Article

Journal of Psychoeducational Assessment
2022, Vol. 40(5) 592–604

DOI: 10.1177/07342829221075121
journals.sagepub.com/home/jpa
SAGE

Assessing Teachers' Strategies in Formative Assessment: The Teacher Formative Assessment Practice Scale

Zi Yan[1] and Serafina Pastore[2]

Abstract

A significant challenge in studying formative assessment is the lack of suitable instruments for assessing teachers' formative assessment practices. This paper reports the development of the Teacher Formative Assessment Practice Scale (TFAPS) and its psychometric properties based on two samples of primary and secondary school teachers: one from Hong Kong (N = 449) and the other from Italy (N = 309). Exploratory factor analysis identified two distinct factors, including teacher-directed formative assessment (TdFA, six items) and student-directed formative assessment (SdFA, four items). The confirmatory factor analysis supported this two-factor structure. Rasch analysis provided further psychometric evidence regarding scale dimensionality and item quality. This study suggests that TFAPS is an appropriate instrument for assessing teachers' formative assessment practices, but the cultural influence on teachers' formative assessment practices should be noted in the applications of the instrument.

Keywords

formative assessment, scale development, validation, teacher professional development

Introduction

Formative assessment plays a strategic role in the teaching and learning process (Black & Wiliam, 1998, 2009; Schneider & Johnson, 2019). The major principles of formative assessment—related to providing feedback and enhancing students' metacognition—are well recognized in their ability to improve student learning (Andrade & Heritage, 2018; Hattie & Timperley, 2007). Nevertheless, formative assessment practice appears to be less frequent than ideal (Bennett, 2011; Elwood,

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

Corresponding Author:
Zi Yan, D1-1/F-49, Department of Curriculum and Instruction, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, N.T., Hong Kong.
Email: zyan@eduhk.hk

2006; Erickson, 2007), and the forms of formative assessment teachers conducted vary widely across contexts (Andrade et al., 2019; Yan et al., 2021).

One significant gap in this field is the lack of instruments suitable for assessing teachers' formative assessment practices, largely due to the lack of a consensus regarding what can be counted as formative assessment practices: while formative assessment principles are generally recognized by teachers as consistent, formative assessment practices, sometimes, are complex, confusing, as well as difficult to implement in the classroom (Harrison & Howard, 2009). This leads to substantial challenges both for research and teacher professional development. First, the lack of an instrument makes it difficult to understand and describe teachers' formative assessment practices and, therefore, makes targeted intervention programs less likely. Second, without an appropriate instrument, we do not know whether and how much a professional development intervention can enhance teachers' formative assessment practices. We also do not know whether enhanced formative assessment practices can result in desirable educational outcomes. While different attempts have been made to assess teachers' assessment literacy (e.g., identifying assessment literacy main components in terms of knowledge and skills) (DeLuca et al., 2016), the assessment of formative assessment practice has been realized predominantly through unpublished survey instruments or observation protocols.

To address this gap in assessing teachers' formative assessment practices, the current study reports the development of a scale based on a clear theoretical framework and its preliminary psychometric evidence.

Formative Assessment Practices

The term formative assessment encompasses significantly variant practices in classrooms (Bennett, 2011). Generally, formative assessment refers to a set of activities carried out by teachers and students in order to collect information "to be used as feedback to modify teaching and learning activities" (Black & Wiliam, 1998, p. 140). Rather than focusing on what has been attained by students, formative assessment helps to identify learning gaps, scaffold new learning, anticipate future teaching steps (Bennett & Gitomer, 2009), and promote self-regulation of student learning (Andrade & Heritage, 2018).

A large body of literature supports teachers in articulating a theoretically sound approach to formative assessment (Andrade & Heritage, 2018; Black & Wiliam, 1998, 2009; Clark, 2012; Stiggins & DuFour, 2009). More specifically, theory and practice in formative assessment have emphasized the key aspects of collecting, evaluating and using evidence of student learning (Heritage, 2010; 2013; McMillan et al., 2013). Over the years, a focus has been on strategies purposed to integrate formative assessment into the process of teaching and learning. In this perspective, Wiliam and Thompson (2008) proposed a framework in which various formative assessment practices are categorized into five key strategies. These five strategies include (1) clarifying and sharing learning intentions and criteria for success; (2) engineering effective classroom discussions, questions, and learning tasks; (3) providing feedback that moves learners forward; (4) activating students as instructional resources for one another; and (5) activating students as the owners of their own learning (Wiliam & Thompson, 2008, p. 64). This framework was used in the current study to guide scale development due to two reasons. First, this framework provides a unifying basis for understanding formative assessment practices. It covers a wide range of essential formative assessment aspects, especially the active role of students in formative assessment, which is a current trend of formative assessment research (Allal, 2020; Yan & Brown, 2021). Second, this framework provides a clear structure to support item development. The five strategies enable a comprehensive and balanced set of items to capture a variety of formative assessment practices at an appropriate level of specificity.

As formative assessment practice is likely to be influenced by the teaching context (Heitink et al., 2016; Yan et al., 2021), we collected data in two different cultures (i.e., Hong Kong and Italy) for a cross-cultural validation with the aim of enhancing the generalization of the developed scale across different contexts. Even though previous studies on teacher formative assessment practices carried out in different educational settings (from pre-k to k-12 and higher education), with different subject matters (e.g., English and Mathematics) and with different targets of students (e.g., special learning needs or English language learners), these studies have not explicitly examined how school sector affects teachers' formative assessment practices. Therefore, we collected data from both primary and secondary school teachers as past studies showed that teachers' assessment practices could differ in primary and secondary schools (Bol et al., 1998).

Assessing Teacher Formative Assessment Practice

Although formative assessment has been a popular topic in educational research in the past two decades, to the authors' knowledge, there is no one widely recognized and applied survey instrument. Most studies have used self-developed instruments to assess teachers' formative assessment practices. For example, McMillan and his colleagues (2010) used seven items to assess teachers' use of formative assessment. Their instrument focused on providing feedback (3 items), but the other items were more like general descriptions of formative assessment (e.g., assessments that were used to guide further instruction) rather than concrete assessment practices. Furthermore, students' involvement in formative assessment was largely neglected. Similarly, Song and Koh (2010) used a questionnaire to assess different assessment modes teachers adopted for formative purposes. These studies did not report a formal process of development and validation for the instruments and, more importantly, they lack an explicit underlying theoretical framework. These shortcomings limit the more generalized use of those instruments.

Compared to surveys, direct observation appears more popular in assessing teachers' formative assessment practices (e.g., Cagasan et al., 2020; Hartmeyer et al., 2016; Lyon et al., 2020). Observation protocols have been very useful to broadly understand formative assessment practices and to support teachers' reflection on formative assessment practice improvement (Goe et al., 2017; Wylie, 2020). However, these attempts to capture formative assessment practices raise different concerns in terms of reliability and validity. Moreover, the lack of comparative studies using the same instrument and the difficulties in using these instruments with large samples (OECD, 2012) negatively impact the conceptualization of formative assessment practices, as well as on the quality of teacher professional development.

Given the need for more research on the implementation and scalability of formative assessment, this paper reports a study aiming to develop and validate a scale that can clearly identify observable aspects of formative assessment practices. The developed instrument should contribute to research and teacher training by supporting profiling and analyzing teachers' formative assessment practices.

Method

Instruments

The Development of the Teacher Formative Assessment Practice Scale (TFAPS). The Teacher Formative Assessment Practice Scale (TFAPS) was developed according to the 5-category framework of formative assessment strategies (Wiliam & Thompson, 2008). After consultation of the available literature and with two focus group interviews with primary ($N = 5$) and secondary ($N = 5$) teachers from Hong Kong, 15 items were developed, with each of the categories having three items. The

initial item pool was subject to review by seven experts in the field of educational assessment. All experts were asked to (a) assign each item to the category to which it belongs, and (b) rate the relevance of each item to teacher formative assessment practice on a scale ranging from "essential," "useful, but not essential," to "not necessary." The proportion of substantive agreement (i.e., the proportion of experts who assign an item to its intended category) was calculated for each item (Anderson & Gerbing, 1991), with a proportion of .7 or above indicating acceptable substantive agreement. As for the item relevance, the content validity ratios (CVRs; Lawshe, 1975) were computed, and a positive CVR (i.e., four or more experts indicating essential) indicated acceptable content validity. The results showed that five items had an unsatisfactory substantive agreement and/or a negative CVR. These items were examined, then removed, and further revisions were made to the remaining 10 items according to expert comments on the item wording. Each formative assessment category in the Wiliam and Thompson (2008) framework had two items. All items are with a six-point Likert-type response scale ranging from "Never," "Rarely," "Seldom," "Sometimes," "Frequently," to "Very frequently." The scale was developed in English and then translated into Cantonese and Italian, respectively, following the forward- and backward-translation procedure (International Test Commission, 2017).

The Formative Assessment Self-efficacy Scale

We examined the relationship between teachers' performance on the TFAPS and their self-efficacy of formative assessment to check the external validity of the TFAPS. Teachers' self-efficacy in conducting formative assessment were assessed with the self-efficacy subscale in the Teacher's Conceptions and Practices of Formative Assessment Questionnaire (Yan & Cheng, 2015) which has six items (Rasch reliability = 0.84; for example, *I can design appropriate assessment tasks for formative assessment*).

Participants and Data Collection

The current study has two samples from Hong Kong and Italy, respectively. A total of 449 teachers from 12 Hong Kong schools participated in the survey. There were 295 (65.7%) females, 151 (33.6%) males, and 3 (0.7%) without gender information. 263 (58.6%) teachers were from primary schools and 186 (41.4%) were from secondary schools.

Data from Italy were collected from 309 teachers in 10 Italian schools in the Apulian region (South of Italy). Most of the teachers were female ($N = 278$, 90.0%). 134 (43.4%) teachers were from primary schools and 175 (56.6%) were from secondary schools.

Data Analyses

As the number of missing values was very small (<0.1%) for all the items, imputation was unnecessary. The Hong Kong sample was used for exploratory factor analysis (EFA) to explore the factor structure, and the Italian sample was used for confirmatory factor analysis (CFA) to confirm the derived factor structure. EFA was conducted using SPSS with the principal components method and promax rotation. The eigenvalue test, scree plots, and interpretability were observed to determine the identified factors. Items were retained when the factor loading was higher than 0.4 and cross-loading was below 0.3. CFA was conducted using AMOS. Multiple fit indices were used to examine the model-data fit: 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). Values of CFI and TLI higher than 0.90, SRMR and RMSEA lower than 0.08 (Fan &

Sivo, 2007; Hu & Bentler, 1999; McDonald & Ho, 2002) were considered indicating acceptable fit.

Rasch analysis was applied to the complete data set combining Hong Kong and Italian samples. The Rasch model adopts a "data fit the model" approach that requires the empirical data to satisfy a priori requirements. In contrast to the two- and three-parameter item response models, the Rasch model has only one item parameter, that is, item difficulty, with other parameters (e.g., discrimination and pseudo-chance parameter) being restricted to constants. By doing so, the Rasch model produces sample-free item calibrations or item-free person measures that are essential for the fundamental measurement (Bond et al., 2020). Due to its theoretical advantages in dealing with ordinal data collected through Likert-type instruments, Rasch analysis has been widely applied in empirical studies to scrutinize the quality of instruments (e.g., Antipkina & Ludlow, 2020; Brann et al., 2021; Yan, 2018, 2020). Multiple indicators were used to examine scale quality, including response category functioning, item fit statistics (i.e., Infit MNSQ and Outfit MNSQ), and differential item functioning (DIF) across gender, educational level (primary vs. secondary), and region (Hong Kong vs. Italy).

In addition, Cronbach's alpha and Rasch reliability were calculated for the TFAPS as the indicators of reliability.

Results

Exploratory Factor Analysis

The EFA was run on the data from the Hong Kong sample. The value for Kaiser–Meyer–Olkin measure of sampling adequacy was 0.834, and Bartlett's test of sphericity was χ^2 (45) = 1834.167, $p < .001$, indicating that the data were appropriate for factor analysis. The results showed a clear two-factor structure which was well supported by the eigenvalues and the scree-test (Cattell, 1966). The six items pertaining to the first three formative strategies (i.e., clarifying and sharing learning intentions and criteria for success; engineering effective classroom discussions, questions, and learning tasks; and providing feedback that moves learners forward) were grouped together. These items are mainly about formative assessment strategies that are usually initiated and implemented by the teacher. Thus, they can be entitled teacher-directed formative assessment (TdFA). The four items about the other two strategies (i.e., activating students as instructional resources for one another; and activating students as the owners of their own learning) formed the second factor. These strategies are mainly peer-assessment and self-assessment for formative purposes and, there, can be appropriately named as student-directed formative assessment (SdFA). The correlation between these two factors was 0.52. Table 1 presents the factor loadings of the 10 items.

Confirmatory Factor Analysis

A CFA with maximum likelihood estimation was applied to the data from the Italian sample. The Kaiser–Meyer–Olkin measure (0.718) and the results of Bartlett's test of sphericity (χ^2 (45) = 508.534, $p < .001$) indicated appropriateness for factor analysis. The initial model-data fit was not satisfactory: CFI = 0.779; and TLI = 0.708; RMSEA = .100; SRMR = 0.072. After checking the modification indices and considering conceptual considerations, the residuals of two pairs of items (Items #1 and #2, both are about clarifying and sharing learning intentions and criteria for success; Items #9 and #10, both are about activating students as the owners of their own learning) were allowed to correlate. The CFA was re-run and the results demonstrated satisfactory model-data fit: CFI = 0.935; and TLI = 0.908; RMSEA = 0.056; SRMR = 0.062. Figure 1 displays the standardized factor loadings of items and correlations between factors.

Table 1. Item factor loadings in the EFA.

Item No	Item Description	Factor 1	Factor 2
Item 1	I share the learning intention before students start working in class	.546	
Item 2	I clarify what is valued for each assessment task	.491	
Item 3	I use various assessment activities in the classroom to check students' mastery of course content	.627	
Item 4	I ensure homework can check students' learning progress	.840	
Item 5	I point out students' strengths and weaknesses in my feedback	.684	
Item 6	I provide suggestions for students to improve their performance	.723	
Item 7	I ask students to evaluate peers' work		.897
Item 8	I ask students to provide feedback to help peers improve		.892
Item 9	I ask students to identify strengths and weaknesses in their own work		.832
Item 10	I ask students to identify strategies that will improve their own work		.737

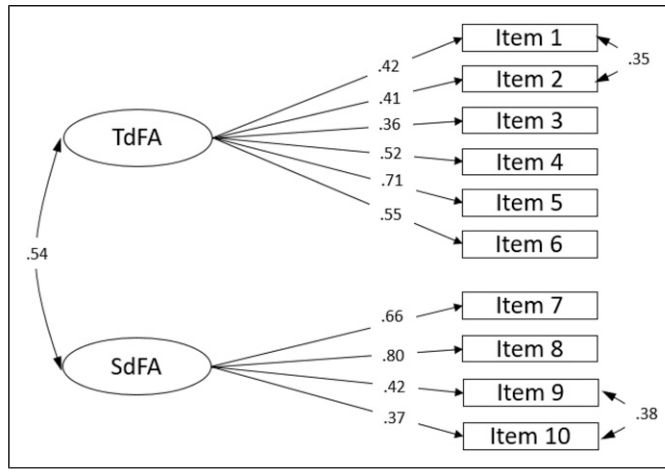

Figure 1. Standardized factor loadings of items and correlations between factors in the CFA.

Rasch Analysis

As the factor analysis identified a two-factor structure, a multidimensional Rasch model (Adams et al., 1997) was employed on the full dataset including both the Hong Kong and Italian samples ($N = 758$). The results (see Table 1) showed that all items fitted well to the Rasch model. Most items had Infit and Outfit MNSQs falling within the desirable range (0.75–1.33) (Wilson, 2005) with only one exception (Item #4), which was within the acceptable range (0.5–1.5) (Linacre, 2006). Item fit statistics (i.e., Infit and Outfit MNSQ) reflect the extent to which each item in a subscale measures a unidimensional latent construct. Their expected value is 1.0. Values less than 1.0 indicate observations are too predictable, while values greater than 1.0 indicate unmodeled noise. The six-point response scale functioned sufficiently well as the step calibrations (i.e., the measures of the transition points between adjacent categories) increased monotonically from −1.84, −1.58, −0.69, 1.21, to 2.90 logits. There was no item showing DIF across gender or educational level. However, five items demonstrated substantial DIF across regions (a difference of item difficulty equal to or greater than 0.5 logits, Wang et al., 2006). Hong Kong teachers found

it much harder to endorse three items (#3, #6, and #9) than did their Italian peers. In contrast, Italian teachers conceived two items (#4 and #7) to be more difficult than did Hong Kong teachers with the same overall level on the measured latent trait. The item difficulty, standard error and item fit statistics for each item are presented in Table 2.

The Pearson correlations between Rasch-calibrated person measures on the two subscales of the TFAPS and their self-efficacy of implementing formative assessment were calculated to test the external validity of the TFAPS. The Cronbach's alphas of the self-efficacy scale were 0.93 (the Hong Kong sample), 0.76 (the Italian sample), 0.88 (the combined sample). As shown in Table 3, formative assessment self-efficacy was significantly and moderately correlated with formative assessment practices, with a stronger correlation for TdFA. Teachers' mean measures on TdFA and SdFA were 1.92 and 0.25 logits, respectively.

Reliability

Cronbach's alpha and Rasch reliability were calculated on the Hong Kong, Italian, and the combined samples, respectively (see Table 4). Both the Cronbach's alphas and Rasch reliabilities for the Hong Kong and the combined sample were satisfactory. The Cronbach's alphas for the Italian sample were relatively low, but still acceptable, considering Cronbach's alpha represents a lower bound to the reliability and often underestimates the true reliability (Sijtsma, 2009).

Convergent and discriminant validity

The average variance extracted (AVE) and composite reliability (CR) coefficients were computed on the combined sample. Although the AVE values were less than 0.5, the CR coefficients for both TdFA and SdFA were higher than 0.7 (see Table 5), indicating that the scale had acceptable convergent validity (Fornell & Larcker, 1981). The square root of AVE (0.53 for TdFA and 0.66 for SdFA) were greater than the inter-factor correlations (0.34), indicating the distinctiveness of each subconstruct and supporting the discriminant validity of the scale (Fornell & Larcker, 1981; Hair et al., 1998).

Table 2. Item difficulty, standard error, and item fit statistics for the 10-item TFAPS.

Item No	Item Measure*	SE	Infit MNSQ	Outfit MNSQ
Teacher-directed formative assessment				
Item 1	0.16	0.04	1.17	1.16
Item 2	0.28	0.04	0.97	0.96
Item 3	0.06	0.04	0.98	0.97
Item 4	0.04	0.04	1.4	1.4
Item 5	−0.08	0.04	0.97	0.95
Item 6	−0.46	0.08	0.86	0.85
Student-directed formative assessment				
Item 7	0.47	0.03	1.06	1.09
Item 8	0.17	0.03	0.94	0.94
Item 9	−0.26	0.03	0.89	0.90
Item 10	−0.38	0.05	0.97	0.96

Note. * All measures are in logits.

Table 3. Correlations between TFALS and formative assessment self-efficacy.

	TdFA	SdFA	Self-efficacy
TdFA	—		
SdFA	.416**	—	
Self-efficacy	.457**	.258**	—
Mean	1.92	0.25	1.67
S.D.	1.14	1.21	2.09

Note. **$p < .01$.

Table 4. Reliabilities on the Hong Kong, Italian, and combined samples.

	Hong Kong Sample		Italian Sample		The Combined Sample	
	Cronbach's Alpha	Rasch Reliability	Cronbach's Alpha	Rasch Reliability	Cronbach's Alpha	Rasch Reliability
TdFA	0.77	0.81	0.60	0.70	0.70	0.74
SdFA	0.87	0.86	0.66	0.69	0.75	0.77

Table 5. The average variance extracted and composite reliability on the combined sample.

	CR	AVE	TdFA	SdFA
TdFA	0.70	0.28	0.53*	
SdFA	0.73	0.43	0.34	0.66*

Note. CR = composite reliability; AVE = average variance explained; * square root of the AVE values.

Discussion

The current study aims to address an important gap in the research and practice of formative assessment by developing and validating an instrument for assessing teachers' formative assessment practices. Both the factor analysis and Rasch analysis demonstrated a clear theory-based two-factor structure, that is, teacher-directed formative assessment (TdFA, six items) and student-directed formative assessment (SdFA, four items).

The development of the TFAPS has the potential to contribute to the research and practice of formative assessment. To date, how teachers effectively implement formative assessment is an important agenda in educational research, as well as in educational reforms. However, the lack of a valid and easy-for-use instrument for assessing formative assessment practices have, for a long time, impacted negatively on teacher development programs, as well as on teacher assessment practices.

In this vein, the development of the TFAPS should be informative in terms of teacher education and teacher practice. Against the backdrop of teacher professionalism, the scale 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 evidence-informed teacher training (Graham et al., 2020). The instrument could also be used to gauge the need for professional development and evaluate the effectiveness of teacher training programs or interventions for enhancing teachers' formative assessment literacy.

Moreover, as the effects of formative assessment vary widely among different implementations and student populations, the TFAPS can facilitate cross-context/cross-cultural research on the implementation and scalability of formative assessment.

Teachers' mean measure on TdFA (1.92 logits) was much higher than that on SdFA (0.25 logits) (see Table 3), indicating teachers still playing the dominant role in formative assessment practices, while their support of students' active role is yet to be enhanced. This is consistent with previous studies (e.g., Fallows & Chandramohan, 2001; Yan, 2021; Yan & Brown, 2021) in that student-directed formative assessment (e.g., self-assessment and peer-assessment) is less implemented than teacher-directed formative assessment, in spite of student-directed formative assessment being well recognized as a regular type of formative assessment in academia (e.g., Heritage, 2010; Wiliam & Thompson, 2008; Yan & Brown, 2017).

Teachers' measures on the two subscales of the TFAPS were significantly and moderately correlated with their self-efficacy on implementing formative assessment, providing reasonable support for the external validity of the TFAPS. Self-efficacy had a stronger correlation with TdFA than with SdFA, suggesting that teachers' confidence with formative assessment is more closely linked to those formative assessment practices dominated by teachers. Following the perspective of teacher identity as assessor/facilitator and the conceptual framework of teacher assessment literacy (Looney et al., 2018; Xu & Brown, 2016) this aspect needs more attention in teacher professional development.

The TFAPS showed relatively low Cronbach's alpha with the Italian sample. This result should be interpreted with caution. We adopt a reflective measurement model in developing and analyzing the scale. That is, the latent construct (i.e., teachers' tendency to use different formative assessment strategies) is reflected in different observable indicators (i.e., formative assessment strategies). However, a teacher who frequently uses one strategy does not necessarily use the other strategies in a similar frequency, which might result in relatively low internal consistency within the scale.

Half of the items demonstrated significant DIF across regions (Hong Kong vs. Italy), supporting the argument that formative assessment practices are context-/culture-dependent (Heitink et al., 2016; Yan et al., 2021). Within the Rasch measurement framework (e.g., Bond et al., 2020), item difficulty is expected to remain consistent across suitable relevant samples; DIF is the loss of item estimate invariance across subsamples of respondents. The diagnostic function of DIF might reveal either weakness in item development or substantive, meaningful differences in the construct(s) of interest across important subsamples. Given that items were invariant (no DIF) across subsamples based on gender and educational level, it is possible that DIF across regions indicates that the understanding and implementation of particular formative assessment strategies are different for Hong Kong and Italian teachers. Taking teachers' performances on the TdFA subscale as an example, Hong Kong teachers found Item #4 (ensure homework can check students' learning progress) easier to endorse than did their Italian peers, while Item #3 (use various assessment activities in the classroom to check students' mastery of course content) and Item #6 (provide suggestions for students to improve their performance) more difficult to endorse for Hong Kong teachers. A speculative explanation is that homework load is very demanding in Hong Kong and teachers are used to checking students' learning progress through homework. The Italian teachers, however, tend to consider homework as not really informative of the student learning progression because the parental help, sometimes, affects the validity of the gathered evidence on student learning. In contrast, Italian teachers tend to use a wider variety of assessment activities in the classroom than Hong Kong teachers. Also, Italian teachers may be inclined to provide more suggestions for students. Future studies might consider analyzing the influence of sociocultural context, as well as of policy framework (e.g. teacher professional standards or teacher education programs) on key strategies of formative assessment. In this perspective,

examining teacher responses to a common scale, like the TFAPS, will offer valuable insights for cross-cultural formative assessment research.

One limitation of the current study is that the consequential validity (see Messick, 1995) has not yet been investigated. It would be appropriate for future studies to examine whether a better understanding and capturing of teachers' formative assessment practices through the use of the TFALS resulted in desirable outcomes, such as teachers' improved self-awareness of formative assessment practices and better effects of teacher training programs that aim to enhance teachers' formative assessment practices.

Conclusions

This study reports the development of the TFAPS and its psychometric properties from one Hong Kong and one Italian sample. The results demonstrated that the TFAPS was a valid and appropriate instrument for assessing teachers' formative assessment practices. This study fills an important methodological gap in the research and practice of formative assessment. The TFAPS can be used to support teachers to clearly recognize how they are carrying out formative assessment in the classroom. It can also help gauge the need for professional training and evaluate the effectiveness of teacher training programs.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work described in this paper was supported by a General Research Fund from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. EDUHK 18607118).

ORCID iD

Zi Yan https://orcid.org/0000-0001-9305-884X

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Allal, L. (2020). Assessment and the co-regulation of learning in the classroom. *Assessment in Education: Principles, Policy & Practice*, *27*(4), 332–349. https://doi.org/10.1080/0969594x.2019.1609411

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*(5), 732–740. https://doi.org/10.1037/0021-9010.76.5.732

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

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

Antipkina, I., & Ludlow, L. H. (2020). Parental involvement as a holistic concept using Rasch/Guttman scenario scales. *Journal of Psychoeducational Assessment*, *38*(7), 846–865. https://doi.org/10.1177/0734282920903164

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

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). Springer. https://doi.org/10.1007/978-1-4020-9964-9_3

Black, P. J., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–148. https://doi.org/10.1177/003172171009200119

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

Bol, L., Stephenson, P. L., O'connell, A. A., & Nunnery, J. A. (1998). Influence of experience, grade level, and subject area on teachers' assessment practices. *The Journal of Educational Research*, *91*(6), 323–330. https://doi.org/10.1080/00220679809597562

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

Brann, K. L., Boone, W. J., Splett, J. W., Clemons, C., & Bidwell, S. L. (2021). Development of the school mental health self-efficacy teacher survey using Rasch analysis. *Journal of Psychoeducational Assessment*, *39*(2), 197–211. https://doi.org/10.1177/0734282920947504

Cagasan, L., Care, E., Robertson, P., & Luo, R. (2020). Developing a formative assessment protocol to examine formative assessment practices in the Philippines. *Educational Assessment*, *25*(4), 259–275. https://doi.org/10.1080/10627197.2020.1766960

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

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

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

Elwood, J. (2006). Formative assessment: Possibilities, boundaries and limitations. *Assessment in Education*, *13*(2), 215–232. https://doi.org/10.1080/09695940600708653

Erickson, F. (2007). Some thoughts on 'proximal' formative assessment of student learning. *Yearbook of the National Society for Study of Education*, *106*(1), 186–216. https://doi.org/10.1111/j.1744-7984.2007.00102.x

Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: Reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education*, *6*(2), 229–246. https://doi.org/10.1080/13562510120045212

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

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.2307/3151312

Goe, L., Wylie, E. C., Bosso, D., & Olson, D. (2017). State of the states' teacher evaluation and support systems: A perspective from exemplary teachers. *ETS Research Report Series*, *2017*(1), 1–27. https://doi.org/10.1002/ets2.12156

Graham, L. J., White, S. L. J., Cologon, K., & Pianta, R. C. (2020). Do teachers' years of experience make a difference in the quality of teaching? *Teaching and Teacher Education*, *96*, Article 103190. https://doi.org/10.1016/j.tate.2020.103190

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis*. Prentice Hall.

Harrison, C., & Howard, S. (2009). *Inside the primary black box: Assessment for learning in primary and early years classrooms*. GL Assessment.

Hartmeyer, R., Stevenson, M. P., & Bentsen, P. (2016). Evaluating design-based formative assessment practices in outdoor science teaching. *Educational Research*, *58*(4), 420–441. https://doi.org/10.1080/00131881.2016.1237857

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

Heitink, M. C., van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review*, *17*(2), 50–62. https://doi.org/10.1016/j.edurev.2015.12.002

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

Heritage, M. (2013). *Formative assessment in practice: A process of inquiry and action*. Harvard Education Press.

Hu, L. T., & 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), 1–55. https://doi.org/10.1080/10705519909540118

International Test Commission (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J. (2006). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. University of Chicago.

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. (2018). Reconceptualising the role of teachers as assessors: Teacher assessment identity. *Assessment in Education: Principles, Policy & Practice*, *25*(5), 442–467. https://doi.org/10.1080/0969594x.2016.1268090

Lyon, C. J., Nabors Oláh, L., & Brenneman, M. (2020). A formative assessment observation protocol to measure implementation: Evaluating the scoring inference. *Educational Assessment*, *25*(4), 288–313. https://doi.org/10.1080/10627197.2020.1766957

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82. https://doi.org/10.1037/1082-989x.7.1.64

McMillan, J. H., Cohen, J., Abrams, L., Cauley, K., Pannozzo, G., & Hearn, J. (2010). *Understanding secondary teachers' formative assessment practices and their relationship to student motivation.* ERIC. https://files.eric.ed.gov/fulltext/ED507712.pdf.

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research & Evaluation*, *18*(2), 1–15. https://doi.org/10.7275/tmwm-7792

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*(9), 741–749. https://doi.org/10.1037/0003-066x.50.9.741

OECD (2012). *Assessment for learning. The case of formative assessment*. OECD.

Schneider, M. C., & Johnson, R. L. (2019). *Using formative assessment to support student learning objectives*. Routledge.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Song, E., & Koh, K. (2010August). *Assessment for learning: Understanding teachers' beliefs and practices*. [Paper presentation]. The 36th Annual Conference of the International Association of Educational Assessment (IAEA) on "Assessment for the Future Generations". Bangkok, Thailand.

Stiggins, R., & DuFour, R. (2009). Maximizing the power of formative assessments. *Phi Delta Kappan*, *90*(9), 640–644. https://doi.org/10.1177/003172170909000907

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*(4), 607–620. https://doi.org/10.1007/s11136-005-4365-7

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.

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. https://doi.org/10.1080/10627197.2020.1766955

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. (2018). The Self-assessment Practice Scale (SaPS) for students: Development and psychometric studies. *The Asia-Pacific Education Researcher*, *27*(2), 123–135. https://doi.org/10.1007/s40299-018-0371-8

Yan, Z. (2020). Developing a short form of the self-assessment practices scale: Psychometric evidence. *Frontiers in Education*, *4*, Article 153. https://doi.org/10.3389/feduc.2019.00153

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., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education*, *42*(8), 1247–1262. https://doi.org/10.1080/02602938.2016.1260091

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., & 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., 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

Author Biographies

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, that is, 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 of Research and Humanities Innovation, University of Bari (Italy). Her research focuses 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.