frontiers in Education

CONCEPTUAL ANALYSIS
published: 26 March 2021
doi: 10.3389/feduc.2021.605999

An Argument-Based Framework for Validating Formative Assessment in the Classroom

*Peter Yongqi Gu**

School of Linguistics and Applied Language Studies, Victoria University of Wellington, Wellington, New Zealand

The embedded and contingent nature of classroom-based formative assessment means that validity in the norm-referenced, summative tradition cannot be understood in exactly the same way for formative assessment. In fact, some scholars (e.g., Gipps, Beyond testing: towards a theory of educational assessment, 1994, Falmer Press, London, UK) have even contended for an entirely different paradigm with an independent set of criteria for its evaluation. Many others have conceptualized the validity of formative assessment in different ways (e.g., Nichols et al., 2009, 28 (3), 14–23; Stobart, Validity in formative assessment, 2012, SAGE Publications Ltd, London, UK; Pellegrino et al., Educ. Psychol., 2016, 51 (1), 59–81). This article outlines a framework for evaluating the argument-based validity of CBFA. In particular, I use Kane (J. Educ. Meas., 2013, 50 (1), 1–73) as a starting point to map out the types of inferences made in CBFA (interpretation and use argument) and the structure of arguments for the validity of the inferences (validity argument). It is posited that a coherent and practical framework, together with its suggested list of inferences, warrants and backings, will help researchers evaluate the usefulness of CBFA. Teachers may find the framework useful in validating their own CBFA as well.

Keywords: formative assessment, classroom-based formative assessment, validity, validation of formative assessment, argument-based validation

OPEN ACCESS

Edited by:
Chris Davison,
University of New South Wales, Australia

Reviewed by:
Susan M Brookhart,
Duquesne University, United States
Ricky Lam,
Hong Kong Baptist University, Hong Kong

***Correspondence:**
Peter Yongqi Gu
peter.gu@vuw.ac.nz

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *14 September 2020*
Accepted: *19 February 2021*
Published: *26 March 2021*

Citation:
Gu PY (2021) An Argument-Based Framework for Validating Formative Assessment in the Classroom. Front. Educ. 6:605999. doi: 10.3389/feduc.2021.605999

INTRODUCTION

Since Black and Wiliam's (1998) review article, formative assessment has gained increasing currency in educational systems as different as Australia (Klenowski, 2011), China (Xu and Harfitt, 2019), New Zealand (Bell and Cowie, 2001), Norway (Hopfenbeck et al., 2015), the United Kingdom (Torrance and Pryor, 1998) and the United States (Ruiz-Primo and Furtak, 2007). Part of the surge of interest comes from its intuitive appeal; part of it comes from claims of its effectiveness in "doubling the speed of student learning" (Wiliam, 2007, 36–37).

Recent years have seen repeated challenges to the effectiveness promise of formative assessment. Dunn and Mulvenon (2009) focused on the lack of consensus on definition. They rightly pointed out that "without a clear understanding of what is being studied, empirical evidence supporting formative evidence will more than likely remain in short supply" (p. 2). Bennet (2011) noted that most of the original claims of effectiveness in Black and Wiliam's (1998) review were exaggerated or misplaced. Kingston and Nash (Kingston and Nash, 2011) did a new meta-analysis of more than 300 studies on the efficacy of formative assessment. They found only 13 studies (42 independent effect sizes) that reported enough information to calculate effect sizes. The average effect size was only 0.20, with formative assessment being more effective in English language arts (effect size = 0.32) than in mathematics (effect size = 0.17) or science (effect size = 0.09). To use Bennet's (2011) words,

the "mischaracterisation" of Black and Wiliam's (1998) conclusions "has essentially become the educational equivalent of urban legend" (p. 12).

It should be noted that none of the challenges denies the potential efficacy of formative assessment. They serve to emphasize a point that formative assessment is not a simplistic issue and that it is not necessarily effective in improving student learning. In addition to different definitions of formative assessment, other factors that influence the effectiveness of formative assessment includes, among others, its domain dependency, teachers' assessment literacy, and support or constraints in the larger educational context.

Most importantly, validity is a necessary but insufficient condition for effectiveness. Even a valid formative assessment task may not lead to intended learning success; invalid formative assessment practices will definitely not be effective. If we follow Kane and Wools (2019) and view validity from both a measurement perspective and a functional perspective, we can reword the previous statement this way: proper assessment procedures and the interpretation and use of assessment results may or may not lead to the functional effect of usefulness. In fact, some forms of formative assessment are more effective than others; and some formative assessment practices may not lead to learning at all. In other words, validating formative assessment is an important step towards ensuring its usefulness.

This article looks at the validity issue of formative assessment, and illustrates how the argument-based framework for test validation (Kane, 2013) can be applied to the validation of formative assessment in the classroom. I will first present an operationalization of classroom-based formative assessment (CBFA), followed by a brief introduction of validity and validation issues in educational measurement in general. Finally, argument-based validation of classroom-based formative assessment will be outlined. I will illustrate how this can be done with a concrete example.

CLASSROOM-BASED FORMATIVE ASSESSMENT

Before we talk about the validity (interpretation and use) of CBFA and its effectiveness, we need to delineate its conceptual boundaries, so that we know exactly what is implemented, summarized as findings, and potentially transferred across contexts (Bennett, 2011). In this section, I will start by operationalizing the construct of formative assessment, and proceed to narrow down the construct into its classroom-based variant. I will also highlight two seminal features as part of this operationalization, i.e., cycle length and a continuum of formality of assessment events, and attempt to locate CBFA as predominantly short-cycle, contingent assessment events that happen in the classroom.

Defining and Operationalizing Formative Assessment

Formative assessment has been understood as instrument, process, and function. The first perspective is in the minority and is represented mostly by test publishers (Pearson Education, 2005). Formative assessment in this sense is reflected in the diagnostic tests they produce. An overwhelming amount of definitions do not view formative assessment as an instrument. Many scholars define formative assessment as a process by which student understanding is elicited and used to adjust teaching and learning (Popham, 2008). Most other definitions see formative assessment as a process aimed at a formative function (Bennett, 2011).

Assessment is formative when evidence of learning is elicited and matched against the learning target to inform the teacher and the learner about the gap between the learner's current state of knowledge or ability and the target. To be helpful at all in closing the gap, a formative assessment event needs to be rounded off with follow-up action (Sadler, 2010). Davison and Leung (2009) outline two basic functions of formative assessment, informing and forming. The former puts emphasis on the necessary but insufficient nature of feedback; while the latter underscores the importance of students' engagement with the feedback they receive in order for learning to take place.

Similarly, Andrade (2010) simply conceptualises formative assessment as "informed action" (p. 345). Expressed in another way, most researchers (Ramaprasad, 1983; Sadler, 1989; Black and Wiliam, 2012) believe that the essence of formative assessment involves establishing 1) where the learners are going; 2) where the learners currently are in their learning; and 3) what needs to be done to get them there.

Formative assessment is hard to operationalize, partly because we normally talk about it being a formative function of assessment rather than a type of assessment with a palpable format. Elsewhere, I have tried to operationalize formative assessment into formative functions and formative practices (Gu, 2020). The former includes a formative purpose before assessment and a formative effect being achieved at the end. The latter includes four crucial consecutive steps: eliciting evidence of learning or understanding, interpreting the evidence, providing feedback, and student/teacher action engaging with the feedback. Each of the four steps is oriented towards achieving a concrete target of learning (**Figure 1**). Ideally, a formative assessment event should include a formative purpose, a formative practice cycle (which I call a formative event), and achieve a formative effect. In most cases, however, we cannot realistically expect to achieve any learning effect with one round of formative practice. Very often we do not have an explicit and conscious formative purpose before we start a round of formative practice inside the classroom. I therefore see one complete round of formative practice involving all four steps moving towards achieving the target of learning as the minimum requirements for the defining features of a formative assessment event. This operationalization allows teachers to catch formative assessment as it appears, as it were, and gives researchers concrete units for analysis (Gu and Yu, 2020).

Classrooms as a major site for learning is a major site for formative assessment as well. However, not all assessment that happens in the classroom is formative. Formative assessment that happens in the classroom can be planned or contingent; and, depending on the task being assessed, classroom-based formative assessment can be completed within short, medium, and long cycles.

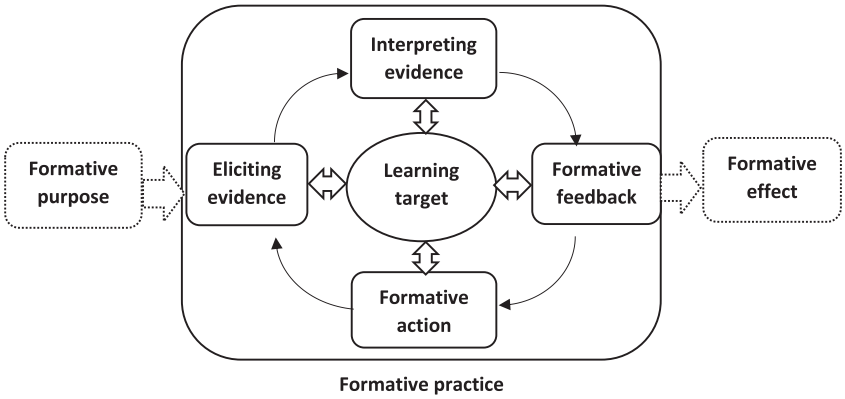

FIGURE 1 | Operationalizing formative assessment.

Delimiting Classroom-Based Formative Assessment

Teachers use a wide range of tools to collect information about student learning in class. Sometimes it can be a formal test; other times it may just be an informal question or an observation of a regular learning task. However, not all classroom tasks are assessment, and classroom-based assessment is not necessarily formative (Black and Wiliam, 2005). Furthermore, formative assessment does not necessarily happen in the classroom. Continuous assessment such as a class quiz, for example, definitely takes the form of an assessment, and it can be done in regular intervals during a period of teaching. Unless the information elicited through the quiz is interpreted and the result relayed back to the students, and unless the students act on the feedback from the quiz result, nothing becomes formative. Likewise, not all alternative forms of assessment such as peer grading can achieve formative functions (Davison and Leung, 2009). Many times, classroom tasks elicit information about student learning that the teacher and the students may not become aware of. Even if this information is noticed by the teacher, and feedback is provided, if the student concerned does not take any action in response to the feedback, the feedback will be wasted.

Classroom-based formative assessment is therefore a teaching/learning event that serves a formative assessment function and which happens within or beyond one class. One complete CBFA event includes 1) elicitation of evidence of students' understanding or learning, 2) interpretation of the elicited information against the learning target or success criteria, 3) feedback based on this interpretation for the student in question, and 4) follow-up action taken by the student and/or teacher to improve learning. All these elements must be present before each CBFA event is complete. And more often than not, learning only takes place after the completion of a series of these cyclical, and spiralling CBFA events.

Cycles of Formative Assessment Events

Classroom assessment practices that involve elicitation of evidence, interpreting the evidence, providing feedback, and student/teacher take-up and action form one complete CBFA event (**Figure 1**). Each event is aimed at a target of learning, teaching, and assessment; and each step or element has the learning target as the reference point. These elements are both sequential and interactive. The completion of one cycle normally will necessitate a readjustment of the target which entails another cycle of assessment practice. The elements, therefore, form spiralling cycles, with each complete cycle moving student understanding or learning closer to the target. This happens continuously until a judgment is made that the target is reached and the success criteria met.

Depending on the scope of the task being assessed, a complete cycle of an assessment event mentioned above can take a few seconds; or it may take a week or much longer to complete. Wiliam (2010) groups the lengths of these cycles into three types: short-, medium-, and long-cycles (**Table 1**).

(Wiliam 2010, 30)

As **Table 1** suggests, CBFA normally belongs to the 'short-cycle' category. This is especially true for those assessments that happen within the classroom. That said, learning usually takes place in timespans longer than a normal class. It is, therefore, often the case that teachers and learners need to check again and again in order to see the effect of learning and see if a course of action works. These actions would take longer than one class and can also be regarded as CBFA. Formative assessment events that go beyond a month or so to complete are normally more formal. For example, information from a formal diagnostic test can be used to guide learning efforts for a whole semester or more. These normally happen well beyond regular classes, and, despite being formative in nature, cannot be counted as CBFA anymore, simply because most of the assessment practices do not happen inside the classroom.

Planned and Contingent Assessment Practices

When formative assessment practices are examined inside the classroom, Cowie and Bell (1999) found largely two types, planned and interactive assessment practices. For planned

TABLE 1 | Short-, medium-, and long-cycle lengths for formative assessment.

Type	Focus	Length
Long-cycle	Across marking periods, quarters, semesters, years	4 weeks to 1 year
Medium-cycle	Within and between instructional units	1–4 weeks
Short-cycle	Within and between lessons	Day by day: 24–48 h minute by minute: 5 s to 2 h

formative assessment, the teacher has a clear but usually general purpose and target before class, s/he deliberately chooses assessment tools to collect information about students' understanding of or performance on the target task, interpret the result on the spot or after class, provides feedback and act on it. A questionnaire before teaching starts would help the teacher gauge the students' current level and expectations, which in turn will help the teacher prepare for more targeted teaching. Likewise, weekly quizzes and many curriculum-embedded tests that are pre-designed for a unit of teaching help the teacher monitor the learning progress of the class and adjust teaching accordingly.

Inside the classroom, many assessment opportunities arise spontaneously without the teacher's preparation. These normally take the form of classroom interactions or the teacher's observations of the students' task performances. Cowie and Bell (1999) labelled these assessment events 'interactive'. Interactive formative assessment events are usually triggered by the teacher noticing an unexpected or erroneous understanding or performance. On the spot interpretation of the deviant understanding would help the teacher recognize the error as a significant point to focus on. The teacher may immediately ask another student the same question and see if the problem is pervasive (both a follow-up action of the previous assessment event and the start of another assessment event), and if the gravity of the problem is deemed serious, the teacher may decide to explain, re-teach, or change a practice activity for the whole class.

The same phenomenon has been observed by Ruiz-Primo and her colleagues who labelled it 'informal formative assessment' (Ruiz-Primo and Furtak, 2006; Ruiz-primo and Furtak, 2007; Ruiz-Primo, 2011). These researchers developed this into an observation framework that included eliciting (E), student response (S), recognizing (R), and using information (U) and called it the 'ESRU cycle'. Interestingly, their studies indicated that informal teacher classroom assessment practices include different configurations in terms of how many elements are practiced. Few complete cycles of informal formative assessment were found. Instead, teachers used ES more often than ESR and ESRU. Those who used more complete ESRU cycles were found to benefit their students better.

Meanwhile, many researchers realize that it is often hard to categorize CBFA events into dichotomies such as planned/unplanned or formal/informal. The dichotomies are in fact two ends of a continuum. Shavelson et al. (2008) outline three anchor points on a continuum: (a) "on-the-fly," (b) planned-for interaction, and (c) formal and embedded in curriculum. Similarly, Bailey and Heritage (2008) also referred to a 'degree of spontaneity' (p. 48) and used 'on the run/in the moment', 'planned for interaction', and 'embedded in curriculum' assessment to describe the continuum. Likewise, Davison (2008) talked about 'a typology of possibilities' which also aligned four types of classroom assessment possibilities along a continuum, ranging from 'in-class contingent formative assessment-while-teaching', 'more planned integrated formative assessment', and 'more formal mock or trial assessments modelled on summative assessments but used for formative purposes', to 'prescribed summative assessments, but results also used formatively to guide future teaching/learning'.

An overwhelming proportion of assessment activities happening in classrooms are contingent, and the cycles are short and often incomplete. The formal, semi-formal, and often curriculum-embedded assessment activities in or out of everyday classes can be used for formative purposes as well.

By nature, formative assessment is meant to support learning. This, however, does not imply that any formative assessment practice will necessarily improve learning. Inside the classroom, many factors influence the validity and the effectiveness of the assessment practice. For example, even if a complete formative assessment event is present, the task being assessed can be irrelevant to the curriculum target being taught and learned. One or even more observations of similar tasks performed by a few students may not be enough to lead to a generalizable conclusion. On the spot interpretations of the evidence of learning may or may not be appropriate. Premature claims can be made about student achievement or ability based on the interpretations. Feedback provided and instructional decisions thereafter can be misguided if the interpretation of learning evidence is inaccurate. In other words, the lack of evidence we discussed previously for the effectiveness of formative assessment can well be due to a lack of validity in the formative assessment that has been studied.

VALIDITY AND VALIDATION

In educational measurement, validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests"; while validation is seen "as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, 11). In this sense, validity of formative assessment is the plausibility of the interpretations and the appropriateness of the feedback and uses based on the evidence of learning elicited. Validation of formative assessment is the process in which interpretations and uses of formative assessment results are specified, justified and supported.

A number of scholars have tried to examine the validity issue of formative assessment. Gipps (1994) contends that assessment for teaching and learning purposes deserves a completely new paradigm for its evaluation. Instead of terminologies such as validity and reliability that belong to the psychometric tradition, new terms such as Curriculum fidelity, Comparability, Dependability, Public credibility, Context description, and Equity represent a set of criteria better suited to formative assessment. Many other scholars (e.g., Stobart, 2012; Kane and Wools, 2019) seem to have come to the conclusion that a validity framework is appropriate for formative assessment, although the emphases in different facets of this framework and the kinds of interpretations and uses of assessment results are very different from psychometric tests (Pellegrino et al., 2016).

Kane and Wools (2019) distinguished between two perspectives on the validity of assessments: a measurement versus a functional perspective. The former focuses on the accuracy of construct scoring, and the latter focuses on the extent to which the assessment serves its targeted purposes. Kane and Wools (2019) argued that, for classroom assessment, "the functional perspective is of central concern, and the measurement perspective plays a supporting role" (p. 11).

A number of scholars (e.g., Stobart, 2012) take a similar position and have placed their emphasis of validity on the effect or the consequential facet of formative assessment, arguing that a major claim is to lead to the improvement of learning. While I do agree that ideally each formative assessment practice leads to targeted learning results, and that this should be the ultimate criterion to evaluate the validity of formative assessment, I do not see it as practical to expect every formative assessment event to result in desired learning consequences. Very simple and concrete learning tasks such as the correct pronunciation of a word may be achievable at the end of a short cycle of formative assessment practice. Most learning tasks, however, will need a much more complex process of teaching, learning and assessment to be completed.

I contend that the "measurement perspective" is equally important for formative assessment, but the emphasis of formative assessment in such a perspective would be very different from traditional tests. Just like the fundamental importance of the psychometric properties of a test in producing the scores for valid interpretations and uses, the basic properties of a formative assessment event (i.e., eliciting evidence of learning, interpreting the results, providing feedback, and acting on feedback) must be carried out appropriately. I would call this an "assessment perspective", and posit that the accuracy and trustworthiness of the information obtained from formative assessment, the correct interpretations and appropriate uses of assessment results determine to a large extent the usefulness of the formative assessment practice.

Most importantly, accurate interpretations and appropriate uses of assessment results very much depend on the assessor's pedagogical content knowledge (Shulman, 1986) which includes, among other things, the learning and assessment target and the success criteria in reaching the target. This domain-specific understanding of the learning target is a crucial facet of classroom formative assessment that makes or breaks any formative assessment practice (Bennett, 2011). Setting the right assessment goal, choosing appropriate tools to elicit the evidence of learning, interpreting the evidence appropriately, providing the right feedback, and embarking on an informed course of action, every stage of an assessment event can go wrong, if the assessor's understanding of the learning target is inappropriate or faulty. For example, in the formative assessment of language learning in class, the teachers' knowledge of curriculum standards, their beliefs in language competence and language learning, and their understanding of the success criteria in performing the language tasks used to elicit evidence of student learning, are as important as, if not more important than the assessment procedures as such.

VALIDATING CLASSROOM-BASED FORMATIVE ASSESSMENT

The Argument-Based Validation Framework

Over the last 2 decades or so, a validation framework that allows all evidences to be presented as a coherent whole (as opposed to a list of fragmented evidences) is getting increasingly accepted by the educational assessment community. The framework is called "argument-based validity". The idea is: in claiming that our assessment is good for its purposes, we are making an argument. Validation is therefore a matter of making this argument convincing enough for people who care about our assessment.

As early as the 1980s, Cronbach (1988) began to see test validation as gathering evidence to support an argument for our design, interpretation, and use of a test. Over the years, Kane (1992), Kane (2001), Kane (2006) and Mislevy et al. (2003) have developed the argument-based approach to test validation into a coherent and practical framework. In language assessment, Bachman (2005) and Bachman and Palmer (2010) have taken up the approach; and one of the major English language tests, TOEFL, has been validated using the argument-based approach (Chapelle et al., 2008). The latest addition is Chapelle's (2020) book-length volume on argument-based validation of language tests.

In an argument-based framework, validation is done in two steps, or to put it another way, we need two sequential arguments to validate an assessment: an interpretation and use argument (IUA) and a validity argument (Kane, 2013). In step 1, we articulate an IUA through a logical analysis of the chain of inferences linking test performance to a judgement or decision, and the assumptions on which they rest. In other words, we outline explicitly the major inferences and claims we are making based on assessment outcomes. In step 2 (validity argument), we provide an overall evaluation of the inferences in the IUA and systematically argue that each claim or inference is true unless proven otherwise. The validity argument uses Toulmin's (2003) argument structure. **Figure 2** shows a simple claim using the Toulmin structure. Since the rebuttal does not overturn the conclusion, the claim stands.

Hopster-den Otter et al. (2019) proposed an argument-based framework to validate formative assessment. They conceptualised

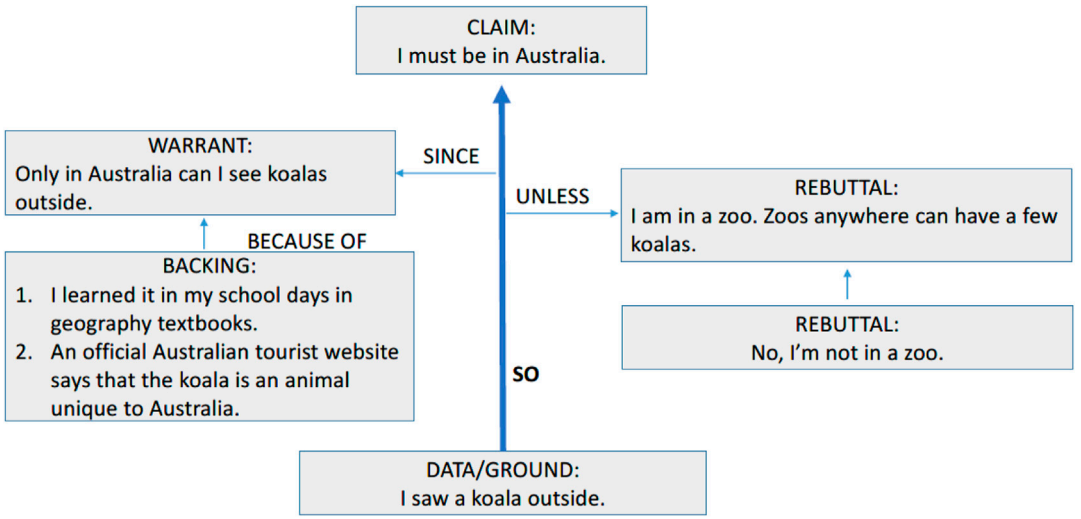

FIGURE 2 | Argument structure for a simple claim/conclusion.

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" (p. 3). This conceptualisation was confined to the curriculum-embedded, pre-defined types of formal assessment tasks (instruments) that resembled summative tests in format, and excluded the majority of classroom-based formative assessments which occur contingently and unplanned. This explains why their "interpretation inferences" in the IUA being identical to those in tests, which is in line with their previous thinking on "formative use of test results" (Hopster-den Otter et al., 2017).

A major contribution of Hopster-den Otter et al. (2019) lies in their conceptualisation of the Use component of the IUA, focusing on the utilisation of test results for instructional purposes. They parsed the use component of IUA into four inferences: Decision, Judgment, Action, and Consequence. These inferences at the end of a diagnostic test make the use of the instrument formative. In their illustrative example, Hopster-den Otter et al. (2019) referred to the validation of an online test of arithmetic which provided subsequent feedback for primary school teachers and learners.

Seeing formative assessment as formative use of tests necessitates the judgment and use of assessment information after a test. However, conceptualising CBFA as both a process and a function but not an instrument (**Figure 1** above) means that most of the judgment, interpretation, and action after feedback are done during the classroom assessment process. As a result, the validation process in CBFA does not have to start after assessment is done; and the Use component of IUA does not need to be parsed the way Hopster-den Otter et al. (2019) did. In other words, the framework presented next is an alternative to Hopster-den Otter et al. (2019) that complements their framework. While the Hopster-den Otter et al. framework is more appropriate for formative use of tests, the framework in this article is more appropriate for CBFA.

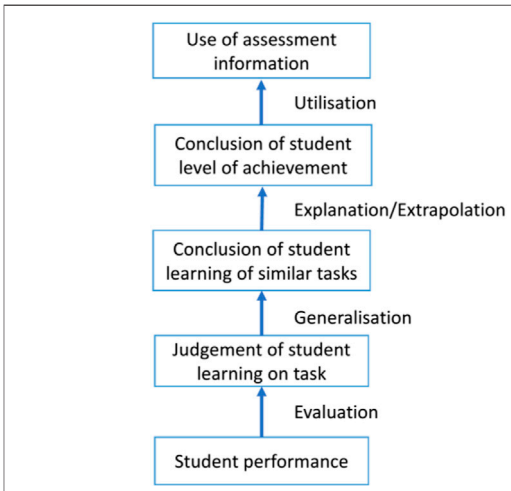

FIGURE 3 | Chain of inferences in CBFA.

Argument-Based Validation of CBFA

Step 1: Interpretation and use argument

The following figure (**Figure 3**) outlines the chain of inferences in CBFA. When we make a judgment of a student's ability in performing a task in class, we are making an evaluation inference. When we conclude that the student is able to do similar tasks across similar situations, we are making a generalization inference. After a number of observations of successful performance on similar tasks, we say that the student has achieved a curriculum criterion (extrapolation), or the student is able to do certain things with language represented by his ability to complete future tasks of a similar nature (explanation). Here we are making two types of the extrapolation inference (extrapolation and explanation). When we use this information to make decisions about this student (e.g.,

TABLE 2 | Claims and inferences in CBFA.

CBFA Claims	Inference links
Claim 1: CBFA judgment is carried out appropriately	Evaluation: linking performance to judgment
Claim 2: CBFA judgment about student achievement is trustworthy	Generalisation: linking individual observation to generalised judgement over all possible observations
Claim 3: CBFA reflects students' expected language achievement	Explanation: linking judgment to interpretation against theoretical construct Extrapolation: linking judgment to interpretation against curriculum targets and teaching
Claim 4: CBFA is used to improve learning outcomes	Utilisation: linking interpretation to use

TABLE 3 | Warrants and their backing in CBFA.

Inference	Assumptions (warrants)	Evidence (backing)
Evaluation	Assessment targets and success criteria are clear; Elicitation tools appropriately chosen and used; and key procedures (elicitation, interpretation, feedback, action) of CBFA have been followed	Interviews of teacher and students to see their understanding of assessment targets and success criteria; Classroom discourse analysis to see assessment types and how they are carried out; and content analysis of classroom recordings to see how elicitation and interpretation are done, what feedback is provided, and what action is taken after feedback.
Generalisation	Classroom performance on language tasks is consistent across similar tasks, assessors, assessment forms and occasions	Multiple sources of evidence; Multiple observations; Sample observation tasks are representative of content domain tasks; and sample observation conditions are representative of content domain conditions
Explanation	Classroom assessment tasks engage the same abilities and processes as those in the theoretical construct of language competence appropriate for the context of teaching	Checking construct relevance and construct representativeness Interviews; Observation of assessment processes; Discourse/conversation analysis; and logical analysis of assessment tasks
Extrapolation	Assessment tasks and materials are representative of the knowledge, skills, and abilities targeted by the curriculum at the relevant level (content domain)	Judgmental evidence that assessment tasks are representative samples of the content domain; and logical analysis of assessment task content
Utilisation	Information provided to users are useful and sufficient (informing); and assessment information is used to adjust learning and teaching (forming)	Analysis of feedback (type, informativeness); Analysis of adjustment to learning and teaching; Analysis of adjustment to learning and teaching; Improved score in exams

he can go to the next level; or he needs more efforts to improve on this standard), we are making a utilization inference.

Since most assessment tasks in CBFA are contingent classroom activities, the assessor (mostly the teacher) makes judgements and decisions on the spot and does not wait till the end of the activity to interpret evidences of student learning. These explanation and extrapolation inferences and the judgements and feedback are much more closely bundled together than those a teacher makes at the end of a test. In addition, since the conceptualisation of CBFA in this framework does not assume formative effects being achieved, for the sake of parsimony, the Utilisation inference in the proposed IUA chain is not further parsed into sub-inferences.

Table 2 elaborates on the four major claims of classroom-based formative assessment. These four claims and their associated inferences make up the interpretation and use argument (IUA).

Step 2: Validity argument

After the articulation of the IUA, the next step is to argue with supporting reasons or warrants that all claims and inferences are plausible. In many cases, we also need to prove that alternative reasoning (rebuttal) is not supported by evidence; otherwise our claims will not stand if evidences are found to back up the rebuttals. **Table 3** lists the warrants and their potential backings for the validity argument of CBFA.

Argument-Based Validation of CBFA: An Example

Let's now look at a CBFA event, and see how it can be evaluated using the argument-based approach. Due to a lack of space, I will be deliberately short, and will not be illustrating all the details in the two-step validation process.

The following classroom assessment event forms a complete assessment cycle and should be counted as CBFA. Is this CBFA good enough for its intended purpose?

For the interpretation and use argument, I have largely indicated the list of inferences and claims for this CBFA event, although the wording is not in the format of a claim or inference. The IUA is illustrated in **Figure 4**.

Validity argument should next be provided for each of the above claims. I will take the explanation claim and show that it is not true (**Figure 5**). In other words, the teacher's interpretation of the assessment outcome is wrong. In these cases, no matter how useful the follow-up actions are, they will not help solve the targeted learning problem, thus not achieving the effect of CBFA.

- We had an in-class shared reading task today. I went around class and observed the students. My observation focused on three groups and I found a number of problems in understanding (evaluation).
- I realized that many students couldn't understand this type of reading (generalization).
- The students' lack of vocabulary is a concern (explanation).
- I told them they needed a larger vocabulary to become better readers; and assigned them a task to memorize 50 words a week from now on (utilization).

After generalising classroom observations of students' reading problems, the teacher could have arrived at the conclusion that the students/class were not achieving a particular curriculum target of reading, or that they would have problems reading similar texts in real world tasks (Extrapolation). She could have also inferred that this evidence in class revealed the students' deficiency or imperfect learning in certain areas of reading competence (Explanation). The teacher opted for the latter but identified a wrong component (vocabulary size) of the construct of reading as the cause of the problem in the Interpretation phase of this CBFA. While the transient nature of many CBFA events would make it unavoidable for some wrong interpretations of assessment data, this example illustrates the importance of teacher pedagogical content knowledge, a crucial aspect of assessment literacy that makes or breaks a formative assessment decision.

The CBFA cycle in this example may take slightly longer than normal to complete, because the action component comes after class. While the consequential aspect of the formative assessment cycle can only become possible after a full round, validity argument for each inference can be done any time during the whole spiralling process. This validity argument during the process as soon as an inference is made explicit in an IUA is a key part of the formative mechanism that makes flexible adjustment of teaching and learning possible. In the example, exercises in explicating the IUA inferences (**Figure 4**) make teachers more aware of their own decision-making processes in making use of assessment during instruction. Likewise, a validity argument (**Figure 5**) for each inference will help teachers decide whether and what changes are needed to achieve the formative effect. Without the validity argument, for example, the students may go on following the teacher's advice to remember more vocabulary items, and the real problem of reading identified at the elicitation stage may never been dealt with.

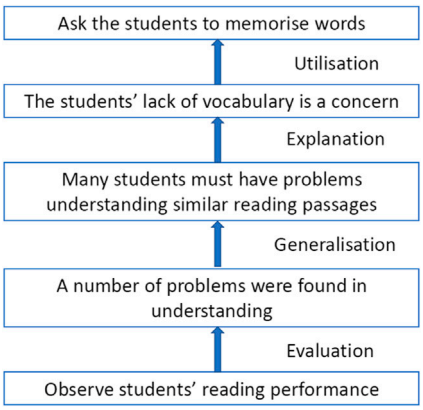

FIGURE 4 | Interpretation and use argument.

Who does CBFA validation, when, how?

Ideally, teachers themselves should validate their own CBFA as and when it happens in class. Teachers should also form communities of assessment practice in and beyond their own schools, so that peer teachers can help each other validate their CBFA. In addition, university researchers should join these communities of assessment practice every now and then to bring further theoretical and empirical expertise and to oversee that CBFA is done appropriately.

Both planned and contingent CBFA should be validated as often as needed, in any case, regularly. After all, as we have seen, despite its powerful potential, CBFA is only as good as the way it is used in class. Informal validation of CBFA should happen as and when it occurs in class. Formal validation can take the form of peer moderations and class observations. Teachers can also video-record their own classes for formal analysis at a later time. In the example above, the wrong interpretation of CBFA evidence could have been caught if the teacher or a peer validated her CBFA practices by going through her own video data of the lesson. She could then reinterpret the evidence available, and provide other alternatives of potential action in future classes. In addition, lesson plans can also be analysed for planned assessment practices and potential contingent CBFA.

CONCLUSION

In this article, I have offered an operational definition of formative assessment and classroom-based formative assessment. I argued that a clear operationalisation is the starting point for researchers and teachers alike to examine the validity and effectiveness of the formative assessment construct. Next, I contended that formative assessment is not necessarily useful in bringing about the desired formative effect, and that validation is needed for even informal and contingent classroom-based assessment events.

The argument-based approach to validation was next introduced. This includes two steps, an explication of the inferences we make from the assessment results followed by an argument for or against each inference using the Toulmin structure of argumentation. In other words, assessment validation is seen as systematically arguing that the interpretations and uses of assessment results are backed up by evidence and theory.

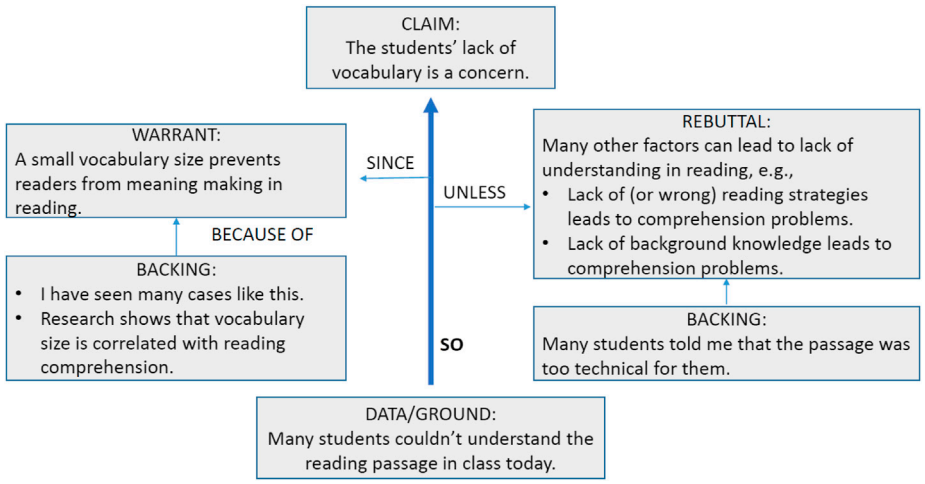

FIGURE 5 | The claim does not stand. Note that this is almost a "slow motion" replay of the argument-based validation process. In real life situations, CBFA validation can be done very quickly and informally without listing all the inferences on paper or drawing up the argument tree of warrants and backings.

Finally, I used an example from an English as a foreign language teacher's CBFA practice to illustrate how validation of CBFA can take place and how overturning one claim can invalidate the overall CBFA inference chain. The article finished by calling for more validations of CBFA not just for research purposes but also for teaching and teacher professional development purposes as well.

A clear operational definition will help teachers implement formative assessment inside their classrooms. A coherent and workable validation framework can assist teachers monitor and evaluate the interpretations and uses of their CBFA practices. This article points to a direction in which CBFA can be validated so that it achieves the formative effect of improved learning.

In using the proposed validation framework, we need to remind ourselves that validation is an ongoing process and that validity is not an either/or concept. Different CBFA events will show different degrees of validity when we go through a validation process. The more confident we are about our assessment outcomes and their interpretations and uses, the more likely we will achieve our intended formative effects.

The validation framework can also be seen as a useful tool for teacher learning. When teachers perform the acts of validation, they will immediately realise that the IUAs are mini-theories in their minds. These mini-theories include the set of criteria teachers make use of on the spot: explicit, latent, and meta-criteria (Sadler, 1985; Wyatt-Smith and Klenowski, 2013) about the nature of the knowledge or competence being assessed and about the criteria for success; they also include the teacher's understanding of how the knowledge is best learned or taught. These mini-theories guide the teacher's interpretation and use of the evaluative task. The more teachers perform validation of their own CBFA practices, the more they become aware of the adequacy of their pedagogical content knowledge behind their assessment. In this sense, validation practices as outlined in this article can also serve as a tool for teacher professional development.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

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

Andrade, H. L. (2010). "Summing up and moving forward: key challenges and future directions for research and development in formative assessment," in *Handbook of formative assessment*. Editors H. L. Andrade and G. J. Cizek (New York, NY: Routledge), p. 344–351.

Bachman, L. F., and Palmer, A. (2010). *Language assessment in practice: developing Language Assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Lang. Assess. Q.* 2 (1), 1–34. doi:10.1207/s15434311laq0201_1

Bailey, A. L., and Heritage, M. (2008). *Formative assessment for literacy, grades K-6: building reading and academic language skills across the curriculum.* Thousand Oaks, CA: Corwin.

Bell, B., and Cowie, B. (2001). *Formative assessment and science education*. New York, NY: Springer.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ. Principles Pol. Pract.* 18 (1), 5–25. doi:10.1080/0969594x.2010.513678

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ. Principles Pol. Pract.* 5 (1), 7–74. doi:10.1080/0969595980050102

Black, P., and Wiliam, D. (2005). Classroom Assessment is not (necessarily) formative assessment (and vice-versa). *Yearbook Natl. Soc. Study Educ.* 103 (2), 183–188. doi:10.1080/0969595980050102

Black, P., and Wiliam, D. (2012). "Developing a theory of formative assessment," in *Assessment and learning*. Editor J. Gardner. 2nd ed. (London, UK: SAGE Publications Ltd), p. 206–230.

Chapelle, C. A., Enright, M., and Joan, J. (2008). *Building a validity argument for the test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Thousand Oaks, CA: SAGE Publications, Inc.

Cowie, B., and Bell, B. (1999). A model of formative assessment in science education. *Assess. Educ. Principles Pol. Pract.* 6 (1), 101–116. doi:10.1080/09695949993026

Cronbach, L. J. (1988). "Five perspectives on validity argument," in *Test Validity Howard wainer and Henry I. Braun, 3–17* (Hillsdale, NJ: Lawrence Erlbaum Associates).

Davison, C., and Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Q.* 43 (3), 393–415. doi:10.1002/j.1545-7249.2009.tb00242.x

Davison, C. (2008). *Assessment for learning: building inquiry-oriented assessment communities*. New York, N.Y.: Routledge.

Hopster den Otter, D. H., Wools, S., Eggen, H. M. J. T., and Veldkamp, B. P. (2019). A general framework for the validation of embedded formative assessment. *J. Educ. Meas.* 56 (4), 715–732. doi:10.1111/jedm.12234

Dunn, K. E., and Mulvenon, S. W. (2009). A critical review of research on formative assessments: the limited scientific evidence of the impact of formative assessments in education. *Pract. Assess. Res. Eval.* 14 (7), 241. doi:10.4324/9780203462041_chapter_1

Gipps, C. V. (1994). *Beyond testing: towards a theory of educational assessment*. London, UK: Falmer Press.

Gu, P. Y., and Yu, G. (2020). Researching classroom-based assessment for formative purposes. *Chin. J. Appl. Linguist.* 43 (2), 150–168. doi:10.1515/cjal-2020-0010

Gu, P. Y. (2020). *Classroom-based formative assessment*. Beijing, BJ: Foreign Language Teaching and Research Press.

Hopfenbeck, T. N., Flórez Petour, M. T., and Tolo, A. (2015). Balancing tensions in educational policy reforms: large-scale implementation of assessment for learning in Norway. *Assess. Educ. Principles Pol. Pract.* 22 (1), 44–60. doi:10.1080/0969594x.2014.996524

Hopster-den Otter, D., Wools, S., Eggen, H. M. J. T., and Veldkamp, B. P. (2017). Formative use of test results: a user's perspective. *Stud. Educ. Eval.*, 52, 12–23. doi:10.1016/j.stueduc.2016.11.002

Kane, M. T., and Wools., S. (2019). "Perspectives on the validity of classroom assessments," in *Classroom Assessment and educational measurement*. Editors S. M. Brookhart and J. H. McMillan (New York, NY: Routledge), p. 11–26.

Kane, M. T. (1992). An argument-based approach to validity. *Psychol. Bull.* 112 (3), 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *J. Educ. Meas.* 38 (4), 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2006). "Validation," in Educational measurement. Editors R. L. Brennan 4th ed. (Westport, CT: American Council on Education/Praeger), p. 17–64.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50 (1), 1–73. doi:10.1111/jedm.12000

Kingston, N., and Nash, B. (2011). Formative assessment: a meta-analysis and a call for research. *Educ. Meas. Issues Pract.* 30 (4), 28–37. doi:10.1111/j.1745-3992.2011.00220.x

Klenowski, V. (2011). Assessment for learning in the accountability era: queensland, Australia. *Stud. Educ. Eval.* 37 (1), 78–83. doi:10.1016/j.stueduc.2011.03.003

Mislevy, R. J., Steinberg, L. S., and Almond, R. G. (2003). Focus article: on the structure of educational assessments. *Meas. Interdiscip. Res. Perspec.* 1 (1), 3–62. doi:10.1207/s15366359mea0101_02

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

Pearson Education. (2005). Achieving student progress with scientifically based formative assessment White paper, Pearson Education Ltd. Available at: http://www.pearsoned.com/wp-content/themes/pearsoned.com_legacy/pdf/RESRPTS_FOR_POSTING/PASeries_RESEARCH/PA1.%20Scientific_Basis_PASeries%206.05.pdf (Accessed September 6, 2014).

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

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

Ramaprasad, A. (1983). On the definition of feedback. *Syst. Res.* 28 (1), 4–13. doi:10.1002/bs.3830280103

Ruiz-primo, M. A., and Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *J. Res. Sci. Teach.* 44 (1), 57–84. doi:10.1002/tea.20163

Ruiz-Primo, A. M., and Furtak, E.M. (2006). 'Informal formative assessment and scientific inquiry: exploring teachers' practices and student learning'. *Educ. Assess.* 11 (3), 205–235.

Ruiz-Primo, M. A. (2011). Informal formative assessment: the role of instructional dialogues in assessing students' learning. *Stud. Educ. Eval.* 37 (1), 15–24. doi:10.1016/j.stueduc.2011.04.003

Sadler, D. R. (1985). The origins and functions of evaluative criteria. *Educ. Theor.* 35 (3), 285–297. doi:10.1111/j.1741-5446.1985.00285.x

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instr. Sci.* 18 (2), 119–144. doi:10.1007/bf00117714

Sadler, D. R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assess. Eval. Higher Educ.* 35 (5), 535–550. doi:10.1080/02602930903541015

Shavelson, R. J, Donald, B. Y., Carlos, C. A., Paul, R. B., Erin, M. F., Maria, A. R. P., et al. (2008). On the impact of curriculum-embedded formative assessment on learning: a collaboration between curriculum and assessment developers. *Appl. Meas. Educ.* 21, 295–314.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15 (2), 4–14. doi:10.3102/0013189x015002004

Stobart, G. (2012). "Validity in formative assessment," in *Assessment and learning*. Editor J. Gardner. (London, UK: SAGE Publications Ltd), p. 133–146.

Torrance, H., and Pryor, J. (1998). *Investigating formative assessment: teaching, learning and assessment in the classroom*. Maidenhead, UK: Open University Press.

Toulmin, S. E. (2003). *The uses of argument*. 2nd edn. New York, NY: Cambridge University Press.

Wiliam, D. (2007). Changing classroom practice. *Educ. Leadersh.* 65 (4), 36–42.

Wiliam, D. (2010). "An integrative summary of the research literature and implications for a new theory of formative assessment," *in Handbook of formative assessment. H. L. Andrade and G. J. Cizek (New York, NY: Routledge),* p. 18–40.

Wyatt-Smith, C., and Klenowski, V. (2013). Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice. *Assess. Educ. Principles Pol. Pract.* 20 (1), 35–52. doi:10.1080/0969594x.2012.725030

Xu, Y., and Harfitt, G. (2019). Is assessment for learning feasible in large classes? Challenges and coping strategies from three case studies. *Asia-Pac. J. Teach. Educ.* 47 (5), 472–486. doi:10.1080/1359866x.2018.1555790

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.