CLASSROOM-BASED ASSESSMENT IN TEFL AS A FORM OF FORMATIVE ASSESSMENT: ENQUIRIES INTO THEORY AND PRACTICE

doi.org/10.61425/wplp.2023.18.80.102

An Hai Trinh
Eötvös Loránd University, Budapest
antrinh@student.elte.hu

Abstract: Closely related to the concept of formative assessment, classroom-based assessment (CBA) has received increasing attention from education researchers and policy makers worldwide. Despite being hailed as an innovative departure from traditional standardized testing, CBA has often been criticized for the lack of research-based evidence to support its purported benefits. This raises concerns about the reliability, validity, and practicality of this approach in mainstream education. By reviewing recent literature on CBA and its application in TEFL classrooms, this article seeks to understand how CBA theory translates into practice and identify potential discrepancies between its claimed advantages and measured efficiency. The discrepancies observed are primarily attributed to teacher assessment identity. Consequently, I propose a CBA literacy model which improves teachers and students' assessment capability in classroom contexts.

Keywords: classroom-based assessment, formative assessment, language assessment, TEFL, assessment literacy, teacher assessment identity.

1 Introduction

Assessment plays an important role in the process of teaching and learning. Based on its purpose, assessment can be classified into two major types: summative and formative assessment (Bachman, 1990; Bachman & Damböck, 2017; Bachman & Palmer, 2010). Summative assessment can take the form of entrance tests, placement tests, diagnostic tests, achievement tests, and summative evaluation, and is often used to determine the test-takers' proficiency at a particular point in time, serving as an endpoint measure. Formative assessment, on the other hand, is continuous assessment integrated into the teaching process. Its purpose, as stated by Bachman and Damböck (2017), is to "improve instruction and learning" (p. 11).

While summative assessment has long been a focus of research and discussion from around the world, the discussion of formative assessment appears to be a relatively more recent development (Lewkowicz & Leung, 2021). Formative assessment is often viewed as a means to reform language assessment (Alderson & Banerjee, 2001; Davison & Leung, 2009; Hill & McNamara, 2012; Turner, 2012; Turner & Purpura, 2016). In what has been called the "alternative assessment movement" (Alderson & Banerjee, 2001, p. 228), proposals have been put forward calling for increased emphasis on improving students' learning rather than on summarily measuring their language ability at a particular point in time. The purpose of these efforts is to overcome the shortcomings and undesirable impacts of formal, high-stakes summative tests (Alderson & Banerjee, 2001).

Formative assessment has taken on different names over the years, such as *assessment for learning*, *teacher-based assessment*, *alternative assessment, informal assessment* (Brown & Hudson, 1998; Clapham, 2000; Hamayan, 1995) and, more recently, *classroom-based assessment* (Lewkowicz & Leung, 2021). In this article, I will use the term classroom-based assessment (CBA) to refer to assessment serving formative purposes in relation to the EFL classroom. More specifically, I define CBA as encompassing all activities undertaken by teachers and/or their students to gather information on a student's performance or language use, which is used to modify the teaching and learning activities in which they are engaged. The formulation of this definition is influenced largely by three sources: Black and William (1998), Turner (2012), and Lewkowicz and Leung (2021), which will be elaborated on further in Section 2 below.

While the benefits of CBA have been frequently cited by its advocates (e.g. Giménez, 1996), critics of CBA highlight that accounts of its efficiency are typically "descriptive and persuasive, rather than research-based" (Alderson & Banerjee, 2001, p. 229). In the context of this paper, efficiency will be used to refer to the extent to which an assessment approach, activity or procedure succeeds in measuring language proficiency. In Section 4, the efficiency of CBA procedures will be discussed in terms of their reliability, validity, and practicality. Clapham (2000) expressed reservations about CBA, arguing that its advocates "do not appreciate the importance of investigating the reliability and validity of their instruments" (p. 150). Similarly, Brown and Hudson (1998) criticised the lack of empiricism in CBA advocacy:

> Certainly, we would agree that credibility, auditability, multiple tasks, rater training, clear criteria, and triangulation of any decision-making procedures along with varied sources of data are important ways to improve the reliability and validity of any assessment procedures used in any educational institution. In fact, these ideas are not new at all. What is new is the notion that doing these things is enough, that doing these things obviates the necessity of demonstrating the reliability and validity of the assessment procedures involved. (p. 656)

The criticisms of how CBA is presented and practised raise suspicions of potential discrepancies between the theoretical advantages of CBA and the efficiency of CBA activities in practise. In light of the above issue, this article reviews recent literature on both the theory and practice of CBA in foreign language education. Following the introduction in Section 1, Section 2 reviews studies aimed at defining CBA and describing its characteristics. In Section 3, I review recent empirical studies on CBA procedures, focusing on three types of CBA activities: portfolio, peer assessment, and self-assessment, with special attention paid to studies related to Teaching English as a Foreign Language (TEFL). However, discussions and findings of other studies related to other areas of language education and general education are also included where applicable. In Section 4, I discuss the observed discrepancies between general theoretical claims and empirical findings of the efficiency of specific CBA procedures in terms of their reliability, validity, and practicality. In Section 5, I attempt to explain the observed discrepancies. In Section 5, I attempt to explain these observed discrepancies through the concept of teacher assessment identity (Looney et al., 2017). In addition, this section offers recommendations on how to minimize these discrepancies in order to better realize the potential of CBA and suggests directions for future research on this assessment approach.

2 Classroom-Based Assessment Theory and related terms

2.1 Historical background

The concept of CBA is closely related to formative assessment, which originated in the field of general education. In the late 1980s, the Task Group on Assessment and Testing (TGAT) in England published a report which highlighted "the conflicts between high-stakes public reporting of student performance in accordance with statutory curriculum standards, and educationally oriented assessment that reports student progression in the context of teaching and learning experience" (Lewkowicz & Leung, 2021, p. 47). The chairs of the TGAT later published a paper titled '*Assessment and classroom learning*', which has since been regarded as seminal work on formative assessment (Black & William, 1998).

Following these developments, researchers in the field of foreign and second language education began to explore issues related to formative assessment, including the quality of teacher-student interaction, the importance of feedback, and the role of the student in the assessment process (Lewkowicz & Leung, 2021). This resulted in a growing, albeit scattered and segmented body of literature demonstrating significant interest in formative assessment and how it can be applied to enhance the learning of foreign and second languages (Turner & Purpura, 2016).

Currently, CBA is a "policy-supported practice in a number of educational systems internationally", though it is often "over-shadowed by national testing programs" (Davison & Leung, 2009, p. 393). There is an increasing trend to devolve the responsibility for assessment to classroom teachers (Hill & McNamara, 2012). Within this context, recent curriculum developments have emphasized the need for English language teachers to be knowledgeable and skilled in CBA, calling upon teachers to perform new instructional and assessment roles in the classroom (Brown & Hudson, 1998; Davison & Leung, 2009).

2.2 Definitions

As shown below, there have been a number of definitions, a bewildering array, put forth to conceptualize CBA, each overlapping with and differing from each other in subtle ways. In reviewing these definitions, I have identified five common dimensions that serve as a framework for comparison, which are listed below. The definition formulated for this study (see Section 1) will serve as reference. For ease of reading, as much as possible, I present all definitions in their original, complete sentences, highlighting keywords in bold.

Dimensions	**Example** (from the definition formulated for this study)
Terminology	Classroom-based assessment (**CBA**)
Reference	refer to all those **activities**
Agents and actions	undertaken by **teachers and/or their students**
Focus of action	to gather information on a **student's performance on curriculum tasks**
Use of information	the information gathered to be used as **feedback for modifying the teaching and learning** in which the teachers and students are engaged.

Table 1. Dimensions of definitions of CBA

The table below analyses the definitions used in the studies reviewed, most of which are referenced in Lewkowicz and Leung's (2021) CBA research timeline.

Researcher	**Terminology**	**Reference**	**Agents and actions**	**Focus of action**	**Use of information**
Black & William (1998, p. 7)	Formative assessment	is to be interpreted as encompassing all those activities	undertaken by teacher and/or their students	which provide information	to be used as feedback to modify the teaching and learning activities in which they are engaged.
McNamara (2001, p. 343)	Classroom-based assessment (CBA)	is any deliberate, sustained, and explicit reflection	by teachers and learners	on the qualities of a learner's work	and the use of this information, for example, as an aid to the formulation of learning goals.
Hill & McNamara (2012, p. 396)	Classroom-based assessment (CBA)	is any reflection	by teachers and/or learners	on the qualities of a learner's or group of learners' work	and the use of that information for teaching, learning (feedback), reporting, management or socialization purposes.
Davison & Leung (2009, p. 395)	Teacher-based assessment (TBA)	refers to practices and procedures	which involves the teacher from the beginning to the end	and allows for the collection of a number of samples of student work over a period of time, using a variety of different tasks and activities	opens up the possibility for teachers to support learner-led enquiry; allows the teacher to give immediate and constructive feedback to students; stimulates continuous evaluation and adjustment of the teaching and learning programme; complements other forms of assessment, including external examinations.
Turner (2012, p. 65)	Classroom assessment	refers to strategies	by teachers to plan and carry out	the collection of multiple types of information concerning	to analyse and interpret; to provide feedback;

				student language use	to make decisions to enhance teaching and learning.
Turner & Purpura (2016, p. 261)	Learning-oriented assessment (LOA)	refers to strategies that involve:	teachers' instructional activities; students' active engagement in the assessment process; teachers and students' responsibility to provide and respond to quality feedback	in the collection of information to support and enhance learning	to guide and support the learning process.
Lewkowicz & Leung (2021, p. 48)	Classroom-based assessment	is any classroom activity	led by teacher	designed to find out about students' performance on curriculum tasks that would yield information regarding their understanding	as well as their need for further support and scaffolding with reference to their situated learning needs.

Table 2. Analysis of definitions of CBA

One distinguishing feature of CBA is its emphasis on utilizing the information gathered about the student's performance to modify teaching and learning activities in ways that support learning (see the "Use of information" column in Table 2 above). Without such use of information, the assessment could not be considered to be formative in function (Black & William, 1998). Sadler (1989) highlighted that:

> If the information is simply recorded, passed to a third party who lacks either the knowledge or the power to change the outcome, or is too deeply coded (for example, as a summary grade given by the teacher) to lead to appropriate action, the control loop cannot be closed. (p. 121)

Due to its focus on leveraging assessment to serve the learning process, CBA and formative assessment are closely related to the concept of Assessment for Learning (AFL), used by the curriculum authority of the United Kingdom (Leung, 2004), which involves both teachers and students in the review of students' learning progress. This process of reflection and feedback is used to identify ways to adjust teaching and learning strategies. AFL is often used in contrast to Assessment of Learning (AOL) (Leung, 2004).

2.3 Forms of CBA

Just as there have been different definitions put forth to conceptualize CBA, there have been different ways used by researchers to classify CBA activities.

For the purpose of helping language teachers decide what types of language tests to use in their institutions and classrooms, Brown and Hudson (1998) classified language assessments into three broad categories: (1) Selected-response, which include true-false, matching, and multiple-choice assessments; (2) Constructed-response, which include fill-in, short-answer, and performance assessments; and (3) Personal-response, which include conference, portfolio, self-assessment and peer-assessments (p. 658).

Hill and McNamara (2012) acknowledged the above forms of assessment and used the term *formal assessment* to classify them. In addition, the authors called for attention to more intuitive, incidental, and embedded forms of assessment, such as those taking place within the discourse structure of teacher questions and pupil responses (p. 398). Consequently, the authors used three terms to distinguish types of assessment, namely: (1) formal assessments, which are planned and evident, e.g. tests and assignments; (2) planned assessment opportunities, which are instruction-embedded and planned, e.g. teaching activities also used for assessment; and (3) incidental assessment opportunities, which are instruction-embedded and unplanned, e.g. unstructured observation (p.403).

2.4 CBA as an instructional activity

Over the past five decades, substantial changes have been seen in the theory and practice of language assessment. Influenced by what can be described as a sociolinguistic revolution, considerable changes have occurred regarding the conceptualisation of teaching (Richards & Rodgers, 2014) and the practice of assessment in the classroom (Bachman & Damböck, 2017). The focus of language assessment has moved beyond standardised testing, where a single score, such as that of a multiple-choice test, is expected to provide a complete picture of students' abilities.

Current trends in language education view language assessment essentially as an instructional activity that is embedded within the process of teaching and learning. Bachman and Palmer (2010, p. 20) described assessment as an evidence-based, systematically planned process of collecting information to interpret language performance. Likewise, in their handbook for language teachers, Bachman and Damböck (2017) reiterated the centrality of the process-oriented approach to assessment. In this sense, assessment is part of a complex process for making interpretive arguments about the learner's ability and making well-informed decisions for the common good of the educational stakeholders, including learners, teachers, and administrators. This emphasis on assessment as a part of the education process contrasts with the traditional view of assessment as a product of teaching and learning.

The idea that CBA can function as an instructional activity embedded within the teaching process might pose a challenge to teachers' conventional views and practices of assessment which view it as an activity distinct and separable from teaching. For example, in the traditional view of assessment, the teacher can usually identify the exact moment of assessment, such as when administering a paper-based test. In contrast, CBA often involves more nuanced activities such as asking oral questions during a lesson, observing students' responses to gauge their understanding, and responding to students' answers on the spot.

2.5 Comparing CBA with large-scale standardized testing

CBA is typically distinguished from "traditional externally set and assessed large-scale formal examinations used primarily for selection and/or accountability purposes" (Davison & Leung, 2009, p. 395). In the table below, I summarise and categorise several common points of comparison (i.e. Alderson & Banerjee, 2001; Brown & Hudson, 1998; Davison & Leung, 2009).

	CBA	**Large-scale standardized testing**
Setting	Ordinary classrooms	Exam hall or specialist assessment centre
Time	Data is gathered over an extended period	Data is taken at one point in time
Conducted by	The students' own teacher	A stranger
Function	Formative in function	Summative in function
Consequences	Low-stakes in consequences	High-stakes in consequences
Washback[1]	Claimed to have beneficial effects on teaching and learning	Criticized for negative effects on teaching and learning
Implementation challenges	Time-consuming Difficult to administer and score	Less time-consuming Usually has clearly defined assessment criteria
Use	Used to enhance teaching and learning	Used primarily for selection and/or accountability purposes
Advantages	Claimed to: - be more easily integrated into day-to-day classroom activities - allow students to be assessed on what they normally do in class everyday - focus on processes as well as products - tap into higher-level thinking and problem-solving skills - be multiculturally sensitive when properly administered - provide easily understood information - can be adapted and modified to match the teaching and learning goals of the particular class and students - allows the teacher to give immediate and constructive feedback to students	(not explicitly mentioned in the studies reviewed)

[1] The term 'washback' refers to the impact that tests have on teaching and learning (Alderson & Banerjee, 2001)

	- stimulates continuous evaluation and adjustment of the teaching and learning programme - complements other forms of assessment, including external examinations	
Purposes	Often developed in an attempt to: - Make testing and assessment more responsive and accountable to individual learners - Promote learning - Enhance access and equity in education	(not explicitly mentioned in the studies reviewed)

Table 3. Comparison of CBA and traditional testing

Table 3 above shows an impressive list of positive characteristics of CBA that should appeal to most language teachers, testers and policy makers alike (Brown & Hudson, 1998). However, critics of CBA often point out the lack of robust research-based evidence to substantiate claims of its benefits and advantages. Clapham (2000), for example, states the following:

> A problem with methods of alternative assessment, however, lies with their validity and reliability. Tasks are often not tried out to see whether they produce the desired linguistic information; marking criteria are not investigated to see whether they 'work'; and raters are often not trained to give consistent marks. (p. 152)

Given the above context, the following section reviews empirical studies of CBA as practiced over the last two decades, focusing on the extent to which CBA procedures succeeded in measuring language proficiency, and what insights they can provide on the practice of CBA in TEFL.

3 The practice of CBA in TEFL

In this section, I review empirical studies on CBA practices in TEFL that have been conducted since 2000. While many quantitative CBA studies have addressed the efficiency of CBA practices (Joo, 2016; Mak & Wong, 2018; Nunes, 2004; Song & August, 2002), qualitative research on the topic often seeks to explore teachers' and students' perceptions of CBA (Davison, 2004; Davison, 2007; Wicking, 2017). Three types of CBA activities are selected for review: portfolio, peer assessment, and self-assessment (Brown & Hudson, 1998; Lewkowicz & Leung, 2021)

3.1 Portfolio

A portfolio is a collection of a student's work, usually constructed by selecting several diverse samples produced at different times. Portfolio assessment programs typically require teachers to plan portfolio tasks and lessons, coach students on drafts, and help them compile and evaluate their portfolios (Song & August, 2002). In foreign language education, particularly with regard to writing assessment, portfolios have been hailed as a major innovation, supposedly overcoming the limitations of one-off impromptu single writing tasks often found in traditional testing (Alderson & Banerjee, 2001).

In traditional writing tests, students are usually given only one or two tasks, on the basis of which generalisations about writing ability across a range of genres are often made. With portfolio assessment, students are given multiple opportunities to showcase their ability across different genres. Students also have more time for planning, researching, editing, redrafting, and revision - activities that are fundamental in much of real-world writing (Alderson & Banerjee, 2001). Brown & Hudson (1998) listed three advantages for portfolio assessment, also falling into three categories: strengthening students' learning, enhancing the teacher's role, and improving testing processes.

However, Black & William (1998) pointed out a lack of research-based evidence, beyond teachers' accounts validating the learning advantages of portfolio assessment. Concerns have also been raised about the reliability of teachers in scoring portfolios. Against this backdrop of criticism, it is important to acquire data that can establish both the reliability and validity of portfolio assessment if it is to develop into a viable alternative to less satisfactory approaches (Hamp-Lyons & Con, 2000). The studies reviewed below attempt to address these concerns.

Song and August (2002) compared the performance of two groups of ESL students at Kingsborough Community College, part of the City University of New York. To be admitted into an advanced writing course, one group was assessed on the basis of both portfolios and a standardized test, while the other group was assessed with standardized test only. The study found no significant difference in score distribution and pass rate between these two groups at the end of the course. This suggests that portfolio assessment is as valid as standardized testing in predicting student success in subsequent English courses. Based on these findings, portfolio assessment emerges as a potential alternative to standardised testing for the evaluation of language proficiency.

Another encouraging outcome from this study was that ESL students were twice as likely to meet the course admission requirements when they were evaluated on the basis of the portfolios as compared to when they were evaluated using standardized testing. In other words, portfolio assessment identified more ESL students who subsequently succeeded in the advanced English course than the standardized test did. This suggests that portfolio assessment may be a more appropriate assessment alternative for evaluating ESL students. Finally, the study attributed the efficiency of this assessment in assessing writing proficiency to the careful design and execution of the assessment along with clear evaluation standards.

Nunes (2004) explored the use of portfolios in EFL high school classrooms in Portugal. The study's findings suggest that portfolio assessment enables teachers to better diagnose their students' skills and competencies. In addition, it increased their awareness of their students' preferences, styles, dispositions, and learning strategies. This information can help teachers make necessary changes to their teaching, which in turn can translate into a more learner-centred approach in education.

Coombe and Barlow (2004) examined the planning and implementation of two portfolio assessment initiatives at two tertiary institutions in the UAE. The student participants were asked to write portfolio entries and fill out survey questions that accompanied each entry. The researchers found that the inclusion of a reflective element in the portfolio strengthened students' EFL writing skills. However, the participants were taken aback by the amount of time and work required to complete the activity. This highlights a potential challenge in implementing portfolio assessment, as it may increase teachers' workloads.

Lam (2017) reviewed 66 studies on second language portfolio assessment and identified two strands of evidence. The positive evidence shows that portfolio assessment is capable of increasing learners' confidence, motivation, and sense of ownership. The negative evidence, on the other hand, points to challenges that impede the implementation of portfolio assessment, such as a fixation on grades, perceptions of fairness, and issues with learner agency (i.e., students' self-regulation). The author suggested three strategies to maximize the application of portfolio assessment in second language classrooms. The first is to improve learner agency by training students to be more autonomous and reflective in monitoring their writing profiles, thereby improving their language skills through the use of portfolios. The second is to develop teachers' expertise in portfolio assessment, requiring a conceptual shift in their understanding of assessment purposes and practices using reflective enquiry. The third strategy calls for the establishment of a "portfolio culture" that supports the practice of reflection and self-assessment in writing.

Mak and Wong (2018) conducted a multiple case study of portfolio assessment practices in EFL elementary classrooms in Hong Kong. Over the course of one academic year, the researchers studied the effects of using portfolio assessment on students' self-regulation. Data sources included classroom observations, interviews, and field notes. The findings couched portfolio assessment as an empowering tool contributing to the development of students' self-regulated learning through goal setting and scaffolding.

Davison (2004) explored the impact of cultural differences on teachers' assessment. The study does not focus on portfolio assessment specifically but on writing assessment in general. The study involved 12 ESL teachers from Australia and Hong Kong, who were asked to assess students' argumentative writings and explain their decisions. Group interviews and discussions with the researcher were conducted one week later. The results showed that while the Australian group assessed students on the basis of their adherence to the published assessment guidelines, the Hong Kong group reached their assessment decisions through community norms. This finding raises the concern that, due to cultural differences, teachers from different regions or countries may possess varying perceptions of assessment, which may lead to differences in their scoring practices (Davison, 2004). This variability could pose a challenge to the consistency and fairness of the assessment when applied in different geographical contexts.

To conclude this section on Portfolio, the reviewed literature suggests that portfolio assessment can be applied efficiently, offering both affective and cognitive benefits to learners. However, for this to happen, special attention must be paid to the design, implementation, and testing of each specific activity. Long-term commitments are needed for teacher training as well as the cultivation of a culture that supports its practice. This approach requires substantial changes in the way teachers manage their workload and classroom roles. Other challenges to the implementation of portfolio assessment in the classroom include the fixation on grades, concerns about fairness, learner agency, and the impacts of geo-cultural differences.

3.2 Peer assessment

Peer assessment requires students to rate the performance of their peers (Brown & Hudson, 1998). Using their capacity to recognize and appraise performance gaps, students collaborate in assessing one another during classroom activities such as group discussion and group project. Cited advantages of peer assessment include, through group collaboration, peer assessment stimulates thinking, increases learning opportunities, and improves individuals' performances (Black & William, 1998).

Issues regarding peer assessment include the reliability of peer assessment, the measurability of learning gains, the setting of criteria, the composition of group, and the training of students in group processes (Black & William, 1998). Some of these issues were investigated in the studies reviewed below. Findings indicate a limited degree of efficiency, which may be enhanced under certain conditions.

Cheng and Warren (2005) studied first-year students (n = 51) at a Hong Kong university, asking them to assess their peers' English language proficiency as exhibited in the seminar, oral presentation, and written report of an integrated group project. These peer ratings were then compared with those of their class teachers' in terms of their means and standard deviations. Findings demonstrated a general agreement between these ratings, although the students tended to mark within a narrower range than their class teacher. The researchers expected that the students might award a wider range of marks if they were given more opportunities to practise and experience peer assessment procedures. The standard deviation of students was approximately half of that of their teacher's. The study reported the positive impact peer assessment had on teachers and students, who found the exercise beneficial in terms of developing students' higher level cognitive thinking and facilitating a deep approach to language learning.

Hirai et al. (2011) examined the reliability of peer assessment among first-year Japanese university students (n = 80), who were asked to rate their peers' performance on an English story retelling speaking test. The performances were also subsequently assessed by the researchers, whose assessments qualified as the teachers' assessment. The scoring scale was based on three criteria: (1) communicative efficiency, (2) grammar & vocabulary, and (3) pronunciation.

The results showed weak correlations between peer and teacher assessment. Only the correlation regarding the communicative efficiency criterion was marginally significant. Regarding the other two criteria (grammar & vocabulary, and pronunciation), peer assessment tended to deviate from teacher assessment. To explain this variation, the researchers suggested that it was relatively easier for students to assess communicative efficiency because it primarily measures the fluency of students' oral performance. The other two criteria were more difficult to assess because they demand that raters have comprehensive knowledge of English grammar, vocabulary, and pronunciation.

Peers tended to be lenient when rating the performances of fellow students, particularly on grammar & vocabulary. This may be because the students were not sure of the correct grammar and vocabulary. Notably, the anonymity of raters was revealed to be an important factor in improving the reliability of peer assessment, a condition which is difficult to maintain in actual classroom settings. The authors therefore concluded that peer assessment cannot always be a reliable substitute for teacher assessment, particularly when single-score ratings are used. Peer assessment may instead be more beneficial for collaborative learning by providing feedback rather than evaluating students (p. 55).

Sato & Lyster (2012) investigated whether students can be trained to provide corrective feedback during peer interaction. The study involved 167 students in four university-level English classes in Japan. The treatment groups participated in a three-week training program with three stages: modelling, practice, and use-in-context. First, two teachers demonstrated how to give feedback. Next, students practised with a role-play scenario. Last, the students applied the learned skills in an authentic context. After one semester of intervention, the students improved in both overall accuracy and fluency. The result indicated that, with well thought out training and frequent practice, students could learn to provide peer feedback in ways that improves performance. By providing corrective feedback to their peers, learners sharpened the ability to monitor both their own language production and that of their interlocutors.

Zhao (2014) explored whether, with the support of teacher intervention, peer assessment can complement teacher assessment in the facilitation of EFL writing instruction. Over the course of four months, 18 second-year English majors at a university in China were asked to provide written feedback for their peers on nine writing tasks across five different genres. Four teacher intervention strategies were employed in the process: (1) training students in providing constructive feedback, (2) analysing the feedback to identify and address problematic areas, (3) commenting on the appropriateness of the feedback to confirm its validity, and (4) providing support to students who sought help in settling disagreements with peer collaborators.

The results showed that teacher intervention had a positive impact on peer assessment. Training students in providing peer feedback improved the amount and quality of peer feedback. Teacher comments on the appropriateness of peer feedback addressed students' concerns over its validity, hence encouraged them to use peer feedback in their revision and learning. The study highlighted the essential role of the teacher in facilitating the efficient use of peer assessment.

Saito (2008) examined the effects of training on peer assessment and peer comments on oral presentations in EFL classrooms. In the first study, 74 Japanese university freshmen were asked to rate and comment on their classmates' oral presentations. Their ratings were then compared with the instructors' ratings. Both the treatment groups and control groups received instruction on the 12 skill aspects of presentation directly associated with the assessment items, but only the treatment groups had an additional 40-minute training on how to rate performances. In terms of ratings, the results of the correlation difference analyses showed no significant differences between the treatment groups and control groups. In terms of comments, the analyses revealed that the treatment groups were superior in both the quality and quantity of comments.

The study concluded that peer assessment is fairly robust (reliable without much training), to the extent that instruction on skill aspects may suffice to achieve a certain level of correlation with the criterion variable (instructor), and rater training may not provide further improvement in correlation. However, rater training may enhance student comments and reduce mis-fitting raters. By drawing student attention to the features of a language learning task, peer assessment training may facilitate the learning process by increasing students' consciousness of the performance criteria.

In sum, the reviewed literature on peer assessment shows a limited degree of its efficiency as a reliable assessment. In the first two studies, Cheng & Warren (2005) and Hirai et al.'s (2011) analyses revealed moderate to weak correlations between peer and teacher assessment. The level of correlation varied across different aspects of assessment, being significant according to some aspects (e.g. the fluency of oral performance), but not according to others, especially not those that require a high level of linguistic knowledge (e.g. grammar & vocabulary). In Hirai et al. (2011), the anonymity of the rater appeared to be an important factor affecting the reliability of peer assessment. Students tended to mark within a narrower range than their class teacher and tended to be more lenient when rating the performances of fellow students. These limitations led some researchers to caution against its use as a reliable substitute for teacher assessment, especially when the assessment carries high-stake consequences. They recommend that peer assessment: (1) may be more beneficial for collaborative learning by providing feedback rather than evaluating student; (2) should be considered for inclusion in academic programmes based on the positive impact it can have in other aspects; (3) can be improved with training, practice, and experience.

The remaining studies investigated these issues and reported positive results. Sato & Lyster (2012) investigated whether students can be trained to provide corrective feedback during peer interaction. Zhao (2014) explored whether, with the support of teacher intervention, peer assessment can complement teacher assessment in the facilitation of EFL writing instruction. Saito (2008) examined the effects of training on peer assessment and peer comments on oral presentations in EFL classrooms. These studies highlight that peer assessment, with appropriate intervention and modification, can be a useful tool that benefits both teachers and students.

3.3 Self-assessment

Self-assessment requires students to rate their own performance. Three common types of self-assessment are: (1) performance self-assessment, which requires students to read a situation and decide how well they would respond in; (2) comprehension self-assessment, which requires students to read a situation and decide how well they would comprehend it; (3) observation self-assessment, which requires students to listen to audio- or videotape recordings of their own language performance and decide how well they think they performed (Brown & Hudson, 1998).

One advantage of self-assessment is that it “require(s) little extra time or resources” (Brown & Hudson, 1998, pp. 53-54). A common motive for including self-assessment in an academic program is the belief that peer assessment can help students to take more responsibility for their own learning (Black & William, 1998). By giving students greater autonomy in the assessment process, self-assessment has the potential to increase students’ commitment and motivation to learn, which may translate into improvement in their learning achievement (Brown & Hudson, 1998). To contrast, one concern with self-assessment is the students’ capacity to assess their own performance. Another concern is the subjective factors that might affect self-assessed scores, such as past academic records, career aspirations, peer-group or parental expectations, high-stake contexts, and students’ vested interests (Brown & Hudson, 1998). The studies reviewed below investigated some of these issues.

Fukazawa (2011) examined the validity of self-assessment of speech performance in a high school setting in Japan. Following Messick’s (1996) interpretation of validity, this study investigated the validity of self-assessment from five aspects: the content, substantial, structural, external, and consequential aspect. Fifty-two students in the 11th grade and three teachers participated in this study. Each student made a two-minute prepared speech, after which he/she was given one minute to assess his/her own performance. Prior to their speeches, the students had received rater training with a training video and completed a proficiency test.

Within a week from the last speech in the class, the students completed a questionnaire on the self-assessment.

The results showed that the content and substantial aspects were considered to have sufficient validity. However, the structural, consequential, and external aspects of validity were not sufficient. Consequently, the researcher suggested that self-assessment has a degree of validity, but it is not sufficient for summative purposes. The researcher also compared these results on self-assessment with that of Fukazawa (2009) on peer assessment, which was conducted and analysed with almost the same procedures using the same validity framework. While the results on peer assessment indicated that peer assessment had comparable validity to teacher assessment, the results on self-assessment seem to show a weaker validity.

Summers et al. (2019) used a validation framework to evaluate the usefulness of a self-assessment survey which were created with consultation from the American Council on the Teaching of Foreign Languages (ACTFL). Participants were 92 students newly enrolled in an Intensive English Program (IEP) at a large private university in the United States of America, all of whom came from foreign countries to learn English in a full-time ESL program. Results showed that the self-assessment instrument can reliably discriminate between examinees. However, the correlations between self-assessment and placement test results were weak. Hence, the use of the instrument for high-stakes purposes is not supported by the study. While the results suggested caution in the use of self-assessment as a sole measure for placement purposes, there might be potential for self-assessment to complement other measures. Additionally, since students were consistent in their self-evaluations, the researchers suggested that self-assessment might be valuable in tracking learning gains over time, a direction which was taken up by the Ma and Winke (2019), as reviewed below.

Ma and Winke (2019) investigated the extent to which students could reliably use self-assessment to track their language gains over time. The study used the data of eighty students at Michigan State University, who took an oral assessment two years in a row (n=80). The students completed: (1) a background questionnaire; (2) an oral skills self-assessment, which was developed in consultation with the American Council on the Teaching of Foreign Languages (ACTFL), and contained five sets of ten Can-Do statements, each set covering a range of ACTFL levels; (3) the computerized Oral Proficiency Interview (OPI - a standardized, global assessment of assessing how well a person speaks a language). Results of the self-assessment and OPI were compared.

While the correlation was moderate to strong with students at the Novice and Advanced levels of proficiency, there was no correlation with the students at the Intermediate level. There was no difference in the self-assessment rating accuracy between Years 1 and 2. Overall, most students' OPI gains were reflected in their self-assessment gains. Considering self-assessment's low cost and its overall benefits for students' motivation, agency and goal-setting, the study suggested that self-assessment based on the Can-Do statements can be a valuable tool for low-stake assessment, such as to monitor students' proficiency gains and to globally track the way in which language programs promote proficiency growth. In the next study, Fan (2016) presented an example of how self-assessment may be a valuable tool for low-stakes assessment.

Fan (2016) investigated the validity of a self-assessment scale developed and used for low-stakes placement decisions at a university in China (n=244). The scale was developed and used to crudely gauge students' English proficiency level before the students select from a variety of optional English courses that were made available to them at the university. Because there was a variety of such optional English courses from which the students can select, the placement decisions based on self-assessment results in this context were deemed low-stakes. Results from Rasch analysis indicated that the scale could reliably distinguish students at different proficiency levels. Structural regression analysis revealed that the association between students' self-assessment and their scores on a standardized proficiency test was moderately strong. The evidence generally supported the validity of the self-assessment scale. The study hence reinforced the potential of using a well-crafted and validated self-assessment scale in language learning and assessment, especially relating to low-stakes assessment.

Li and Zhang (2021) conducted a meta-analysis that explored the correlation between self-assessment and language performance. The analysis included 67 studies with 97 independent samples involving more than 68,500 participants. The overall correlation found between self-assessment and an externally administered language measure was 0.446 (p<.01). Given this moderate overall correlation, the study suggested that self-assessment has the potential to be used as a complementary assessment approach to the "external" approach (e.g. language tests and teacher assessment) in measuring language proficiency.

The study also identified six factors that seem to have significant moderating effects, based on which the following suggestions were made to improve the correlation between self-assessment and language performance: (1) The self-assessment criteria should be highly specific; (2) The self-assessment criteria can be broken down into very detailed items that improve learners' understanding; (2) The choice of format, for example computer-assisted adaptive instruments, may help to produced stronger correlations; (3) Pre-assessment training should be carefully designed to enhance learners' familiarity with the criteria and format; (4) The reliability issue of instruments (self-assessment and external measures) should be taken into account as it may affect the correlation.

In summary, based on their empirical findings, the reviewed studies seem to concur that self-assessment might not be a suitable tool for summative purposes and high-stakes contexts, mainly due to limitations in the validity of the instruments studied (Fukazawa, 2011; Summers et al., 2019). However, since students were consistent in their self-evaluations, self-assessment might be used to monitor students' proficiency gains and to globally track the way in which language programs promote proficiency growth (Ma & Winke, 2019; Summers et al., 2019). Self-assessment can also be a cost-effective, valuable tool to support low-stakes placement decisions (Fan, 2016). Most studies saw the potential of self-assessment as an approach that complements other approaches, such as the "external" approach (e.g. language tests and teacher assessment) (Li & Zhang, 2021; Summers et al., 2019). Its potential benefits as a pedagogical tool that promotes learner agency, motivation and autonomy were reiterated throughout all the studies (Fukazawa, 2011). Suggestions were offered to improve the validity and reliability of self-assessment instruments, focusing on the importance of: (1) careful design to improve learners' understanding of each assessment item; (2) rigorous testing to validate the instrument, and (3) pre-assessment training to enhance learners' familiarity with the criteria and format (Fan, 2016; Li & Zhang, 2021).

4 Discrepancies between theory and practice

The review of the CBA literature outlined above suggests discrepancies between theoretical claims of its advantages and the degree to which specific types of CBA activities, when put into practice, succeed in measuring language proficiency. In order to address these discrepancies, I organize them into three categories: reliability, validity, and practicality. In this paper, I use the traditional, time-honoured definition of test validity: "A test is said to be valid if it measures accurately what it is intended to measure" (Hughes, 2007, p. 26). Reliability is used to refer to the "consistency of measurement" (Bachman & Palmer, 1996, p. 23).

I do not employ alternative conceptualizations of validity offered by CBA proponents such as Huerta-Macias (1995, p. 10), who use the term 'credibility' instead of validity and 'auditability' instead of reliability. Huerta-Macias (1995) claimed such assessments are "in and of themselves valid, due to the direct nature of the assessment", and that "consistency is ensured by the auditability of the procedure", "using multiple tasks", "training judges to use clear criteria", and by triangulating any decision-making process with varied sources of data (for example, students, families and teachers" (p. 10). The discrepancies of the CBA approach are summarised below on the basis of the findings of the literature reviewed.

4.1 Reliability issues

As highlighted in the discussion above, the reliability of CBA varies by its sub-type, with portfolio assessment showing the most positive results while self-assessment tends to be identified as the least. Variation is also present across different aspects of assessment (Hirai et al., 2011), and between different groups of learners (Ma & Winke, 2019). Geo-cultural differences may impact the reliability of CBA when applied across different regions or countries.

Type of CBA	Summary of findings from the literature reviewed
Portfolio	Teachers from different countries or regions might have different perceptions of CBA, which may lead to differences in their practices (Davison, 2004). This may affect the reliability of CBA when applied across different geographical contexts.
Peer assessment	Level of reliability varies across different aspects of assessment. In Hirai et al. (2011), among the three assessment criteria of an oral presentation, communicative efficiency was the most consistently rated, while grammar & vocabulary, and pronunciation deviated from teacher assessment.
Self-assessment	Level of reliability varies across students of different proficiency levels. In Ma & Winke (2019), while the correlation was moderate to strong with students at the Novice and Advanced levels of proficiency, there was no correlation with students at the Intermediate level.

Table 4. Reliability issues based on summary of findings from the literature

4.2 Validity issues

CBA data from peer assessment and self-assessment tended to show only moderate to weak correlations with externally administered language measures (Cheng & Warren, 2005; Fan, 2016; Fukazawa, 2011; Hirai et al., 2011; Li & Zhang, 2021; Summers et al., 2019). In the following table (see next page), validity issues are illustrated with findings from the literature reviewed.

Type of CBA	Summary of findings from the literature reviewed
Portfolio	(No evidence found from the studies reviewed in section 3.1)
Peer assessment	Students tended to mark within a narrower range than their class teacher (Cheng & Warren, 2005). Peers tended to be lenient when rating the performances of fellow students, particularly on grammar & vocabulary (Hirai et al., 2011).
Self-assessment	The association between students' self-assessment and an externally administered language measure ranged from moderate (Fan, 2016; Li & Zhang, 2021) to weak (Summers et al., 2019). Regarding the different aspects of validity, Fukazawa (2011) found that the content and substantial aspects were considered to have sufficient validity. However, the structural, consequential, and external aspects of validity were not sufficient.

Table 5. Validity issues based on summary of findings from the literature

4.3 Practicality issues

In theory, CBA aims to allow teachers and students to actively take charge of the assessment process. However, in practice, there are challenges related to the implementation of such approaches, such as grade fixation, fairness, and learner agency (i.e., learners' active role in the writing process) (Lam, 2017). For CBA to be efficient, several conditions need to be met which common classrooms and existing educational structures may not yet be ready to provide (Hirai et al., 2011; Joo, 2016; Saito, 2008; Zhao, 2014). In addition, if the adoption of CBA implies a heavier workload (Coombe & Barlow, 2004; Song & August, 2002), teachers might develop negative views towards its practice (Crusan et al., 2016). These issues pose challenges to the successful implementation of CBA in typical classroom settings.

Type of CBA	Summary of findings from the literature reviewed
Portfolio	The amount of time and efforts it took to complete a portfolio activity may translate to a substantially heavier workload for the teacher (Coombe & Barlow, 2004). Other challenges to the implementation of CBA include the fixation on grades, concerns about fairness, and learner agency (Lam, 2017).
Peer assessment	The efficiency of peer assessment may be dependent on certain conditions which a typical classroom setting or an existing curriculum structure may not be able to accommodate. In Hirai et al. (2011), the anonymity of raters was revealed to be an important factor in improving the reliability of peer assessment. In Saito (2008), the time required to train students may take up a significant part of teaching time (i.e., 200 minutes spanning over five sessions).
Self-assessment	(No evidence found from the studies reviewed in section 3.3)

Table 6. Practicality issues based on summary of findings from the literature

5 Discussion

5.1 Significance of findings

The findings above are drawn from this author's review of theoretical and empirical studies on CBA. As the number of studies reviewed are limited, they certainly do not represent the whole of the CBA research field. Nevertheless, they serve to highlight some important issues about the CBA approach.

Firstly, as innovative or radical as reforms relating to CBA may sometimes be described as, CBA should not be viewed as "somehow magically different" (Brown & Hudson, 1998). As the findings in sections 3 and 4 reveal, the reliability, validity and practicality of CBA procedures vary greatly from one measure to another, being influenced by arrays of possible factors. Hence, for the purpose of responsible assessment and decision-making, it is important that each CBA procedure should be carefully constructed and rigorous tested before it is implemented. It should also be frequently reviewed to detect problems as well as opportunities for improvement.

Secondly, it is important to understand the characteristics of each CBA activity in order to make the most of its use. For example, on peer assessment, Hirai et al. (2011) concluded that, due to the weak correlations between peer and teacher assessment, peer assessment may not always be a reliable substitute for teacher assessment. However, the author suggested that peer assessment may be beneficial for collaborative learning by providing feedback to students. This suggestion was taken up in Sato and Lyster (2012), who investigated whether students can be trained to provide corrective feedback during peer interaction. The reported results were positive. In another effort to make the best use of peer assessment, Zhao (2014) explored whether, with the support of teacher intervention, peer assessment can complement teacher assessment in the facilitation of classroom instruction. The results revealed the potential of peer assessment in such a complementary role.

Thirdly, this paper has focused on analysing the efficiency of CBA procedures in measuring language proficiency. However, as the reviewed literature showed repeatedly, decisions on the use of CBA should not be solely based on this asspect. Instead, considerations should include the positive impact of CBA in other respects, such as its potential for strengthening students' learning and enhancing the teacher's role (Brown & Hudson, 1998), developing higher level cognitive thinking and facilitating a deep approach to learning (Cheng & Warren, 2005), encouraging learner autonomy as well as providing more opportunities for communication in a language classroom (Fukazawa, 2011).

5.2 Teacher assessment identity

Although the discrepancies outlined in section 4 can be attributed to several factors, one central reason is the complexity inherent in teacher assessment identity. This idea is conceptualised as "a dynamic and interactive teacher assessment identity constituted by beliefs, feelings, knowledge and skills" (Looney, et al., 2017, p. 14). The dimensions of teacher assessment identity are illustrated in Figure 1 (see next page), including 'I know', 'I feel', 'I believe', 'I am confident', and 'my role'.

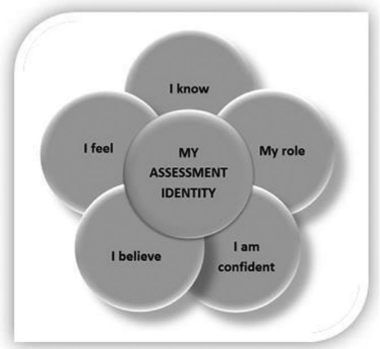

Figure 1. Teacher assessment identity (Looney et al., 2017, p. 15)

Teacher assessment identity accounts for numerous factors involved in the teachers' decision-making process in relation to assessment. For example, although teachers might recognize the usefulness of peer assessment (Saito, 2008), the practicality issues (e.g., time constraints) they experience may discourage them from putting such methods into practice. Such a situation illustrates that, despite possessing knowledge and skills related to assessment practice, teachers might "not be confident in their enactment of such practice" (Looney, et al., 2017, p. 14). Teachers might also "have mixed feelings about assessment" due to socio-psychological factors (Looney et al., 2017, p. 14). This is evidenced by Davison (2004), where a Hong Kong teacher participant had little faith in the CBA system being introduced in his school. The Hong Kong teachers, when faced with the role of assessors, were more concerned with social judgment than with what is required by the professional community (Davison, 2004, p. 322).

5.3 CBA literacy

In recognition of the need to improve teachers and students' CBA capability, I propose a literacy model aimed at enhancing their knowledge and skills in this practice. This model draws from Taylor's (2009) advocacy for providing teachers and their students with professional knowledge and skills for assessment.

Taylor (2013, p. 410) suggested that a teacher who is literature in language assessment should have eight levels of knowledge: knowledge of theory, technical skills, principles and concepts, language pedagogy, sociocultural values, local practices, personal beliefs/attitudes, and scores and decision-making. Although Taylor (2009) was hesitant about applying this eight-level framework as a model for every classroom teacher, it paves the way for the modification and extension of the concept of language assessment literacy.

For a CBA literacy model to be successfully implemented, two factors need to be considered: (1) the local infrastructure where CBA is implemented; (2) teachers assessment identity.

(1) Local infrastructure

Although rater training has been shown to be useful in increasing the efficiency of assessment, such training courses may pose logistical challenges, especially in underprivileged areas. It is therefore important to understand the context in which teachers are situated in order to design a suitable CBA literacy program.

(2) Teacher assessment identity

While teachers have been shown to be generally aware of the contribution of CBA to learning, their implementation of CBA procedures may fail due to a lack of resources. Despite these difficulties, some teachers persist as they believe that students will benefit from self-regulation and self-discipline. As explained above, understanding teacher assessment identity requires acknowledging the role of the teacher as an assessment practitioner, whose practice is informed by the multiple dimensions of beliefs, feelings, knowledge, and skills. It is also crucial to bear in mind that teachers often juggle multiple and demanding roles. As Pryor and Crossouard (2008) observed, "the different identities of the educator as assessor, teacher, subject expert and learner all involve different divisions of labour and rules shaping their interaction with students" (p. 11).

5.4 Conclusion and directions for future CBA research

In this paper, I have reviewed the literature on the theory and practice of CBA in TEFL. In theory, I have described CBA as an evolving concept that has grown out of historical shifts in education policy towards the improvement in student learning. Embedded in dynamic sociocultural contexts, constrained by limited resources, and hampered by uncoordinated action, the concept of CBA intermingles with other educational concepts such as formative assessment and assessment for learning (Davison & Leung, 2009). Yet, these concepts converge on the assumption that assessment and learning are interrelated.

In practice, analyses of CBA procedures revealed the varying degrees to which they succeed in measuring language proficiency. While some procedures, such as the portfolio assessment studied by Song and August (2002), showed a level of reliability and validity that suggested their potential to be used for summative purposes and high-stakes contexts, most other procedures showed moderate to weak results, suggesting their suitability for formative purposes and low-stakes contexts. Overall, findings point to the potential of CBA as a pedagogical tool to support teaching and learning. This use of CBA as a pedagogical tool resonates with the conceptualization of CBA as an instructional activity that is embedded within the process of teaching and learning, which was discussed in the theory section of this paper (section 2.3).

The diversity of perspectives within the field of language assessment does not relegate CBA to a subordinate status compared with other paradigms. Findings and insights from the reviewed recent literature suggest a perpetual belief in the efficiency of CBA in the education process. However, for CBA to become a viable force in education, CBA procedures need to be subjected to the same requirements of responsible testing and evaluation as other types of assessment procedures are. Their reliability, validity and practicality need to be established before conclusions can be drawn on whether a CBA procedure is fit for a specific purpose. One reason is because assessment, regardless of its type, may have important impacts on the decisions made on teaching, learning, and people's lives (Brown & Hudson, 1998; Hamayan, 1995). Another reason for establishing psychometric integrity in CBA procedures is because these types of data are needed to convince policy makers to take action (Song & August, 2002).

A future research agenda in CBA, arguably, should place learning as the central aim within the social processes that involve teachers, learners, administrators, and society, as proposed by Jones et al. (2016). Technological advances in assessment, particularly online assessment, might put CBA in a more complex and interactive network of a globalised ecology. Sophisticated statistical techniques, together with the growth of machine learning, may prove helpful in enhancing the evaluation of learner progress throughout their schooling years (Van

Moere & Hanlon, 2020). Considering the above aspects and adapting Hill and McNamara's (2012) framework, I suggest the following questions for future research:

- What approaches do language teachers adopt when they carry out CBA in multimodal settings (e.g., digital modes of assessment)?
- What specific criteria do teachers focus on when assessing learners?
- What theory or 'standards' do teachers use?
- How do teachers and learners perceive the process of learning, teaching, and assessment?
- What is the role of feedback in CBA?
- How does technology (e.g., machine learning) affect CBA practices and research?
- What are the ethical considerations related to the use of technology in CBA (e.g., artificial intelligence)?

These questions, which aim to facilitate research on CBA processes, emphasize the view of assessment as a part of the process, rather than a measurement of the product, of education.

Acknowledgements: I thank Dr. Gergely A. Dávid for his support and guidance during this research. I also thank Dr. Jamil Toptsi Eftim, Department of English Applied Linguistics, Eötvös Loránd University, Budapest, for proofreading this manuscript.

References

Alderson, J. C., & Banerjee, J. (2001). Language testing and assessment (Part 1). *Language Teaching, 34*(4), 213-236. https://doi.org/10.1017/S0261444800014464

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Damböck, B. (2017). *Language assessment for classroom teachers*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Black, P. & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653-675. https://doi.org/10.2307/3587999

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93-121. https://doi.org/10.1191/0265532205lt298oa

Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics, 20*, 147-161. https://doi.org/10.1017/S0267190500200093

Coombe, C. & Barlow, L. (2004). The reflective portfolio: Two case studies from the UAE. *English Language Teaching Forum, 42*(1), 18-23.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43-56. https://doi.org/10.1016/j.asw.2016.03.001

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), 305-334. https://doi.org/10.1191/0265532204lt286oa

Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly, 4*(1), 37-68. https://doi.org/10.1080/15434300701348359

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

Fan, J. (2016). The construct and predictive validity of a self-assessment scale. *Papers in Language Testing and Assessment, 5*(2), 69-100. https://doi.org/10.58379/jdlz9308

Fukazawa, M. (2009). Validity of peer assessment of speech performance: Using item response theory. *Society for Testing English Proficiency Bulletin, 21*, 31-47.

Fukazawa, M. (2011). Validity of self-assessment of speech performance: A case of Japanese high school students. *JLTA (Japan Language Testing Association) Journal, 14*, 61-79. https://doi.org/10.20622/jltajournal.14.0_61

Giménez, J. C. (1996). Process assessment in ESP: Input, throughput and output. *English for Specific Purposes*, *15*(3), 233-241. https://doi.org/10.1016/0889-4906(96)00007-5

Hamayan, E. V. (1995). Approaches to alternative assessment. *Annual Review of Applied Linguistics*, *15*, 212-226. https://doi.org/10.1017/s0267190500002695

Hamp-Lyons, L., & Con, W. (2000). *Assessing the portfolio principles for practice, theory and research*. Hampton Press.

Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing, 29*(3), 395–420. https://doi.org/10.1177/0265532211428317

Hirai, A., Ito, N., & O'Ki, T. (2011). Applicability of peer assessment for classroom oral performance. *JLTA (Japan Language Testing Association) Journal, 14*, 41-59. https://doi.org/10.20622/jltajournal.14.0_41

Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal, 5*(1), 8-11.

Hughes, A. (2007). *Testing for language teachers.* Cambridge University Press.

Jones, N., Saville, N., & Salamoura, A. (2016). *Learning oriented assessment*. Cambridge University Press.

Joo, S. H. (2016). Self-and peer-assessment of speaking. *Studies in Applied Linguistics and TESOL, 16*(2). https://doi.org/10.7916/salt.v16i2.1257

Lam, R. (2017). Taking stock of portfolio assessment scholarship: From research to practice. *Assessing Writing, 31*, 84-97. https://doi.org/10.1016/j.asw.2016.08.003

Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice, and change. *Language Assessment Quarterly: An International Journal*, *1*(1), 19-41. https://doi.org/10.1207/s15434311laq0101_3

Lewkowicz, J., & Leung, C. (2021). Classroom-based assessment. *Language Teaching, 54*(1), 47-57. https://doi.org/10.1017/s0261444820000506

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189–218. https://doi.org/10.1177/0265532220932481

Looney, A., Cumming, J., Van Der Kleij, F., & Harris, K. (2017). Reconceptualising the role of teachers as assessors: Teacher assessment identity. *Assessment in Education: Principles, Policy & Practice, 25*(5), 1-26. https://doi.org/10.1080/0969594X.2016.1268090

Ma, W., & Winke, P. (2019) Self-assessment: How reliable is it in assessing oral proficiency over time? *Foreign Language Annals. 52*(1), 66–86. https://doi.org/10.1111/flan.12379

Mak, P., & Wong, K. M. (2018). Self-regulation through portfolio assessment in writing classrooms. *ELT Journal, 72*(1), 49–61. https://doi.org/10.1093/elt/ccx012

McNamara, T. (2001). Language assessment as social practice: challenges for research. Language Testing, 18(4), 333-349. https://doi.org/10.1177/026553220101800402

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256. https://doi.org/10.1177/026553229601300302

Nunes, A. (2004). Portfolios in the EFL classroom: Disclosing an informed practice. *ELT Journal, 58*(4), 327-335. https://doi.org/10.1093/elt/58.4.327

Pryor, J., & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment. *Oxford Review of Education, 34*, 1–20. https://doi.org/10.1080/03054980701476386

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching: a description and analysis*. Cambridge University Press.

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

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553–581. https://doi.org/10.1177/0265532208094276

Sato, M., & Lyster, R. (2012). Peer interaction and corrective feedback for accuracy and fluency development: Monitoring, practice, and proceduralization. *Studies in Second Language Acquisition*, *34*(4), 591-626. https://doi.org/10.1017/S0272263112000356

Song, B., & August, B. (2002). Using portfolios to assess the writing of ESL students: A powerful alternative? *Journal of Second Language Writing, 11*(1), 49–72. https://doi.org/10.1016/s1060-3743(02)00053-x

Summers, M. M., Cox, T. L., McMurry, B. L., & Dewey, D. P. (2019). Investigating the use of the ACTFL can-do statements in a self-assessment for student placement in an Intensive English Program. *System*, 80, 269–287. https://doi.org/10.1016/j.system.2018.12.012

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21-36. https://doi.org/10.1017/s0267190509090035

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403-412. https://doi.org/10.1177/0265532213480338

Turner, C. E. (2012). Classroom assessment. In The Routledge handbook of language testing (pp. 79-92). Routledge. https://doi.org/10.4324/9780203181287-12

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of Second Language Assessmen*t (pp. 255-274). De Gruyter Mouton. https://doi.org/10.1515/9781614513827-018

Van Moere, A., & Hanlon, S. (2020). A Bayesian approach to improving measurement precision over multiple test occasions. *Language Testing*, *37*(4), 482-502. https://doi.org/10.1177/0265532220934203

Wicking, P. (2017). The assessment beliefs and practices of English teachers in Japanese universities. *JLTA (Japan Language Testing Association) Journal, 20,* 76-89. https://doi.org/10.20622/jltajournal.20.0_76

Zhao, H. (2014). Investigating teacher-supported peer assessment for EFL writing. *ELT Journal, 68*(2), 155-168. https://doi.org/10.1093/elt/cct068