Studies in Educational Evaluation 74 (2022) 101183

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Are teachers literate in formative assessment? The development and validation of the Teacher Formative Assessment Literacy Scale

Zi Yan [a,*], Serafina Pastore [b]

[a] Department of Curriculum and Instruction, The Education University of Hong Kong, Hong Kong
[b] Department of Research and Humanities Innovation, University of Bari, Italy

ARTICLE INFO

Keywords:
Formative assessment
Formative assessment literacy
Scale development
Validation
Teacher professional development

ABSTRACT

Despite the critical role of formative assessment in instruction, there is a lack of theory-driven instruments that specifically assess teachers' formative assessment literacy. This paper reports the development and validation of the Teacher Formative Assessment Literacy Scale (TFALS). The instrument was developed on a three-dimensional model of formative assessment, aiming at assessing the conceptual, practical, and socio-emotional aspects of formative assessment literacy. Survey data were collected from 585 teachers in Hong Kong primary and secondary schools. Exploratory factor analysis suggests a three-factor structure, and confirmatory factor analysis supports this structure. Rasch analysis results further support its scale dimensionality and item quality. In addition, the relations between TFALS scores and teachers' formative assessment practices demonstrate the instrument's external validity. Overall, the results suggest that the TFALS is an appropriate instrument for assessing teachers' formative assessment literacy. The potential of using the TFALS in research and practice is discussed.

1. Introduction

The emphasis on the use of assessment data represents a crucial aspect in the current discourses on education: data collection and data-use allow educators to 'make many types of decisions from school improvement to classroom and instructional decision-making' (Mandinach & Gummer, 2016, p. 1). Therefore, the need for robust inferences about student knowledge (i.e., information and data on students' learning progression as well as on their misconceptions gathered through discussions, observations, homework, etc.) posits assessment as an integral part of instruction and learning. Moving from the testing tradition (e.g., assessment as functional to produce accurate estimations of student learning progress), formative assessment has been increasingly recognized along with the related strategies teachers can use to support their instructional practice (Andrade et al., 2019; Brown, 2019; Heritage & Wylie, 2020; Wiliam, 2017). While early studies attempted to define the features of formative assessment among various assessment activities (Wiliam & Thompson, 2008), recent research has concentrated on the extent to which teachers actually implement formative assessment, and to what extent formative assessment impacts student learning outcomes (Wylie, 2020). The current study extended this line of research and focused on defining teacher formative assessment literature and developing an instrument to assess it.

The features of assessment literate teachers have been developed progressively, by mapping out the changes teachers need to deal with accountability requirements and to ensure sound instructional decision-making and better learning for students (Herppich et al., 2018; Popham, 2018). Assuming that teachers should use different assessment data to make decisions about guiding students and adjusting teaching (classroom level), developing curriculum and school improvement (school level), and balancing and aligning with the national evaluation system (system level), the concept of assessment literacy has been reshaped and enlarged by scholars over the years. However, some studies (e.g., Boardman & Woodruff, 2004; Brookhart, 2011) have demonstrated that many teachers are reluctant to change their assessment practices or conceptions of assessment (Brown, 2004; Remesal, 2007). While there is an increasing interest in researching formative assessment literacy and designing professional development programs to develop teachers' formative assessment literacy, these attempts have to be supported with a valid and reliable instrument that fits those purposes. Unfortunately, such an instrument, to our knowledge, is currently lacking. Most available assessment literacy instruments are supported by weak psychometric evidence (Gotch & French, 2014), emphasizing summative and standardized assessments, thereby downplaying the formative

* Correspondence to: Department of Curriculum and Instruction, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, N.T., Hong Kong.
E-mail address: zyan@eduhk.hk (Z. Yan).

https://doi.org/10.1016/j.stueduc.2022.101183
Received 9 September 2021; Received in revised form 4 June 2022; Accepted 24 June 2022
Available online 2 July 2022

assessment component of teaching and learning (DeLuca et al., 2016a).

To contribute to this crucial aspect of modern assessment theory and practice, the present study is designed to: (1) develop a self-report instrument to comprehensively assess teachers' formative assessment literacy; and (2) examine the reliability and validity of that instrument. More specifically, in the first section, the paper reviews the most relevant formative assessment research studies and proposes a specific theoretical framework of teacher formative assessment literacy by drawing insights from formative assessment research and theories of assessment literacy. Secondly, it reports the development of the Teacher Formative Assessment Literacy Scale (TFALS) and a validation study carried out with 585 teachers in primary and secondary schools in Hong Kong. Lastly, the paper focuses on the implications for developing formative assessment literate teachers. At the same time, it proposes new research paths in the education assessment domain.

2. Theoretical framework

Two broad areas have guided the development of the conceptual framework for the present study: formative assessment and teacher assessment literacy. Against the backdrop of the existing frameworks of assessment literacy, formative assessment is currently recognized as a relevant aspect. Unfortunately, Bennett (2011) concludes that 'it is doubtful that the average teacher has that knowledge, so most teachers will need substantial time and support to develop it' (p. 20). Thus, to situate the current research project, a framework of formative assessment literacy is then proposed by drawing insights from both areas. The particular focus of this study on formative assessment literacy is, therefore, justified by an awareness of the difficulties teachers have in translating formative assessment theory into practice, and by the need to review teacher education discourses in the light of the importance of assessment literacy.

2.1. Formative assessment

Due to its role in informing instruction and scaffolding student learning, formative assessment has been advocated as a powerful tool for educational reforms (e.g., Assessment Reform Group in the UK, or No Child Left Behind in the US). Furthermore, it has been recognized as a fundamental component in raising student achievement and ensuring innovative classroom environments (Leenknecht et al., 2021; Wylie, 2020). In their seminal work, Black and Wiliam (1998) defined formative assessment as "encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (p. 7). Subsequently, researchers, practitioners, and policy-makers have worked towards a more comprehensive understanding of an articulated set of formative assessment practices to support student learning (Antoniou & James, 2014; Heritage, 2010).

Despite the documented benefits of formative assessment on students' learning improvement, one of the most critical issues is its scarce use by teachers in real classrooms. Too often, teachers struggle to define what students know and the extent of any learning progress. While some studies (e.g., Boström & Palm, 2020; Kennedy, 2016; Smith, 2011) have identified features of effective professional development programs on formative assessment (e.g., active learning, teacher collaboration, professional learning community, alignment between educational policies and practice, appropriate duration of time spent on the program), other studies have pointed out that formative assessment is seldom observed in actual classroom practice (Desimone, 2009; Wylie & Lyon, 2015), or that its implementation is far less than satisfactory (Yan & Cheng, 2015; Yan & Brown, 2021). These studies have also addressed how teacher formative assessment practices are often misaligned with educational policies (e.g., professional standards) or educational principles (Klinger et al., 2012; Stiggins, 2017).

Attempts to reduce and mitigate the difficulties in implementing high-quality formative assessment practices have revealed the assessment illiteracy of both novice and expert teachers. More specifically, recent studies confirmed the difficulties teachers have in understanding the value of assessment, and demonstrated a lack of knowledge and ability in educational assessment (DeLuca, 2012; Leighton et al., 2010; Lysaght & O'Leary, 2013; Schneider & Bodensohn, 2017). Albeit teachers recognize the importance of using assessment evidence, they are not able to manage several sources of information to do that, including formative assessment data. Consequently, teachers do not necessarily possess the assessment literacy, nor the knowledge needed, to use assessment data to inform instruction (Will et al., 2019; Yan & Brown, 2021).

2.2. Teacher formative assessment literacy

In earlier conceptualizations, teacher assessment literacy was recognized as a set of knowledge and skills required to ensure appropriate design, selection, interpretation, and use of assessment for instructional purposes. In this perspective, formative assessment has been progressively considered a pivotal component of teacher assessment literacy (Brookhart, 2011; DeLuca et al., 2016a; Xu & Brown, 2016). However, the prolonged focus on practical aspects of assessment and formative assessment has marginalized other aspects, which are currently recognized as linked to teacher professionalism and to the teacher role as assessor (e.g., social and emotional impacts of formative assessment on student learning; sensitiveness to the ethical aspects of assessment; assessment responsibility). For example, DeLuca and colleagues (2016a), through a thematic analysis of 15 assessment standards and an examination of 8 assessment literacy measures, showed considerable shifts in standards over time. At the same time, they found the persistence of old conceptions of assessment literacy in most measures. The socio-cultural shift addressed by Willis et al. (2013) led to diversely considered assessment literacy. Suggesting a process-oriented conceptual framework, Xu and Brown (2016) developed a hierarchical model of assessment literacy based on six components (teacher knowledge base; conceptions of assessment; institutional and socio-cultural contexts; assessment literacy in practice; teacher learning; and teacher identity as an assessor). Other scholars, instead, have recently stressed the importance of teachers' identities as assessors. In their systematic review of self-reported scales on teacher assessment literacy, Looney et al., (2018) focus on the ethical aspects of assessment practice and consider assessment literacy intertwined with teacher identity. Therefore, teachers' conceptions, beliefs, experiences, and feelings are considered relevant to better understand assessment practice. Integrating processes and practices research approaches, Herppich et al. (2018) have developed a sophisticated framework of professional assessment literacy for teachers and pointed out the need for more integrative research aimed not only to describe or explain but to predict and foster assessment literacy. In the holistic perspective, assessment literacy, made by personal, social, and contextual elements, is considered entangled with teacher identity: teachers can directly manage these elements and work on them if they are perceived as ineffective, difficult, or not sustainable. In this vein, the current attempts to develop a conceptual framework of proficiency progression for assessment literate teachers (Adie et al., 2020) are aligned with the recognition of the need for valid measurement instruments, as well as with the design of more effective and responsive teacher education paths in the assessment domain.

Thus, this research aimed to develop a self-reporting scale on formative assessment literacy, adopting a three-dimension model of teacher assessment literacy (Pastore & Andrade, 2019) as the theoretical framework for the scale development. Aligned with the recent conceptualizations of assessment literacy as a dynamic component of teacher professionalism, this model encompasses an assessment knowledge base, technical skills, personal, social, and contextual elements. With its main dimensions (i.e., Conceptual, Praxeological, and Socio-emotional), the selected model provides for an accurate operationalization of the

domains related to formative assessment and a clear description of its components.

Formative assessment literacy is therefore defined as an interrelated set of knowledge, skills, and dispositions that a teacher can use to design and implement appropriate, context-based assessments with an aim to promote learning and improve teaching. This definition is used to guide the development of the TFALS in the current study. Formative assessment encompasses any activity, formal (e.g., planned, embedded with the curriculum instructional design, and involving the administration of assessment) or informal (unplanned, or "on-the-fly"), that provides information about student learning to be used by teachers and/or students to promote learning towards the learning goals. Teacher formative assessment literacy then has three main constructs. A combination of the three maximizes the potential for teachers to conduct effective formative assessments in the classroom (e.g., documenting student learning, diagnosing and monitoring student learning and enhancing student achievement) and school level (e.g., using school-wide data on student performance to modify, adjust, and differentiate school-policies in order to better align with national curriculum standards).

1. *Conceptual dimension* refers to the content knowledge and guiding principles regarding formative assessment, including its purposes and methods.
2. *Practical dimension* refers to the formative assessment practices a teacher uses to monitor, judge, and manage the teaching and learning process, assuring its soundness and quality, with an aim to promote learning and teaching. Following the Boudieu's theory of practice that conceptualizes practice as a complex mix of professional's habitus, capital, and field, the expression "praxeological component" in the model proposed by Pastore and Andrade (2019) refers to teachers assessment practice. For the sake of legibility, however, we decided to use the expression "practical".
3. *Socio-emotional dimension* refers to a teacher's awareness of the social and emotional aspects of formative assessment. Assuming assessment as a social practice, this dimension includes aspects relevant to formative assessment that are beyond academic learning (e.g., emotion, motivation, and well-being). At the same time, this dimension involves teachers working with different stakeholders to create a shared sense-making of assessment practices and enhance assessment systems in the service of student learning.

Modelling this comprehensive conceptualization of assessment literacy, the TFALS is based on a dynamic model of formative assessment. Thus, the instrument development aims to effectively integrate a formative assessment rationale, applicable to a wide range of assessment contexts, as well as strategies and methods relevant to teachers' daily practice in the classroom.

3. Instruments focusing on teacher formative assessment literacy

To our knowledge, there is no available theory-driven instrument specifically assessing teachers' formative assessment literacy. Although a number of instruments have been developed to assess teachers' assessment literacy either in a specific domain (e.g., Fulcher, 2012; Soh & Zhang, 2017 for language teachers) or in general (see Gotch & French, 2014, for a review), none of them was designed explicitly for *formative* assessment literacy. This is probably because those instruments were based on theoretical frameworks or standards that did not highlight formative assessment. With the aim of supporting teachers to effectively implement formative assessment, some instruments such as observation protocols and rubrics have been designed to account for formative assessment practices observed "in vivo" in the classroom (e.g., Cagasanned et al., 2020; Wylie, 2020), which represents only one aspect of formative assessment literacy.

In their systematic review of 36 assessment literacy instruments, Gotch and French (2014) concluded that "the psychometric evidence available to support assessment literacy measures is weak" even though evaluation of teaching effectiveness - including assessment literacy - was at the top of the education agenda. Some of the most known and used teacher assessment literacy instruments existing in the literature are reported in the following:

- The Measurement Competency Test (MCT) of Mayo (1967). Four dimensions and 60 items articulated in terms of importance for teachers;
- The Teacher Assessment Literacy Questionnaire (TALQ) of Plake et al., (1993). This 35 items questionnaire is strictly linked to *Seven Standards of Teacher Competencies in Educational Assessment of Students* (AFT, NCME, & NEA, 1990);
- The Assessment Practice Inventory (API) of Zhang and Burry-Stock (1997). In this inventory 67 items are related to assessment practice;
- The Assessment Literacy Inventory (ALI) of Campbell et al., (2002) consists of 5 different scenarios linked to the professional standards of 1990. For each scenario there are 7 questions; and
- The Classroom Assessment Literacy Inventory (CALI) of Mertler (2003), based on the previous version of the ALI (above) and further revised in 2005 (Mertler & Campbell, 2005).

More recently, DeLuca et al. (2016a) have provided a comprehensive review of assessment literacy standards and assessment literacy instruments developed after 1990. They found that assessment literacy standards before 2000 focused on summative and standardized assessments, and emphasized teachers' psychometric understandings. Since 2000, however, formative assessment has emerged as a new theme and has become a more dominant theme in modern assessment standards since 2010. They analyzed eight assessment literacy instruments developed between 1993 and 2012 and concluded that the reviewed instruments were mostly based on the early conceptions of assessment literacy, especially the 1990 Standards for Teacher Competence in the Educational Assessment of Students (AFT et al., 1990). DeLuca et al. (2016a) concluded, echoing Brookhart (2011) criticism, that these assessment literacy instruments failed to incorporate the significant shifts in assessment standards over time, especially those relating to formative assessment practices and standards-based education.

In order to construct an instrument reflective of contemporary assessment practices and contexts, DeLuca et al., (2016b) developed the Approaches to Classroom Assessment Instrument (ACAI) according to the Classroom Assessment Standards (Joint Committee on Standards for Educational Evaluation [JCSEE], 2015). The ACAI is a comprehensive but relatively lengthy survey which might limit its utility for large-scale studies. Furthermore, the ACAI can provide an overall profile of teachers' "approach to classroom assessment", but not specific to formative assessment per se. All three parts of the ACAI cover elements related to formative assessment, but there is no subscale focused on teachers' formative assessment literacy, a significant drawback if that were the purpose of adopting the survey.

To address the methodological gap, we intend to develop an instrument specifically focused on assessing teachers' formative assessment literacy. This instrument would acknowledge the identified subcomponents of formative assessment literacy, as discussed earlier, and that internal structure would enable a more nuanced understanding of the distinct ways in which a teacher might be literate in formative assessment (Gotch & French, 2014). This instrument would then be expected to assist teachers' classroom assessment practices and professional development programs.

4. The present study

Based on the proposed formative assessment literacy framework, this study aims to: (1) develop a self-report instrument to comprehensively assess teachers' formative assessment literacy; and (2) examine the

reliability and validity of that instrument based on a sample of 585 Hong Kong teachers. More specifically, the current study examined the content, substantial, structural, generalizability, and external validity of the instrument (Messick, 1995). In particular, content validity was ensured by the theory-driven procedure of scale development and expert review, as described above. The substantive validity was ensured by the theory-driven procedure of scale development and the guiding theoretical framework, i.e., the Pastore & Andrade, 2019 three-dimension framework of assessment literacy. The structural validity was investigated by exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The generalizability validity was inferred from the results of the DIF analyses across gender and educational level (primary vs. secondary). Given the close relationship between teachers' assessment literacy and their assessment practices, the external validity was examined by the correlations between the TFALS subscales and the Teacher Formative Assessment Practice Scale (TFAPS), a self-reported scale on formative assessment practice (Yan & Pastore, 2022).

5. Method

5.1. Participants

Data were collected from a convenience sample of 585 teachers from 14 Hong Kong schools (9 primary and 5 secondary schools). The sample comprised 339 (57.9%) primary school teachers and 246 (42.1%) secondary school teachers. There were 267 (45.6%) female, 122 (20.9%) male teachers, and 196 (33.5%) without gender information as their data could not be matched with the 1st wave data that had the gender information in the larger project. The average teaching experience was 14.3 years (SD = 10.1), ranging from 1 to 38 years.

5.2. Procedure

The data used in this study were the 2nd wave of data collected in a larger project about formative assessment. Ethics approval was sought and given by the authors' affiliated University. After obtaining the official consent from the principal of each participating school, the research team approached teachers in participating schools and asked for their informed consent to participate. Each school had a coordinating teacher who collected the completed questionnaires and returned them back to the research team. All participants were informed about the purpose of the study, the confidentiality of their data, and their right to withdraw at any time.

5.3. Instruments

5.3.1. The development of the teacher formative assessment literacy scale (TFALS)

As discussed earlier, we adopted a deductive approach (Hinkin, 1995) to develop the TFALS based on the framework of formative assessment literacy, which drew insights from formative assessment research and the assessment literacy theoretical model introduced by Pastore & Andrade, 2019. The TFALS is intended to provide a holistic indication of teachers' formative assessment literacy covering three dimensions: Conceptual, Practical, and Socio-emotional. After consulting the literature regarding formative assessment and focus group interviews with ten Hong Kong teachers (5 primary and 5 secondary school teachers), we generated an initial item pool of 32 items to cover all three dimensions. The initial item pool was then sent to a panel of twelve experts to examine the content validity of the instrument. Nine of these experts provided responses. All experts are experienced researchers and practitioners in the field of educational assessment, and many also have experience in front-line teaching or teacher training. All experts were asked to (a) sort the 32 items (presented in random order) according to the three dimensions, and (b) rate the relevance of each item to teacher formative assessment literacy on a scale ranging from 'essential', 'useful, but not essential', to 'not necessary'. Based on the item sort, the proportion of substantive agreement (i.e., the proportion of experts who assign an item to its intended dimension) was calculated for each item (Anderson & Gerbing, 1991). A proportion of.7 or above was adopted as an indication of substantive agreement. As for the item relevance, the content validity ratios (CVRs; Lawshe, 1975) were computed, and positive CVR (i.e., five or more experts indicating essential) indicated acceptable content validity. The review results showed that five items had both an unsatisfactory substantive agreement and a negative CVR. These items were considered and removed. Three other items demonstrated unsatisfactory substantive agreement only, while another two showed negative CVR only. After discussion within the research team as well as getting more feedback from the review experts, we concluded that the misclassifications of these items were due largely to awkward wording. The items were therefore revised and retained after consultation. Comments were also solicited from the review experts regarding the readability of the items, and modifications were made accordingly. The resultant instrument had 27 items: 8 items representing the Conceptual dimension, 11 items in Practical, and 8 items for Socio-emotional. A six-point Likert-type response scale (1 = Strongly disagree, 2 = Disagree, 3 = Slightly disagree, 4 = Slightly agree, 5 = Agree, 6 = Strongly agree) was adopted for all items. All items were initially developed in English and then were translated into Chinese following the translate-back translate procedure (International Test Commission, 2017).

A pilot study using the Chinese version of the instrument was conducted with 26 volunteer Hong Kong teachers (14 primary and 12 secondary school teachers). They were asked to complete the questionnaire and provide qualitative feedback regarding the clarity and readability of items. Some minor revisions were then made according to the teachers' feedback.

5.3.2. The teacher formative assessment practice scale (TFAPS)

To provide evidence of the external validity of the TFALS, we examined the relations between teachers' formative assessment literacy and their self-reported formative assessment practices. Teachers' actual formative assessment practices were assessed with a 10-item Teacher Formative Assessment Practice Scale (TFAPS) (Yan & Pastore, 2022). The TFAPS was developed following the Wiliam and Thompson (2008) theoretical framework of formative assessment. The scale has two subscales: teacher-directed formative assessment (TdFA; 6 items) focuses on formative assessment practices usually dominated by teachers, such as sharing the learning goals and providing feedback; while student-directed formative assessment (SdFA; 4 items) targets at formative assessment practices over which students have more control, such as self-assessment and peer-assessment. A six-point Likert-type response scale (1 = Never, 2 = Rarely, 3 = Seldom, 4 = Sometimes, 5 = Frequently, 6 = Very Frequently) was adopted for all items. The confirmatory factor analysis on the two-factor solution using the sample of this study showed a satisfactory fit: $\chi^2/df = 2.153$; RMSEA = .065; SRMR = .049; CFI = .968; and TLI = .955. The TFAPS demonstrated satisfactory reliability. Cronbach's alpha for the two subscales were 0.79 and 0.86, respectively. The Rasch reliabilities for the two subscales were 0.79 and 0.83, respectively.

5.4. Data analyses

We followed the 5% rule on missing data (Tabachnick et al., 2007). Participants with more than 5% of item-level missing data ($N = 4$) were excluded, and the remaining responses with item-level missing data were imputed using maximum likelihood methods with the expectation-maximization algorithm (Allison, 2003).

The sample was randomly split into two sub-samples. The first sample was used for EFA and the second was used for CFA. The first subsample included 169 (57.5%) primary school teachers and 125 (42.5%) secondary. There were 127 (43.2%) female, 67 (22.8%) male

teachers, and 100 (34.0%) without gender information. The average teaching experience was 15.1 years (SD = 9.96), ranging from 1 to 36 years.

The second subsample consisted of 167 (58.2%) primary school teachers and 120 (41.8%) secondary school teachers. There were 137 (47.7%) female, 54 (18.8%) male teachers, and 96 (33.4%) without gender information. The average teaching experience was 13.57 years (SD = 10.13), ranging from 1 to 38 years. In order to ensure the comparability of these two subsamples, we compared the distribution of gender and educational level between them. Chi-square test results showed that there were no significant differences regarding the distribution of gender (χ^2 = 1.752, p = 0.186), nor the distribution of educational level (χ^2 = 0.030, p = 0.863) between these two subsamples. A *t*-test also indicated that no significant differences were found for teaching experience (t [376] = 1.477, p = 0.140). Thus, the two subsamples are not significantly different in terms of participants' gender, educational level, and teaching experience.

EFA, using SPSS version 26 with maximum likelihood method and promax rotation, was conducted on the data from the first subsample to explore the factor structure underlying the items. The eigenvalue test, scree plots, and interpretability of the rotated factors were observed to determine the number of factors in the data matrix. Items were retained when the factor loading was higher than 0.4, without cross-loading above 0.3.

CFA using AMOS version 26 was conducted on the data from the second subsample to examine the fit between those empirical data and the factor structure derived from the EFA results. The model-data fit was evaluated by multiple fit indices, including the comparative fit index (CFI), the Tucker-Lewis index (TLI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA), which minimize Type I and Type II errors under many conditions (Hu & Bentler, 1999). We also checked the χ^2/df ratio. Values of CFI and TLI higher than 0.90, SRMR and RMSEA lower than 0.08, and χ^2/df ratio less than 3 (Fan & Sivo, 2007; Hu & Bentler, 1999; McDonald & Ho, 2002) were considered acceptable.

To provide more comprehensive validity evidence, Rasch analysis was applied to the complete data set. Rasch analysis has been widely applied to examine scale quality (e.g., van der Lans et al., 2018; Coniam & Yan, 2016). The Rasch model adopts a "data fit the model" approach and examines the extent to which items that form a scale reflect a single underlying latent construct (Andrich, 2004; Bond et al., 2020). A multidimensional Rasch model (Adams et al., 1997) was employed with these data because the proposed three dimensions in formative assessment literacy are inter-correlated. The advantage of the multidimensional Rasch model, compared with the conventional unidimensional Rasch model, is that it calibrates all subscales simultaneously so that the correlations between the subscales could be considered to increase measurement precision on each subscale. Multiple indicators were used to examine the scale quality. *Response category functioning* shows whether response options function well in parallel with the underlying variable. *Item fit statistics* (i.e., Infit MNSQ and Outfit MNSQ) reflect the extent to which each item in a subscale measures a unidimensional latent construct. Values of Infit/Outfit MNSQ between 0.75 and 1.33 indicate sufficient fit to the Rasch model (Wilson, 2005). *Differential item functioning* (DIF) across gender and educational level was checked by calculating the difference of item difficulty across groups after controlling the levels of the latent trait. A difference equal to or greater than 0.5 logits indicates the existence of substantive DIF (Wang et al., 2006).

Furthermore, we calculated the correlations between the TFALS and teachers' self-reported formative assessment practices to examine the external validity of the TFALS. With regard to the reliability of the TFALS, we assessed both Cronbach's alpha and Rasch reliability for each subscale.

6. Results

6.1. Exploratory factor analysis

An initial EFA with all items on the first subsample (n = 294) was conducted. The value for Kaiser–Meyer–Olkin measure of sampling adequacy (0.95) was well above the recommended value of 0.60, and Bartlett's test of sphericity was also significant (χ^2 (351) = 4533.375, p < .001), indicating that the data were suitable for factor analysis. The EFA results identified three factors with eigenvalues higher than 1. The scree test (Cattell, 1966) also supported the three-factor solution. Four items appeared problematic as they showed cross-loading (Item #6: *I understand that formative assessment can work only if students take action on assessment feedback information*; Item #9: *I review my teaching practices according to assessment results;* Item #11: *I align my assessment activities with learning goals*; and Item #20: *I am aware of my power as assessor to influence students' reactions to formative assessment*). These four items were removed and a second EFA also indicated a three-factor solution. The factor loadings of the remaining 23 items can be found in Table S1 of the Supplementary Materials. The three factors were in line with the theoretical proposal and could be meaningfully interpreted. The Conceptual dimension is about the content knowledge and guiding principles regarding formative assessment. The Practical dimension refers to formative assessment practices a teacher uses to promote learning and teaching. Lastly, the Socio-emotional dimension gauges a teacher's awareness of the social and emotional aspects of formative assessment.

6.2. Confirmatory factor analysis

A CFA with maximum likelihood estimation was applied to the data from the second subsample (n = 287). The value for Kaiser–Meyer–Olkin measure of sampling adequacy (0.95) and the result of Bartlett's test of sphericity (χ^2 (351) = 4593.635, p < .001) indicated that the data were suitable for factor analysis. The results of the CFA showed that some fit statistics were satisfactory: χ^2/df = 2.642; RMSEA = .076; and SRMR = .061; while others were merely marginal: CFI = .897; TLI = .886. We removed Item #19 based on the Rasch analysis results, (presented in the following section) and re-ran the CFA on the remaining 22 items. The fit statistics were slightly improved: χ^2/df = 2.714; RMSEA = .077; SRMR = .061; CFI = .900; and TLI = .888. Taking into account the modification indices and conceptual considerations, the residuals of two pairs of items (Items #1 and #2 under the Conceptual factor; Items #12 and #13 under the Practical factor) were allowed to correlate. The CFA was conducted again and the results demonstrated a satisfactory fit between the hypothesized model and the observed data: χ^2/df = 2.370; RMSEA = .069; SRMR = .058; CFI = .921; and TLI = .910. Fig. 1 displays the standardized factor loadings of items and correlations between factors.

6.3. Rasch analysis

A multidimensional Rasch analysis was applied to the 23 items, retained in the EFA, on the complete data set (n = 581). The results identified one misfitting item (Item #19: I communicate assessment results to other people involved (e. g., colleagues, parents) to support student learning). Its Infit and Outfit MNSQ were 1.42 and 1.52 respectively, which were out of the acceptable range of 0.75–1.33 (Wilson, 2005). Furthermore, this item also demonstrated substantial DIF across the educational level (the item difficulty for primary school teachers was 0.9 logits lower than that for secondary school teachers), indicating that it was much easier for primary school teachers to endorse this item than their peers in secondary schools. We, therefore, removed this item and re-ran the Rasch analysis on the remaining 22 items. All items fitted quite well to the Rasch model except Item #18 (I share assessment criteria with students; Infit MNSQ: 1.38; Outfit MNSQ: 1.39). Since the misfit of Item #18 was marginal and this is the first validation of the TFALS, we tended to keep it. No items demonstrated substantial

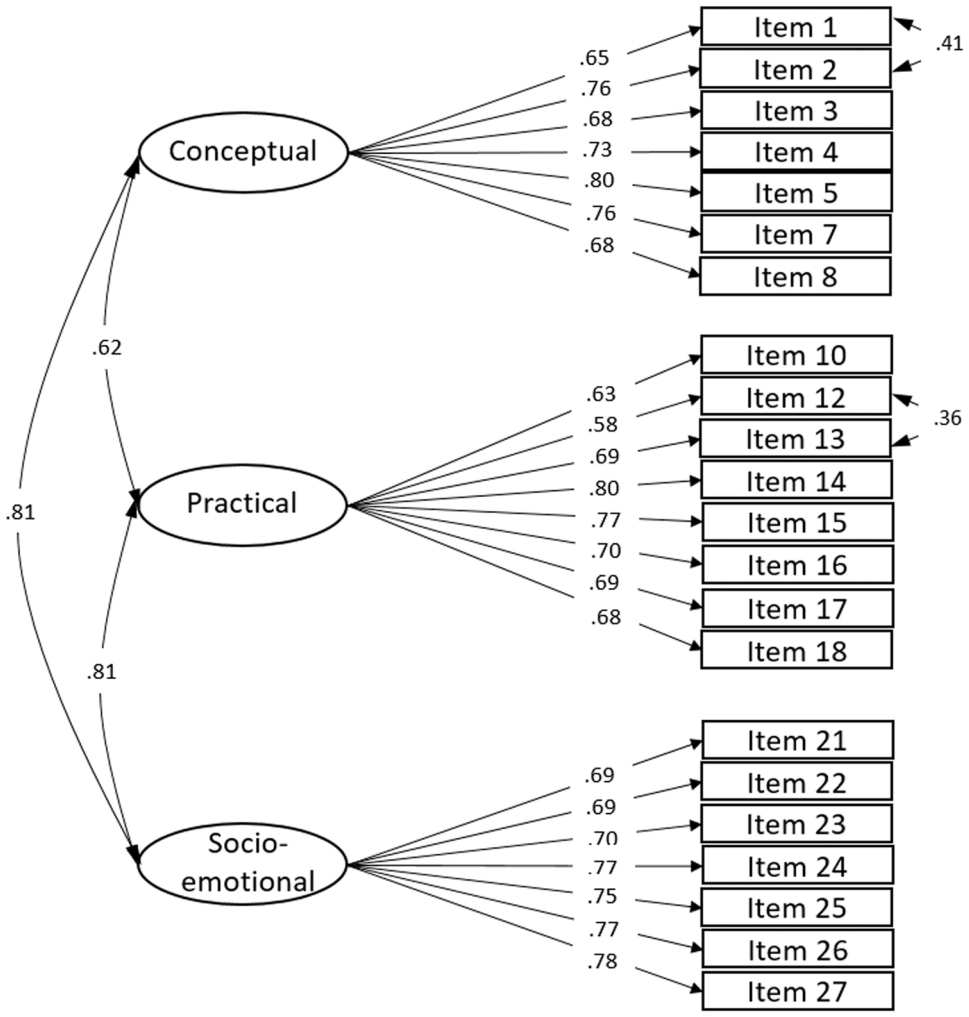

Fig. 1. Standardized factor loadings of items and correlations between factors in the confirmatory factor analysis.

DIF across gender and educational level. The step calibrations (the measures of the transition points between adjacent categories) of the six-point response scale increased monotonically from − 6.50, − 2.40, − 0.95, 2.53, to 7.32 logits, indicating that the response scale functioned well (Linacre, 2002). The item difficulty, standard error and item fit statistics for the retained 22 items are presented in Table 1. The final version of TFALS is provided in the Appendix.

Fig. 2 presents the item-person map where person measures and item difficulties are displayed on the same metric. The three continua on the left side represent person measures on the three TFALS subscales. The items and the associated thresholds, indicated with the notion of *x.y*, are placed on the right side. For instance, 1.5 refers to the fifth threshold of Item #1. It can be seen that the TFALS items and their thresholds cover a wide range of latent trait that is aligned with the teachers' formative assessment literacy.

6.4. Correlations with teachers' formative assessment practices

To examine the external aspect of validity of the TFALS, the Pearson correlations between Rasch-calibrated person measures on the three subscales of the TFALS and those on the two subscales of the TFAPS were calculated. On the one hand, these two scales measure two different concepts. On the other hand, teachers' assessment literacy is supposed to correlate positively with their assessment practices. Thus, we expected moderate, rather than low or high, correlations between these two concepts. As shown in Table 2, all correlations are significant. The

Table 1
Item Difficulty, Standard Error, Item Fit Statistics for the 22-item TFALS.

Item No.	Item Measure*	SE	Infit MNSQ	Outfit MNSQ
Conceptual Dimension				
Item 1	1.01	0.06	1.06	1.05
Item 2	-0.02	0.06	0.83	0.74
Item 3	-0.87	0.07	1.11	1.06
Item 4	0.51	0.06	0.93	0.93
Item 5	-0.24	0.06	0.89	0.85
Item 7	-0.09	0.06	1.11	1.10
Item 8	-0.32	0.16	1.21	1.21
Practical Dimension				
Item 10	-0.27	0.06	1.06	1.05
Item 12	0.99	0.06	1.23	1.30
Item 13	0.96	0.06	1.04	1.04
Item 14	0.06	0.06	0.95	0.89
Item 15	-0.62	0.06	0.87	0.80
Item 16	-0.05	0.06	0.91	0.89
Item 17	-0.32	0.06	1.07	1.08
Item 18	-0.75	0.16	1.38	1.39
Socio-emotional Dimension				
Item 21	0.10	0.06	0.91	0.86
Item 22	-0.04	0.06	1.10	1.09
Item 23	-0.20	0.06	1.06	1.04
Item 24	-0.20	0.06	0.86	0.85
Item 25	0.58	0.06	1.16	1.16
Item 26	0.09	0.06	0.87	0.84
Item 27	-0.33	0.15	0.80	0.76

Note. *All measures are in logits.

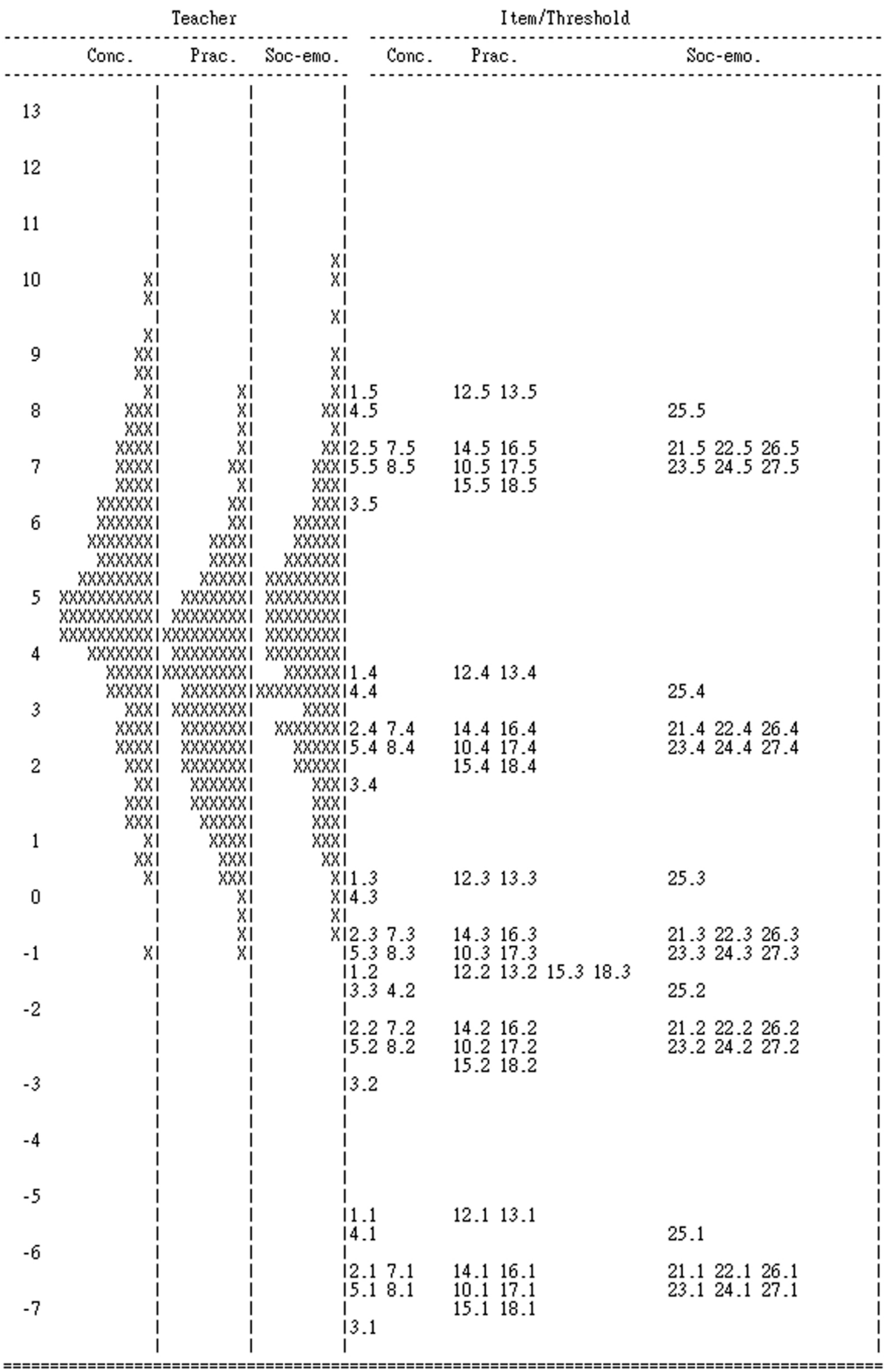

Fig. 2. The item-person map of the TFALS.

correlations between teachers' performance on the three subscales of the TFALS ranged from 0.524 to 0.665. All three subscales of the TFALS moderately correlated to TdFA, *r* ranging from 0.385 to 0.515. The correlations between each of the three subscales of the TFALS and SdFA were weak to moderate, *r* ranging from 0.164 to 0.409.

6.5. Reliability

Reliability analysis was performed for the final 22-item TFALS on the complete data set (n = 581). Cronbach's alpha for the Conceptual, Practical, and Socio-emotional subscales of the TFALS were 0.88, 0.88,

Table 2
Correlations between TFALS and TFAPS.

	Conceptual	Practical	Socio-emotional	TdFA	SdFA
Conceptual	-				
Practical	.524**	-			
Socio-emotional	.635**	.665**	-		
TdFA	.385**	.515**	.438**	-	
SdFA	.164**	.409**	.215**	.446**	-
Mean	4.59	3.36	4.03	3.33	0.71
SD.	1.91	2.02	2.05	1.66	2.16

Note. * * $p < .01$.

and 0.89, respectively, indicating good internal consistency. The Rasch analysis also showed good reliabilities (i.e. EAP/PV reliabilities) for the three subscales (Conceptual: 0.86, Practical: 0.89, and Socio-emotional: 0.90), corresponding to the separation index of 2.48, 2.83, and 3.00.

7. Discussion

The current study aimed to develop and validate an instrument for assessing teachers' formative assessment literacy. The EFA on the first half sample identified a three-factor solution, in line with the theoretical framework, and was confirmed by the CFA on the second half sample. Rasch analysis provided further support to the dimensionality of the three subscales, the functioning of the rating scale, item fit statistics, and invariant item measures across gender and educational level. All subscales had satisfactory Cronbach's alpha coefficients and Rasch reliabilities. The expected correlations with teachers' formative assessment practices supported the external aspect of validity of the TFALS. The final version of the TFALS has 22 items accessing three dimensions of formative assessment literacy, i.e., Conceptual dimension (7 items), Practical dimension (8 items), and Socio-emotional dimension (7 items).

The development of the TFALS addresses an important methodological gap in the research and practice of formative assessment. Improving teacher assessment literacy is one of the most pressing and contested contemporary educational policy and practice issues. There is compelling evidence that assessment, more specifically formative assessment, represents a key leverage point for improving student outcomes (Andrade & Heritage, 2018; De Simone, 2020). The efforts made by policy-makers, teacher educators, and administrators clearly show that the intersection of teacher education and assessment practice matters in fostering the quality of instruction in national school systems (Darling-Hammond et al., 2017; Desimone, 2009; Kennedy, 2019; Stiggins, 2017). However, the increased importance accorded to formative assessment is not linked to effective professional development paths that can encourage a systemic review of teacher assessment practices (DeLuca & Volante, 2016; Will et al., 2019). Teachers struggle to transfer what they have learnt through professional development into the real classroom context. Therefore, it is fundamental to support teachers to clearly recognize how to transfer their learning and how to better carry out formative assessment in the classroom.

In this vein, the development of the TFALS should be informative in terms of teacher education and teacher practice. Against the backdrop of teacher professionalism, the scale, in fact, can provide data on teachers' profiles (strengths and weaknesses) in formative assessment, which could be used to support teachers in developing their formative assessment literacy as well as enable purposefully designed teacher training. The instrument, indeed, could be used to gauge the needs for professional development and evaluate the effectiveness of teacher training programs or interventions for enhancing teachers' formative assessment literacy. In addition, compared to an overall measure, the three-dimension model of formative assessment literacy underpinning the TFALS enables a nuanced diagnosis of training needs and evaluation of the training programs. Previous studies have shown that continuing development and learning is a critical lever for the effectiveness of policy for teachers and teaching practice and improving student achievement (Desimone, 2009; Gore et al., 2021). In the formative assessment domain, instead, an unclear definition of this concept and a lack of valid instruments for measuring this concept have, for a long time, negatively impacted teacher development programs and teacher assessment practices. Now, educational research, closely linked to professional development models, must demonstrate that formative assessment literate teachers can make a difference. In this perspective, more research is needed in order to identify scalable models, strategies, and instruments supporting teachers for improved formative assessment practice. The TFALS clearly intersects this research stream.

The results showed that teacher performance on the TFALS was high: the Rasch-calibrated person measures ranged from 3.36 to 4.59 (the mean item difficulty was set to 0 so that person measures higher than zero indicate a positive response, while person measures lower than zero indicate a negative response), corresponding to mean raw scores from 4.64 to 4.97 on a 6-point rating scale, indicating a high level of formative assessment literacy. On the one hand, this finding is not surprising for Hong Kong teachers considering that formative assessment has been promoted in the Hong Kong education system for two decades, and many teacher training programs have been provided to in-service teachers aiming to enhance their formative assessment literacy (Yan & Brown, 2021). On the other hand, this finding is alerting when we realize that the implementation of formative assessment is not satisfactory in practice in both the Hong Kong context (Lam, 2018; 2019) and other contexts (Desimone, 2009; Wylie & Lyon, 2015). Such a literacy-practice gap is probably due to the fact that, as review studies (e.g., Fulmer et al., 2015; Heitink et al., 2016; Yan et al., 2021) consistently pointed out, the implementation of formative assessment is not only influenced by teachers' personal factors, but also constrained by contextual factors, such as internal school support, working conditions, student characteristics, external policy, and cultural norms. For instance, in another study of Hong Kong teachers' formative assessment practices, Yan and Cheng (2015) found that the teacher reported frequency of formative assessment was low even though they had relatively high levels of self-efficacy to conduct formative assessment. The authors speculated that teachers' formative assessment practices were susceptible to contextual factors, in addition to personal factors, such as large class sizes and heavy workload. This finding also echoes Yan's (2021) argument that teacher training might enhance teachers' knowledge and skills for new assessment methods, but might not substantially increase its implementation in practice if there is a lack of systematic support, such as curriculum re-design and individual student support. Future studies can examine whether teachers' high level of formative assessment literacy and a similar literacy-practice gap exists in other teaching contexts.

The correlation analysis showed a clear relationship between practical and conceptual dimensions of the TFALS. Previous studies have repeatedly identified theory and practice (or knowledge and skills) as the main pillars of teacher assessment literacy (Brookhart, 2011; DeLuca & Volante, 2016; Popham, 2018). Interestingly these two dimensions result also as constituents of formative assessment literacy. For years educational researchers and teacher educators have demonstrated a strong commitment to helping teachers understand what formative assessment is and how to practise it in the classroom. A relevant part of studies has constantly pointed out that the improvement of teacher assessment literacy is related to the curriculum of teacher education courses and the quality of professional development paths. In this vein, the interplay of the conceptual and practical dimensions of formative assessment continues to be a "must" for reviewing and re-designing teacher education. The socio-emotional dimension of TFALS also correlated with the conceptual and practical ones, indicating that the social and cultural nature of formative assessment is recognized not only at the practical level but also in theory. In other words, within the framework of teacher formative assessment literacy, some aspects such as values, beliefs, attitudes, power dynamics, or student well-being are

conceived by teachers as relevant knowledge and as a fundamental basis for effective practice. This aspect is clearly aligned with the current attempts to define the general concept of teacher assessment literacy.

The correlations between the three TFALS subscales and teachers' self-reported formative assessment practices were of the expected size and direction, providing additional validity evidence. The relationship between teacher-directed formative assessment and the Practical dimension of TFALS is, not surprisingly, the highest in value, highlighting that teachers need to identify the constituents of formative assessment and think about what they need in terms of methods, strategies, and instruments to effectively practice it in the classroom. The results suggest that these two concepts (formative assessment literacy and practice) are related, but distinct, constructs. The links between TFALS and the student-directed formative assessment (i.e., self-assessment and peer-assessment) were much lower than their correlations with teacher-directed formative assessment. Also, the Rasch-calibrated person measure on student-directed formative assessment was only 0.71, corresponding to a mean raw score of 3.96 on a 6-point rating scale. This finding implies that, although student-directed formative assessment has been integrated into the definition of formative assessment (see Andrade & Cizek, 2010; Black & Wiliam, 1998) and recognized as a regular type of formative assessment in academia (e.g., Black & Wiliam, 2009), in practice teachers implement it less frequently and its implementation is less influenced by teachers' formative assessment literacy. A speculative explanation is that the implementation of student-directed formative assessment is hindered by the lack of appropriate competency in students to conduct formative assessment for themselves, or by the lack of teacher confidence in them to do so.

As a newly developed instrument, more studies are encouraged to further test its validity and utility. First, this study used a sample of primary and secondary school teachers from a Confucian culture. As formative assessment is likely to be influenced by the teaching context (Heitink et al., 2016; Yan et al., 2021), future studies may consider validating the scale among teachers from different teaching environments (e.g., higher education or kindergarten) or from other cultures. Second, it is worth noting that the implementation of formative assessment may vary across disciplines. The formative assessment strategies used by a mathematic teacher may be different from that of a language teacher, although the principle remains the same. The TFALS aims to capture the essence of formative assessment independent of disciplines to ensure its generalizability. Still, it is open to adjustments, if necessary, in future studies to cater for discipline-based needs. Third, the users of the TFALS are advised to attend to the potential response bias. For example, respondents may overestimate their formative assessment literacy either because of a lack of metacognitive ability to recognize their deficiencies (Duning-Kruger effect), or because of an intention to be viewed favourably by others (social desirability bias). Fourth, the consequential aspect of validity (see Messick, 1995) could not be examined in the current study. Future studies could usefully explore whether and how the use of the TFALS leads to important outcomes such as adaptive teaching or better student learning outcomes.

8. Conclusions

With a recognition of the crucial role of teachers' formative assessment literacy, the present study aimed to develop and validate an instrument (TFALS) for assessing the conceptual, practical, and socio-emotional aspects of teachers' formative assessment literacy. The results of both factor analysis and Rasch analysis demonstrated satisfactory psychometric properties of this instrument for use with primary and secondary school teachers. The development of the TFALS addresses an important methodological gap in the research and practice of formative assessment. The information provided by this instrument can be used to describe teachers' profile of formative assessment literacy, inform the design of teacher education by identifying the training needs and evaluate the effectiveness of teacher training programs or interventions for enhancing teachers' formative assessment literacy.

Acknowledgements

The first author was supported by a General Research Fund (GRF) (Project No: EDUHK18607118) from the Research Grants Council of Hong Kong. We also wish to thank (in alphabetical order) Prof. David Boud, Prof. Phillip Dawson, Dr. Ricky Lam, Prof. Anastasiya Lipnevich, Dr. Ernesto Panadero, Dr. Kim Schildkamp, Dr. Joanna Tai, Dr. Shirley Xiao, and Prof. Yueting Xu for their invaluable feedback on the original items.

Appendix

Teacher Formative Assessment Literacy Scale.

Conceptual Dimension	
Item 1	I can explain the rationale for formative assessment.
Item 2	I know that students' learning needs can be identified through formative assessment.
Item 3	I think assessment activities should be aligned with learning goals.
Item 4	I understand that formative assessment tasks should elicit evidence about students' learning.
Item 5	I know that formative assessment results are useful for teachers to cater for student learning needs.
Item 6	I think students should be engaged in the formative assessment in order to promote learning
Item 7	I know diverse assessment methods that allow students to demonstrate their learning.
Practical Dimension	
Item 8	I use a variety of assessment methods that allow students to demonstrate their learning.
Item 9	I teach students to engage in peer feedback processes.
Item 10	I help students to develop self-assessment skills.
Item 11	I engage students in using feedback information in subsequent tasks.
Item 12	Based on assessment results, I show students what they need to do in order to improve their learning.
Item 13	I train students to act on assessment feedback information to improve their learning.
Item 14	I clarify assessment purposes to students.
Item 15	I share assessment criteria with students.
Socio-emotional Dimension	
Item 16	I am aware of the need to create a common understanding of formative assessment among teachers and students.
Item 17	I attend to students' emotional responses to assessments.
Item 18	I recognize that students' values, beliefs, and attitudes impact how they experience the process of formative assessment.
Item 19	I am aware of the impact that assessment feedback information might have on students' learning motivation.

(*continued on next page*)

(*continued*)

Item 20	I am sensitive to the ethical aspects of formative assessment, such as fairness and student privacy.
Item 21	I am aware of my responsibilities to cater for students' well-being during the formative assessment process.
Item 22	I am conscious of the fact that students have the right to benefit from formative assessment practices.

Appendix B. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.stueduc.2022.101183.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

Adie, L., Stobart, G., & Cumming, J. (2020). The construction of the teacher as expert assessor. *Asia-Pacific Journal of Teacher Education, 48*(4), 436–453.

AFT, American Federation of Teachers, National Council on Measurement in Education, & National Education Association, NCME, & NEA, & NEA. (1990). *Standards for teacher competence in educational assessment of students*. National Council on Measurement in Education.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545–557. https://doi.org/10.1037/0021-843X.112.4.545

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology, 76*, 732–740.

Andrade, H. L., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning, achievement, and academic self-regulation*. Routledge,.

Andrade, H. L., Bennett, R. E., & Cizek, G. J. (Eds.). (2019). *Handbook of formative assessment in the disciplines*. Routledge.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*, 1–16.

Antoniou, P., & James, M. (2014). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation and Accountability, 26*(2), 153–176.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74.

Boardman, A. G., & Woodruff, A. L. (2004). Teacher change and "high stakes" assessment: what happen to professional development? *Teaching and Teacher Education, 20*(6), 545–557.

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Boström, E., & Palm, T. (2020). Expectancy-value theory as anexplanatory theory for the effect of professional development programmes in formative assessmenton teacher practice. *Teacher Development, 24*(4), 539–558.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practices, 30*(1), 3–12.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles. Policy & Practice, 11*, 301–318.

Brown, G. T. L. (2019). Is assessment for learning really assessment. *Frontiers in Education, 4*(64). https://doi.org/10.3389/feduc.2019.00064

Cagasan, L., Care, E., Robertson, P., & Luo, R. (2020). Capturing formative assessment practices in the Philippines through an observation tool. *Educational Assessment Journal Special Issue: Classroom Observations of Formative Assessment Practice, 25*, 4. https://doi.org/10.1080/10627197.2020.1766960

Campbell, C., Murphy, J.A., & Holt, J.K. (2002). *Psychomtric analysis of an assessment literacy instrument: Applicability to preservice teachers*. [Paper presentation]. Annual meeting of the Mid-Western, Educational Research Association, Columbus, OH.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276.

Coniam, D., & Yan, Z. (2016). A comparative picture of the ease of use and acceptance of onscreen marking by markers across subject areas. *British Journal of Educational Technology, 47*(6), 1151–1167. https://doi.org/10.1111/bjet.12294

Darling-Hammond, L., Hyler, M.E., & Gardner, M. (2017). Effective teacher professional development. Learning Policy Institute.

De Simone, J. J. (2020). The roles of collaborative professional development, self-efficacy, and positive affect in encouraging educator data use to aid student learning. *Teacher Development, 24*(4), 443–465.

DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action in Teacher Education. XXXIV, 5/6*, 356–372.

DeLuca, C., & Volante, L. (2016). Assessment for learning in teacher education programs: Navigating the juxtaposition of theory and praxis. *Journal of the International Society for Teacher Education, 20*, 19–31.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016aa). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251–272.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016bb). Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy. *Educational Assessment, 21*(4), 248–266.

van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education, 86*(2), 247–264.

Desimone, L. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529. https://doi.org/10.1080/00273170701382864

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.

Fulmer, G. W., Lee, I. C. H., & Tan, K. H. K. (2015). Multi-level model of contextual factors and teachers' assessment practices: An integrative review of research. *Assessment in Education: Principles. Policy & Practice, 22*(4), 475–494. https://doi.org/10.1080/0969594X.2015.1017445

Gore, J., Miller, A., Fray, L., Harris, J., & Prieto, E. (2021). Improving student achievement through professional development: Results from a randomized controlled trial of Quality Teaching Rounds. *Teaching and Teacher Education, 101*. https://doi.org/10.1016/j.tate.2021.103297

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice, 33*(2), 14–18.

Heitink, M., van der Kleij, F., Veldkamp, B., Schildkamp, K., & Kippers, W. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review, 17*, 50–62.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Corwin Press,.

Heritage, M., & Wylie, E. C. (2020). *Formative assessment in the disciplines: Framing a continuum of professional learning*. Harvard Education Press,.

Herppich, S., Praetorius, A., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76*, 181–193.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*, 967–988.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55. https://doi.org/10.1080/10705519909540118

International Test Commission. (2017). The ITC guidelines for translating and adapting tests (2nd ed.). Retrieved on June 21, 2021 from ⟨https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf⟩.

Kennedy, M. K. (2019). How we learn about teacher learning. *Review of Research in Education, 43*(1), 138–162.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research, 86*(4), 945–980.

Klinger, D. A., Volante, L., & Deluca, C. (2012). Building teacher capacity within the evolving assessment culture in canadian education. *Policy Futures in Education, 10*(4), 447–460.

Lam, R. (2018). Testing, drilling and learning: What purpose does the Grade 3 Territory-wide System Assessment serve? *Asia Pacific Education Review, 19*(3), 363–374.

Lam, R. (2019). Teacher assessment literacy: Surveying knowledge, conceptions and practices of classroom-based writing assessment in Hong Kong. *System, 81*, 78–89.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2021). Formative assessment as practice: the role of students' motivation. *Assessment & Evaluation in Higher Education, 46*(2), 236–255.

Leighton, J.P., Gokiert, R.J., Cor, M.K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom-versus large-scale tests: Implications for assesment literacy. Assessment in Education: Principles, Policy & Practice, 17(1), 7–21.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.
Looney, A., Cummuning, J., van Der Kleij, F., & Haris, K. (2018). Reconceptualizing the role of teachers as assessors: Teacher assessment identity. *Assessment in Education: Principle, Policy & Practice, 25*(5), 442–467.
Lysaght, L., & O'Leary, M. (2013). An instrument to audit teachers' use of assessment for learning. *Irish Educational Studies, 32*(2), 217–232.
Mandinach, E. B., & Gummer, E. S. (2016). *Data Literacy for Teachers: Making It Count in Teacher Preparation and Practice*. Teachers College Press.
Mayo, S. T. (1967). *Pre-service preparation of teachers in educational measurement*. US Department of Health, Education and Welfare.
McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*, 64–82.
Mertler, C.A. (2003, October). Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference? [Paper presentation]. Meeting of the Mid-Western Educational Research Association. Columbus, OH.
Mertler, C.A., & Campbell, C. (2005). Measuring teachers' knowledge & application of classroom assessment concepts: development of the assessment literacy inventory. In Annual meeting of the American Educational Research Association, Montreal.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50*, 741–749.
Pastore, S., & Andrade, H. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education, 84*, 128–138. https://doi.org/10.1016/j.tate.2019.05.003
Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10–12.
Popham, W. J. (2018). *Assessment literacy for educators in a hurry*. ASCD.
Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: the Spanish instance, building upon Black and Wiliam. *The Curriculum Journal, 18*(1), 27–38.
Schneider, C., & Bodensohn, R. (2017). Student teachers' appraisal of the importance of assessment in teacher education and self-reports on the development of assessment competence. Assessment in Education: Principles, Policy & Practice, 24(2), 127–146.
Smith, L. G. (2011). *Advancing formative assessment in every classroom: A guide for instructional leaders*. American Association of School Administrators.
Soh, K. C., & Zhang, L. (2017). The development and validation of a teacher assessment literacy scale: A trail report. *Journal of Linguistics and Language Teaching, 8*(1), 91–116.
Stiggins, R. J. (2017). *The perfect assessment system*. ASCD.
Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics*. Pearson.
Wang, W. C., Yao, G., Tsai, Y. J., Wang, J. D., & Hsieh, C. L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Quality of Life Research, 15*, 607–620.
Wiliam, D. (2017). Assessment and learning: some reflections. *Assessment in Education: Principles, Policy & Practice, 24*(3), 394–403.
Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Erlbaum.
Will, K. K., McConnell, S. R., Elmquist, M., Lease, E. M., & Wackerle-Hollman, A. (2019). Meeting in the middle: Future directions for researchers to support educators' assessment literacy and data-based decision making. *Front Educ, 4*, 106.
Willis, J., Adie, L., & Klenowski, V. (2013). Conceptualizing teachers' assessment literacies in an era of curriculum and assessment reform. *The Australian Educational Researcher, 40*(2), 241–256.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Erlbaum.
Wylie, E. C. (2020). Observing formative assessment practice: Learning lessons through validation. *Educational Assessment, 25*(4), 251–258.
Wylie, E. C., & Lyon, C. J. (2015). The fidelity of formative assessment implementation: Issues of breadth and quality. *Assessment in Education: Principles, Policy & Practice, 22*(1), 140–160.
Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010
Yan, Z. (2021). Assessment-as-learning in classrooms: The challenges and professional development. *Journal of Education for Teaching, 47*(2), 293–295. https://doi.org/10.1080/02607476.2021.1885972
Yan, Z., & Cheng, E. C. K. (2015). Primary teachers' attitudes, intentions and practices regarding formative assessment. *Teaching and Teacher Education, 45*, 128–136. https://doi.org/10.1016/j.tate.2014.10.002
Yan, Z., & Brown, G. T. L. (2021). Assessment for learning in the Hong Kong assessment reform: A case of policy borrowing. *Studies in Educational Evaluation, 68*, Article 100985. https://doi.org/10.1016/j.stueduc.2021.100985
Yan, Z., & Pastore, S. (2022). Assessing teachers' strategies in formative assessment: The Teacher Formative Assessment Practice Scale. *Journal of Psychoeducational Assessment*. https://doi.org/10.1177/07342829221075121
Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. Assessment in Education: Principles. *Policy & Practice, 28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042
Zhang, Z., & Burry-Stock, J. (1997). Assessment practices inventory: A multivariate analysis of teachers' perceived assessment competency. [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL.

Zi Yan is an Associate Professor in the Department of Curriculum and Instruction at The Education University of Hong Kong. His publications and research interests focus on two related areas, i.e., educational assessment in the school and higher education contexts with an emphasis on student self-assessment; and Rasch measurement, in particular its application in educational and psychological research. A recent book co-edited with Lan Yang is entitled Assessment as learning: Maximising opportunities for student learning and achievement.

Serafina Pastore is an Associate Professor in the Department ofResearch and Humanities Innovation at the University of Bari (Italy) where she also teaches in PGCE courses. Her research interests include the intersections of assessment practice, teacher education and educational policy as they operating in the context of school and university innovation. Her recent work focuses on teacher assessment literacy.