number of individuals rather than for individual cases (Kane, 2016).

Operational Example of the Validation Framework for Formative Assessment

To clarify the proposed validation framework for formative assessment, Bennett (2011) argues that we need one or more operational examples that show what formative assessment built on the basis of this theory looks like. This section contains such an example, to which we will apply the framework. We used the embedded formative assessment platform Groeimeter (GM), which was developed by the Cito Institute for Educational Measurement in the Netherlands. We will start with a description of the components of GM, followed by a description of how it is used. Then, we will apply the proposed validation framework to GM and will provide some examples as a means of validating it.

Description of GM

GM is aimed at supporting primary school teachers and guiding students in learning arithmetic. It consists of embedded formative assessment tasks, a teacher dashboard, and a student dashboard. The formative assessment tasks are related to the learning objectives of the Dutch arithmetic curriculum. Each predefined task is supposed to measure one learning objective. There are two types of assessment tasks, depending on what best fits the learning objective to be measured. The first type is a digital test in which students answer seven predefined items online. The number of items was chosen to make the tests practical. Digital tests are used for learning objectives that can be operationalized into automatically scored items, for example, "The student is able to calculate additions and subtractions up to 20." The items could be short answer, multiple choice, multiple response, hotspot, or matching items. For example, students fill in the right answer to the short-answer item: "How many balls do John and Mike have together?" or they need to select the coins that amount to 15. For the digital test, mastery is assigned to six correct items (Béguin & Straat, 2019). The second type is an assignment, for instance, having a group discussion or making a drawing. It is used when the learning objective is not suitable for automatic scoring because it requires more cognitively complex thinking. An example of such a learning objective is, "The student can think and reason critically about length and perimeter in meaningful problem situations." In the assignment, students were asked to come up with three different rectangles with a 16-meter perimeter and to explain their choices. In another assignment, they had to calculate the perimeter of a new fence for the parcels of land belonging to the farmer, James. For this assignment, mastery or nonmastery needed to be manually assigned after scoring the assignment.

GM contains a teacher dashboard that shows students' performance on completed assessments as a green or orange block, indicating mastery or nonmastery, respectively, of the measured learning objective. The program allows the teacher to manually change this status. Furthermore, the dashboard displays the students' icons, with information about their individual progress and item responses. Finally, it shows all the learning objectives of the Dutch arithmetic curriculum including an explanation and item example of the accompanying assessment. It is possible to assign a learning objective to an individual student or to the whole group of students.

GM also contains a student dashboard that shows the learning objectives assigned to the student. In this dashboard, the student can complete the assessment, and his or her performance is again shown as a green or orange block. It is possible to view the individual item responses on the digital test and compare them with the correct answers.

Use of GM

Teachers and students are supposed to view the students' mastery and individual item responses on the completed digital test and compare them with the correct answers. They can also analyze the students' answers on the assignment. In this way, teachers and students can judge the results themselves. Teachers can try to explain the results by linking them to the students' individual circumstances. When

Table 1
Inference, Assumptions, and Possible Sources of Evidence That Can Be Collected in the Validation of GM

Inferences	Assumptions	Sources of evidence
Scoring: from student performance to score profile	- Teachers are able to consistently mark performance on the assignments - The scoring rules provide insights into student learning strategies and mistakes	- Interrater reliability analysis of teachers' descriptions regarding the same student undertaking an assignment - Analyzing whether the distractors correspond to common learning strategies and mistakes
Generalization: from score to test domain	- Both types of tasks reflect the depth of student learning - Both types of tasks are sufficiently large to control sampling error	- Evaluation of test content matrices with regard to content and difficulty - Analysis of whether (a) different (number of) items provide similar inferences about students' thinking - Calculating a reliability coefficient
Extrapolation: from test domain to practice domain	- The tasks result in students performing the expected thinking processes - The tasks include all critical aspects of the learning objective	- Think-aloud protocols with students, which investigate whether they perform at the level of the expected thinking processes while completing the items - Study the relationship with other measures of the learning objective, for example, observations, standardized tests, etc.
Decision: from practice domain to decision	- The decision is in line with students' actual mastery of the learning objective.	- Comparing students' performance on a specific learning objective to other learning objectives of the same level of difficulty. - Comparing the decision on an external criterion such as oral exams or think-aloud studies. - Log-file analysis investigating how many times the decision has been overruled by the teacher

(Continued)

Table 1
Continued

Inferences	Assumptions	Sources of evidence
Judgment: from decision to diagnosis	- The assessment information supports teachers and students in correctly interpreting the decision in the teacher and student dashboards	- Think-aloud protocols that analyze how teachers and students interpret the decision - Set up an experiment where teachers are asked to interpret assessment information in different scenarios
Action: from diagnosis to action	- The measured learning objective is recognizably connected to teaching and learning	- Interviews that investigate whether teachers were able to correctly explain the meaning of the learning objectives - Analysis of the connection between the learning objectives in GM and the teaching methods used - Background documents of test developers that specify the relation between teaching and learning
	- The assessment information from GM supports teachers and students in selecting actions that enhance the teaching and learning process	- Classroom observation and/or log -file analysis that show what actions teachers and students perform - Interviews or questionnaires about how teachers and students experience the usability of GM
Consequence: from action to student learning.	- The performed actions have a positive impact on student learning	- Longitudinal study comparing schools that utilize GM and those that do not
	- There are no obvious obstacles within the educational context	- Evaluating the characteristics of schools in which GM works well

teachers determine that the automatically assigned status (mastery/nonmastery) does not reflect reality, they can overrule the status.

Assessment results are supposed to be used to guide follow-up action. For example, teachers are expected to provide additional instruction if they conclude that a learning objective has not been mastered due to a particular misconception. Students could undertake additional assignments to exercise a learning objective. It is assumed that the implementation of these actions supports student learning.

Designing a Validation Study for GM

The GM example illustrates the two distinctive characteristics of embedded formative assessment. First, it consists of an instrument that provides fine-grained information about students' performance vis-à-vis the learning objectives defined in the Dutch curriculum. Second, this information is supposed to be used for actions in the teaching and learning process.

This conceptualization requires an IUA that consists of inferences regarding both a score interpretation and a score use. Table 1 shows the inferences and its underlying assumptions. Furthermore, it provides examples of analytical and empirical evidence that can be collected to evaluate validity.

Since validation is a major activity, it is important to provide most attention to the most questionable or critical inferences. In our opinion, the most questionable and critical assumption of the score interpretation would be the need for fine-grained information. It should be made plausible that the assessment results provide enough insight into the depth of student thinking processes. In terms of the assumptions regarding score use, it should be made plausible that teachers and students are able to use the score interpretation to inform instructional actions that support learning.

Conclusion and Discussion

In this article, we proposed an extension of the argument-based approach (Kane, 2006, 2013) to the validation of embedded formative assessment. Embedded formative assessment was defined as both an instrument and a process, whereby evidence from a purposefully designed instrument is gathered, judged, and used for decisions about actions to support student learning. This conceptualization requires an IUA consisting of inferences regarding both a score interpretation and a score use. The score interpretation connects the specific task performance from the assessment instrument with an interpretation about the student's general performance. The score use connects that interpretation to decisions about actions in the teaching and learning process that are intended to support student learning. The validity argument should focus on critical claims regarding score interpretation as well as score use, since both are critical to the effectiveness of formative assessment.

In comparing this proposed framework in Figure 2 to the existing validation framework exemplified in Figure 1, the proposed structure of the inferences regarding the score interpretation is identical. However, the content of the score interpretation regarding formative assessment differs because the alignment with the teaching and learning process requires a different level of information granularity. This would result in different kind of tasks with different formulations regarding the scoring,

generalization, and extrapolation inferences. For example, the scoring inference often implies a way of scoring that provide insight into student learning strategies and mistakes, meaning that an aggregated score would usually not be sufficient. Furthermore, the generalization and extrapolation links may be less far-stretching than for summative assessment due to a narrowly defined practice domain (Crooks, 2004; Crooks et al., 1996; Dorans, 2012; Stobart, 2012). Therefore, generalization and extrapolation are less problematic and pose problems that are different from those of summative assessments, which often address broad constructs such as language literacy. For broad constructs, generalization and extrapolation could be so important that there is a need to add inferences (see, e.g., Kane, 2004; Wools et al., 2010). In addition to the score-interpretation inferences, we included three additional use inferences to make the use more visible (Bennett, 2011; Kane, 2016): a judgment inference, an action inference, and a consequence inference.

Adjustments in the IUA also changed the validity argument that evaluates the IUA; for different uses (e.g., formative vs. summative), different issues tend to become more salient. These differences demonstrate that an assessment instrument cannot be used interchangeably for both summative and formative purposes. The formative use of summative assessment and vice versa can only be applied after extensive and careful research.

Noteworthy, the GM system was used as an operational example to illustrate how the proposed framework suits the definition of curriculum-embedded formative assessment. It would be interesting to perform validation studies that provide analytical and empirical evidence with regard to the underlying assumptions.

In addition, the framework could be applied to other examples of curriculum-embedded assessment. This assumption might be investigated in a follow-up study, as an IUA needs to be developed and evaluated for each assessment in a particular context of practice (Kane, 2004). This could result in the specification of a somewhat different network of inferences and assumptions in another case-specific IUA, with the evaluation in the accompanying validity argument.

Furthermore, we developed a framework that suits the definition of curriculum-embedded assessment, which are the most formal category of formative assessment. However, a significant number of formative assessment is informal, such as a diagnostic conversation indicating a student's strengths and weaknesses. In a follow-up study, it would be interesting to investigate whether this framework could be applied to more informal formative assessment. To do this, we would need to further specify the differences between formal and informal formative assessment and identify the consequences for validation.

The general framework could be a meaningful contribution to guide the design and evaluation of formative assessment and to enhance our reasoning on validity. For example, it emphasizes the importance of actual use by teachers and students, placing substantial demands on teachers' content and pedagogical knowledge (Herman, Osmundson, Dai, Ringstaff, & Timss, 2015). To support the judgments and actions of the user, understandable score reports could be an important tool requiring careful design. This tool could meaningfully communicate the assessment scores and reduce the demands on users' knowledge and skills (Hattie & Brown, 2008; Matuk, Linn, & Eylon, 2015; Ryan, 2006; Zapata-Rivera & Katz, 2014).

Finally, this article opens up the discussion about the scope of validity theory, which is currently under intense debate (Newton & Shaw, 2016). The perspectives surrounding this debate range from those who insist that validity should remain a technical evaluation of measurement procedures (Borsboom, Mellenbergh, & van Heerden, 2004) to those who insist that it should become a broad concept to evaluate use of assessment results in the larger system (P. A. Moss, 1998). Although it seems possible to limit the scope of "validity" to a technical evaluation of summative assessment, this is impossible for formative assessment. The actual use and educational context of formative assessment are essential aspects of the effectiveness of these assessments. Shepard (2016) thus gets to the point in her remark that "Just as test design is framed by a particular context of use, so too must validation research focus on the adequacy of tests for specific purposes" (p. 273). Therefore, we felt the need to incorporate use inferences in the IUA for formative assessment, thus making the proposed use of tests an integral part of validation. The currently developed validation frameworks for summative assessment, however, do not include such use inferences. These differences could result in confusion around the concept of validity, which is not desirable. Therefore, the necessary incorporation of use inferences for formative assessment leaves the question of whether the concept of validity should be expanded to an overall evaluation of the score interpretation as well as of the score use. Referring to all of this as validation would make it possible to strive for a uniform conceptual framework within validity theory for both summative and formative assessment.

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Antoniou, P., & James, M. (2014). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation and Accountability*, *26*, 153–176. https//doi.org/10.1007/s11092-013-9188-4

Béguin, A. A., & Straat, J. H. (2019). On the number of items in learning goal mastery testing. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 121–134). Cham, Switzerland: Springer International Publishing.

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

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

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. https//doi.org/10.1037/0033-295X.111.4.1061

Broadfoot, P. M., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13. https//doi.org/10.1111/j.1745-3992.2009.00165.x

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different methods. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 212–225. https//doi.org/10.1080/0969594X.2015.1063479

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(2), 205–249. https//doi.org/10.1007/s10648-011-9191-6

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Crooks, T. J. (2004, March). *Tensions between assessment for learning and assessment for qualifications*. Paper presented at the third conference of the Association of Commonwealth Examinations and Accreditation Bodies (ACEAB), Nadi, Fiji.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, *3*(3), 265–286. https//doi.org/10.1080/0969594960030302

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice*, *31*(4), 20–37. https//doi.org/10.1111/j.1745-3992.2012.00250.x

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation*, *14*(7), 1–11. http://pareonline.net/getvn.asp?v=14&n=7

Falk, A. (2012). Teachers learning from professional development in elementary science: Reciprocal relations between formative assessment and pedagogical content knowledge. *Science Education*, *96*(2), 82–99. https//doi.org/10.1002/sce.20473

Forbes, C. T., Sabel, J. L., & Biggers, M. (2015). Elementary teachers' use of formative assessment to support students' learning about interactions between the hydrosphere and geosphere. *Journal of Geoscience Education*, *63*, 210–221. https//doi.org/10.5408/14-063.1

Furtak, E. M. (2006). *Formative assessment in K-8 science education: A conceptual review*. Commissioned paper for the Committee on Science Learning, Kindergarten through Eighth Grade, National Research Council, Ontario, Canada.

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

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendliski, T., … Bernbaum, D. J. (2006). Developing expertise with classroom assessment in K-12 science: Learning to interpret student work. Interim findings from a 2-year study. *Educational Assessment*, *11*(3–4), 237–263. https//doi.org/10.1080/10627197.2006.9652990

Goertz, M. E., Olah, L. N., & Riggan, M. (2009). Can interim assessments be used for instructional change? Policy brief. RB-51. *CPRE Policy Briefs*. http://repository.upenn.edu/cpre_policybriefs/39

Gulikers, J., & Baartman, L. (2017). *Doelgericht professionaliseren: Formatieve toetspraktijken met effect! Wat DOET de docent in de klas*? [Targeted professionalization: Formative assessment practices with effect! What DOES the teacher do in the classroom?]. PPO-NRO 405-15-722. Den Haag, The Netherlands: NRO

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy and Practice*, *4*(3), 365–379. https//doi.org/10.1080/0969594970040304

Hattie, J., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems*, *36*(2), 189–201. https//doi.org/10.2190/ET.36.2.g

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review*, *17*, 50–62. https//doi.org/10.1016/j.edurev.2015.12.002

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

Herman, J. L. (2013). *Formative assessment for next generation science standards: A proposed model* (CRESST Resource Paper No. 16). Los Angeles, CA: CRESST.

Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timss, M. (2006). *The nature and impact of teachers' formative assessment practices* (CSE Technical Report 703). Los Angeles, CA: CRESST.

Herman, J. L., Osmundson, E., Dai, Y., Ringstaff, C., & Timss, M. (2015). Investigating the dynamics of formative assessment: Relationships between teacher knowledge, assessment practice and learning. *Assessment in Education: Principles, Policy and Practice*, *22*(3), 344–367. https//doi.org/10.1080/0969594X.2015.1006521

Kahl, S. (2005). Where in the world are formative tests? Right under your nose! *Education Week*, *25*(4), 38.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535. https//doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 135–170. https//doi.org/10.1207/s15366359mea0203_1

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. https//doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 198–211. https//doi.org/10.1080/0969594X.2015.1060192

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*(1), 127–135. https//doi.org/10.7334/psicothema2013.258

Maciver, R., Anderson, N., Costa, A. C., & Evers, A. (2014). Validity of interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment*, *22*(2), 149–164. https//doi.org/10.1111/ijsa.12065

Matuk, C. F., Linn, M. C., & Eylon, B. (2015). Technology to support teachers using evidence from student work to customize technology-enhanced inquiry units. *Instructional Science*, *43*, 229–257. https//doi.org/10.1007/s11251-014-9338-1

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: Macmillan.

Moss, C. M., Brookhart, S. M., & Long, B. A. (2013). Administrators' roles in helping teachers use formative assessment information. *Applied Measurement in Education*, *26*(3), 205–218. https//doi.org/10.1080/08957347.2013.793186

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, *17*(2), 6–12.

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word "validity" and options for reaching consensus. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 178–197. https//doi.org/10.1080/0969594X.2015.1037241

Nichols, P. D., Meyers, J. L., & Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice*, *28*(3), 14–23. https//doi.org/10.1111/j.1745-3992.2009.00150.x

Popham, W. J. (2008). *Transformative assessment in action*. Alexandria, VA: ASCD.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265–273. https//doi.org/10.1080/08878730.2011.605048

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

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Mahwah, NJ: Lawrence Erlbaum.

Sabel, J. L., Forbes, C. T., & Zangori, L. (2015). Promoting prospective elementary teachers' learning to use formative assessment for life science instruction. *Journal of Science Teacher Education*, *26*(4), 419–445. https//doi.org/10.1007/s10972-015-9431-6

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

Shavelson, R. (2003). *On the integration of formative assessment in teaching and learning with implications for teacher education*. Stanford, CA: Stanford Education Assessment Laboratory and University of Hawaii Curriculum Research and Development Group.

Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 268–280. https//doi.org/10.1080/0969594X.2016.1141168

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy and Practice*, *23*(2), 226–235. https//doi.org/10.1080/0969594X.2015.1072084

Sluijsmans, D., Joosten-ten Brinke, D., & Van der Vleuten, C. (2013). Toetsen met leerwaarde: Een reviewstudie naar de effectieve kenmerken van formatief toetsen [Assessment with learning value: A review study into the characteristics of effective formative assessment]. NWO-PROO 411-11-697. Den Haag, The Netherlands: NWO.

Stobart, G. (2012). Validity in formative assessment. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 233–242). London, UK: Sage.

Timperley, H. (2009, August). *Using assessment data for improving teaching practice*. Paper presented at the ACER research conference on assessment and student learning, Perth, Western Australia.

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal*, *27*, 615–631. https//doi.org/10.1080/01411920120095780

Trumbull, E., & Lash, A. (2013). *Understanding formative assessment: Insights from learning theory and measurement theory*. San Francisco, CA: WestEd.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. H. M. (2015). Integrating data-based decision making, assessment for learning, and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy and Practice*, *22*(3), 324–343. https//doi.org/10.1080/0969594X.2014.999024

Wiliam, D. (2011a). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.

Wiliam, D. (2011b). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https//doi.org/10.1016/j.stueduc.2011.03.001

Wools, S., Eggen, T. J. H. M., & Béguin, A. A. (2016). Constructing validity arguments for test combinations. *Studies in Educational Evaluation*, *48*, 10–18. https//doi.org/10.1016/j.stueduc.2015.11.001

Wools, S., Eggen, T. J. H. M., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO*, *8*, 63–82.

Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy and Practice*, *21*(4), 442–463. https//doi.org/10.1080/0969594X.2014.936357

Authors

DORIEN HOPSTER-DEN OTTER is a PhD candidate at RCEC, Cito, University of Twente, Postbus 217, 7500 AE Enschede, The Netherlands; d.denotter@utwente.nl. Her primary research interests include formative assessment, validation, and score report development.

SASKIA WOOLS is a manager in CitoLab at Cito, Amsterdamseweg 13, 6814 CM, Arnhem, The Netherlands; saskia.wools@cito.nl. Her primary research interests include the validity, validation and innovations of educational assessments.

THEO J. H. M. EGGEN is a senior research scientist at the Psychometric Research Center of Cito and at the University of Twente, Amsterdamseweg 13, 6814 CM, Arnhem, The Netherlands; theo.eggen@cito.nl. His primary research interests include the quality of educational testing and computerized (adaptive) testing.

BERNARD P. VELDKAMP is the head of the Department of Research Methodology, Measurement and Data Analysis at the University of Twente, Postbus 217, 7500 AE Enschede, The Netherlands; b.p.veldkamp@utwente.nl. His primary research interests include measurement optimization and behavioral data science.