Exploring the dimensionality of self-perceived performance assessment literacy (PAL)

Michael P. Kelly 1 · Richard Feistman 2 · Emily Dodge 3 · Andresse St. Rose 4 · Josh Littenberg-Tobias 5

Received: 5 October 2019 / Accepted: 28 October 2020 / Published online: 4 November 2020
© Springer Nature B.V. 2020

Abstract
Performance assessments attempt to provide a practical and authentic demonstration of students’ learning. Despite growing investments in performance assessments by states, as well as researchers’ theorized value of this type of assessment, the field has not developed a measure of assessment literacy specific to performance assessments that has sufficient psychometric evidence to support it. This study begins important research on developing a quantitative measure that can be used by educational practitioners to self-evaluate their own performance assessment literacy (PAL). Using the Quality Performance Assessment (QPA) framework from the Center for Collaborative Education as a foundation, this study explores and confirms the dimensionality of a 27-item survey instrument that assesses educational practitioners’ perceptions of their PAL using exploratory and confirmatory factor analysis. Our findings provide evidence that the instrument captures five reliable dimensions of PAL: valid design, reliable scoring, data analysis, fair assessment, and student voice and choice.

Keywords Assessment literacy · Performance assessment · Survey research · Scale development · Factor analysis

The Every Student Succeeds Act (2015) allows up to seven states, or groups of states, to apply for flexibility under section 1204: Innovative Assessment Demonstration Authority. With greater assessment flexibility, states have made efforts to create

✉ Michael P. Kelly
michaelpkelly99@gmail.com

1 Boston College, Chestnut Hill, USA
2 Teach Plus, Boston, USA
3 Harvard, Cambridge, USA
4 The Center for Collaborative Education, Boston, USA
5 Massachusetts Institute of Technology, Cambridge, USA
balanced assessment systems (Wylie and Lyon 2017) with the key feature of not relying exclusively on one single type of assessment (e.g., statewide standardized multiple-choice assessments). Instead, these systems incorporate multiple types of assessments, especially formative and summative classroom assessments. One type of classroom assessment that is growing in popularity among states is performance assessment (Hoffman et al. 2015). For example, in academic year 2014–2015, the New Hampshire Department of Education gained federal approval to operate the Performance Assessment of Competency Education (PACE) pilot program in which participating districts administer common and local performance assessments instead of the Smarter Balanced assessment in most grades and subjects. Additionally, the New York Performance Assessment Consortium, with 38 member high schools in New York City, Rochester, and Ithaca Public Schools, has a waiver from the state to forego all but one state Regents exams, and instead, students must pass performance assessments across academic subjects to graduate. More recently, in Massachusetts, school and union leaders across multiple districts have formed the Massachusetts Consortium for Innovative Education Assessment (MCIEA) with the goal of creating a new accountability model that measures academic learning using performance assessments.

Performance assessments are multistep assessments that focus on the creation or refinement of some artifact to promote use of higher-order thinking skills and address real-world problems (Frey and Schmitt 2007; Tung 2010). Several studies have found that performance assessments provide valuable and fair learning experiences to all students, as well as useful assessment data to educational practitioners (Brown and Mednick 2013; Brown and Mevs 2012; Gagnon 2010; Hess 2009). However, there has been little research on whether educational practitioners who work with performance assessments have the ability, or even confidence, to create, use, and interpret performance assessments. The present study aims to contribute to this body of research by reporting on the development and validation of a new scale to measure educational practitioners’ self-reported assessment literacy in performance assessment. Specifically, the study explores and confirms five unique dimensions of performance assessment literacy (PAL) captured by using this new survey scale. The aim of this scale is to provide multiple stakeholders in an assessment system with a way to evaluate their performance assessment abilities that is theory-driven, practitioner-informed, holistic, and reliable.

1 Performance assessments and performance assessment literacy

Performance assessments are referred to by many names (Linn et al. 1991; Koh 2011; Darling-Hammond and Pecheone 2009) and can be defined in different ways (Arter 2005). This study defines performance assessments as multistep assessments that involve creating or refining a product, typically involving open-ended questions, that promote use of higher-order thinking skills to address real-world problems (Frey and Schmitt 2007; Tung 2010). Examples of performance assessments include science labs; speeches, presentations, debates, and exhibitions; and artistic works and performances. Beginning in the 1990s, performance assessments became popular alternatives to standardized tests since they have been found to more authentically capture student learning (Goldschmidt et al. 2007; Meisels et al. 1995; Pellegrino et al. 2001) and
minimize bias (Goldschmidt et al. 2007; Stemler et al. 2009). Most research on performance assessments has focused on their value for teachers and students (Brown and Mednick 2013; Brown and Mevs 2012; Gagnon 2010), best practices for their construction and implementation (Stiggins 1987; Zane 2009; Johnson et al. 2008), and their psychometric properties (Dunbar et al. 1991; Gallant 2009; Meisels et al. 1995).

Like all assessments, performance assessments can range in their quality based on how they are designed, used, and interpreted. An individual’s understandings and abilities as they relate to “fundamental assessment concepts and procedures deemed likely to influence educational decisions” (Popham 2011, p. 267) are referred to as assessment literacy. These concepts include understanding validity, reliability, and test bias, as well as understanding testing procedures such as identifying, constructing, administering, scoring, interpreting, and using assessment data to promote student learning (Brookhart 2011; Popham 2004; Stiggins 1991). Assessment literacy is important because it benefits the many stakeholders in a balanced assessment system, such as students, teachers, family members, and the broader community of educational stakeholders (Henderson et al. 2007). Strong empirical evidence suggests links between teacher assessment literacy and improvements in both student achievement and metacognitive skills (Black and Wiliam 1998; Darling-Hammond et al. 2013; Popham 2011). Increased assessment literacy also gives teachers better evidence of whether students have learned the concepts that they are teaching, and those concepts that students will be evaluated on later using summative assessments (Popham 2009). Assessment literacy also helps teachers to ensure that their assessments are free from biases so that they can be confident that all their students are being assessed fairly (Popham 2009).

Improving the assessment literacy of administrators and policymakers could improve their interpretations of both classroom and standardized test results. This is integral for building arguments about school and teacher quality in the current accountability context and in determining the specific types of assessment literacy supports that would benefit teachers the most (Stiggins and Duke 2008). For families, increased assessment literacy can assist them in better evaluating the need and purpose of different student assessments. For example, assessment-literate families would understand the different roles and purposes of summative, formative, and benchmark assessment, and better understand the strengths and limitations of various types of assessments that teachers can use to assess student learning.

1.1 Assessment literacy measures

Since being highly assessment literate is theorized to have wide benefits for educational stakeholders, it is important to test those theories by empirically investigating the potential benefits of increased assessment literacy. The first step in that process is designing psychometrically valid measures of assessment literacy. In terms of measuring general assessment literacy, researchers have primarily evaluated this skillset through quantitative tests of assessment knowledge, self-reported measures, and rubrics (Gotch and French 2014). Self-reports are used most often because they are the easiest to administer, because they have been shown to correlate with actual assessment practices, while still helping educational
leaders and administrators to identify appropriate targeted interventions (DeLuca et al. 2016; Zhang and Burry-Stock 1997, 2003).

Yet, many measures of assessment literacy, including self-reports, have little psychometric support. In their systematic review of assessment literacy measures published between 1991 and 2012, Gotch and French (2014) found that most of the 36 included measures did not have content validity, score reliability, or consistent internal structure. Similarly, DeLuca et al. (2016) found that of the eight assessment literacy measures they studied, only five had acceptable internal consistency ratings (greater than 0.7). Additionally, they found that the alignment between observed themes from the standards on which the measures were based and the content of the measures was poor. For example, only two instruments considered assessment for learning (e.g., formative assessments) or addressed assessment fairness (e.g., consideration of student diversity and exceptional learners) which is a particularly common weakness among assessment literacy measures (Looney et al. 2017). The authors suggest that this is because most instruments are based on the 1990 Standards for Teacher Competence in the Educational Assessment of Students, which were developed prior to the widespread adoption of standard-based reform and formative assessments (Brookhart 2011). Many other measures of assessment literacy are also based on these outdated standards (Koloi-Keakitse 2016; Levy-Vered and Nasser-Abu Alhija 2015; Mertler 2009; Williams 2015).

Despite recognition of the importance of performance assessment, almost no research has been conducted on how to best measure teachers’ PAL (Popham 2009), that is, the ability of educational practitioners to create, use, and interpret performance assessments. Qualitative work in this area has considered how professional development interventions improve teachers’ conceptual understanding of the definition of performance assessments and skills, including choosing and designing appropriate assessments, recording information when observing students, and developing and using scoring rubrics (Koh 2011; Borko 1997; Borko et al. 1997). However, measuring PAL also should involve measuring whether practitioners are designing assessments that are fair for all students, including those students who are English Language Learners and students who have disabilities. Another important aspect of PAL that has recently grown in its importance for assessment use and development is incorporating students’ agency in the assessment experience by allowing them to personalize their assessment experience (Adie et al. 2018). Very little quantitative work has been conducted in PAL, and the existing literature focuses almost exclusively only on teachers’ PAL rather than teachers along with administrators and coaches (Popham 2009). In one study, teacher PAL was measured through an increase in the quality of assessment tasks (Koh 2011). In another, teacher PAL was merely one component of a six-factor measure of assessment skills (Zhang and Burry-Stock 2003). Like general assessment literacy instruments, this instrument was based on 1990 standards, and therefore did not consider the importance of teachers’ ability to develop performance assessments that are fair, are free of bias, address diverse learners’ needs, and encourage student agency.

In order to address the need for a measure of PAL that is both grounded in current assessment standards and statistically sound, this study asks the following research question: can a self-report PAL measure be developed for education practitioners that
measures multiple, theory-based dimensions of PAL while also providing practitioners with targeted, reliable information about their unique strengths and weaknesses in performance assessment?

1.2 Quality performance assessment framework

The theoretical framework that guides our study is the Quality Performance Assessment (QPA) Framework (Brown and Mednick 2013). Thousands of educational practitioners in more than a dozen states have been trained in the QPA Framework to guide their work. The framework considers important criteria of assessments that few previous measures have included, such as (1) having a valid design, (2) having reliable scoring mechanisms, and (3) promoting fairness as well as (4) requiring data analysis and (5) prioritizing student voice and choice. The framework theorizes that good performance assessments, in combination with local policy support and continued learning through professional development, can provide educational stakeholders with broad, but still targeted, information about student learning while allowing students to gain experience with real-world tasks and opportunities for applying their learning in complex ways. This framework provides us with an explicit theory of what constitutes high-quality performance assessments and serves as the foundation of our scale development and validation efforts.

2 Methods

2.1 The performance assessment literacy scale

A practitioner–researcher collaboration was the central organizing element in the creation of the PAL scale. As part of their internal evaluation process, the Research, Evaluation, and Policy team at the Center for Collaborative Education utilized the QPA Framework to guide their efforts in developing the PAL scale items. The process for item creation focused heavily on practitioner experience and need. To validate the content of our scale items, we interviewed teachers, coaches, and administrators who were exploring how to improve performance assessment skills. This bottom-up process ensured that the measure reflected the experiences of practitioners who use performance assessments in their daily work. After receiving input from these relevant PAL stakeholders, the authors designed items focused on assessing educator perceptions of the five dimensions that the QPA Framework identifies as essential components of performance assessment practices: valid design, reliable scoring, data analysis, fair assessment, and student voice and choice. Items were developed and revised by the research team as well as members of the performance assessment practice area at the Center for Collaborative Education iteratively until a final version of 27 items was agreed on (Table 1).

In our first sample, items were answered using a 5-point Likert-type scale: Not confident (0), Somewhat confident (1), Moderately confident (2), Confident (3), and Extremely confident (4). However, based on the results of the descriptive analysis of that sample, we hypothesized that there was a ceiling effect of the measure. Therefore, for the second and third samples, we added a sixth point to the upper end of the scale:
Completely confident (5). For all samples, the research team used Survey Monkey to design and distribute the PAL scale using participants’ email addresses.

### 2.2 Sample

Our study drew upon a large sample of educational professionals who planned to attend, and/or did attend, one of several performance assessment professional development workshops sponsored by the Center for Collaborative Education on designing

---

**Table 1** PAL scale items

| Item name | Item text |
|-----------|-----------|
| alv1      | Use backwards design/planning to organize my units and lessons |
| alv2      | Create assessments that are clearly aligned to Work-Study-Practices |
| alv3      | Create assessments designed to give students the opportunity to demonstrate high levels of cognitive rigor |
| alv4      | Identify student work products that can be used as exemplars for other students |
| alv5      | Create assessments that are clearly aligned with state competencies |
| alr1      | Calibrate scoring with colleagues using a common rubric |
| alr2      | Use a rubric to score student work |
| alr3      | Create rubrics that have clear criteria and descriptions of performance at each level. |
| alr4      | Create standardized rubrics for multiple assessments so my students can easily track their progress and growth from one assessment to the next |
| alr5      | Develop common rubrics with other educators |
| alr6      | Identify student work samples that can be used as anchors for scoring |
| alda1     | Create performance assessments that provide actionable feedback about your students’ learning |
| alda2     | Analyze and reflect on student assessment data on my own |
| alda3     | Discuss and interpret student assessment data with colleagues |
| alda4     | Modify instruction for students based on student assessment data |
| alda5     | Adjust instruction for particular groups of students based on student assessment data |
| alda6     | Personalize instruction for individual students based on student assessment data |
| alfl      | Design performance assessments that provide students with multiple pathways to demonstrate their knowledge |
| alf2      | Develop performance assessments that incorporate content on diverse cultures and traditions |
| alf3      | Incorporate accommodations into assessments for students with disabilities |
| alf4      | Design assessments that are free of stereotypes about cultural and linguistic groups |
| alsvc1    | Create performance assessments that focus on addressing authentic problems |
| alsvc2    | Design performance assessments that provide students with feedback to make decisions about their learning |
| alsvc3    | Design performance assessments that allow students to exercise ownership and decision making |
| alsvc4    | Develop performance assessments that provide students with opportunities to reflect on their learning |
| alsvc5    | Create performance assessments that allow students to set their own learning goals |
| alsvc6    | Develop assessments that promote an academic growth mindset (i.e., the belief that their academic abilities are not fixed and can improve with effort) |
performance assessments between 2015 and 2017 \((N=1080)\). The PAL scale was included as a part of larger surveys used to either collect information on participants prior to the professional development workshops to help inform the curriculum, or after the professional development workshops to collect program evaluation data. Participation in the evaluative aspects of the study was voluntary, and all data collection was approved by the New England IRB.

These participants were particularly appropriate for this investigation because they represent a wide range of PAL abilities. For example, novice educators who were surveyed prior to attending a professional development workshop on PAL could reasonably be expected to have low PAL. Contrarily, experienced educators who have already participated in the professional development workshop could reasonably be expected to have higher PAL.

We took this total sample and divided it into three subsamples: two for our replicability exploratory factor analysis (EFA) and one for our confirmatory factor analysis (CFA). To divide the sample, we first separated out the educational professionals that responded to the 5-point PAL scale to use as one of our samples for the replicability EFA \((N=275)\). We refer to this sample as EFA1. The sample was reduced after removing duplicate cases (e.g., participants who attended more than one professional development workshop between 2015 and 2017) based on participants’ email addresses \((N=242)\). We divided the remaining sample, which all answered the 6-point version of the scale, into two approximately equal samples based on the surveys that they responded to (i.e., participants who attended the same professional development were kept together). Our second EFA sample (EFA2) contained 314 respondents after removing duplicate records \((N=144)\), and our CFA sample contained 347 respondents.

Table 2 presents the demographic information across the three samples. As shown, most of the participants in each sample were white, female, and had at least a Master’s degree. The percentages for EFA1 sample and the CFA sample align closely with the latest US teacher demographic information provided by the National Center for Educational Statistics (2013, Table 209.10), although teachers with a Bachelor’s degree or less were slightly underrepresented. There was a higher amount of missing demographic information for the EFA2 sample, but that sample showed similar trends as it also included a high percentage of white female teachers with Master’s degrees. Importantly, EFA2 sample did not have the same degree of missing responses for the PAL items as it did for the demographic information.

### 2.3 Data analysis

**Missing data** Prior to conducting the exploratory factor analysis, we examined the missing data to ensure it did not violate the statistical assumptions of the analytic model. This was not necessary for the confirmatory factor analysis sample because we used Full Information Maximum Likelihood methods to estimate the missing cases, which works well for all data missing at random (Enders and Bandalos 2001). We found that both of our EFA samples contained missing data. In our first EFA sample, 22% of our cases contained at least one missing value on the PAL scale items. However, we conducted Little’s missing completely at random (MCAR) test (1988) and determined that we could not reject the null hypothesis that our missing data were not MCAR \((\chi^2 = 476.302, p = 0.663)\). Therefore, we chose to proceed using listwise
deletion and reduced our sample from 242 to 188 complete cases. In our second EFA sample, we found that 10% of cases were missing data from the PAL scale items and that the missing data was not MCAR (Little 1988). However, since our sample size was larger than the EFA1 sample, and the overall number of cases with missing values was particularly small (N = 33), we decided to proceed using the 314 complete cases under the untestable assumption that our data were missing at random.

### Table 2 Demographic information of survey sample

| Demographic                  | EFA1 sample | EFA2 sample | CFA sample |
|------------------------------|-------------|-------------|------------|
|                              | N   | %    | N   | %    | N   | %    |
| Gender                       |     |      |     |      |     |      |
| Female                       | 146 | 60.3%| 150 | 42.9%| 214 | 61.7%|
| Male                         | 43  | 17.8%| 46  | 13.1%| 55  | 15.9%|
| Transgender                  | 0   | 0.0% | 0   | 0.0% | 0   | 0.0% |
| Other                        | 0   | 0.0% | 0   | 0.0% | 0   | 0.0% |
| Missing                      | 53  | 21.9%| 154 | 44.0%| 78  | 22.5%|
| Race                         |     |      |     |      |     |      |
| White                        | 163 | 67.4%| 167 | 47.7%| 252 | 72.6%|
| Asian or Pacific Islander    | 6   | 2.5% | 3   | 0.9% | 0   | 0.0% |
| Latino, Hispanic, or Chicano | 5   | 2.1% | 11  | 3.1% | 5   | 1.4% |
| African-American or Black    | 5   | 2.1% | 5   | 1.4% | 3   | 0.9% |
| Other                        | 2   | 0.8% | 1   | 0.3% |      |      |
| Multiracial                  | 1   | 0.4% | 5   | 1.4% | 5   | 1.4% |
| American Indian or Native Alaskan | 1 | 0.4% | 0   | 0.0% | 0   | 0.0% |
| Missing                      | 59  | 24.4%| 158 | 45.1%| 82  | 23.6%|
| Degree                       |     |      |     |      |     |      |
| High school graduate         | 0   | 0.0% | 0   | 0.0% | 1   | 0.3% |
| Some college                 | 0   | 0.0% | 0   | 0.0% | 1   | 0.3% |
| Associate’s degree           | 1   | 0.4% | 1   | 0.3% | 3   | 0.9% |
| Bachelor’s degree            | 48  | 19.8%| 15  | 4.3% | 37  | 10.7%|
| Master’s degree              | 110 | 45.5%| 146 | 41.7%| 197 | 56.8%|
| Post-Masters credits or certification | 15 | 6.2% | 18  | 5.1% | 32  | 9.2% |
| Professional school          | -   | -    | 1   | 0.3% | 1   | 0.3% |
| Doctoral degree              | 16  | 6.6% | 11  | 3.1% | 2   | 0.6% |
| Other                        | 1   | 0.4% | 1   | 0.3% | 0   | 0.0% |
| Missing                      | 51  | 21.1%| 157 | 44.9%| 73  | 21.0%|

For our EFA, we specified the use of Principal Axis Factoring as the extraction method. First, we conducted an EFA on each subscale using only the items of that subscale and extracted one factor. We then used multiple sources of data to determine alternative plausible factor solutions, including the Kaiser rule (1960), a scree test (Cattell 1966), the percentage of variance extracted by each factor, and a parallel analysis (Watson 2017) with 27 simulated variables drawn from a normal
distribution with mean 0 and variance 1. Afterwards, we conducted an EFA on all items of the PAL scale, specifying the extraction of five factors based on our theory that the items of each of our PAL subscales are elicited by a unique latent dimension of PAL: validity, reliability, data analysis, fairness, and student voice and choice. We used the same tests (e.g., Kaiser rule, scree test) to determine whether additional factor solutions were plausible. Since we expected that any multiple factor solutions, including our original 5-factor solution, would have correlated factors, we specified a Promax rotation (Kappa = 4).

After determining which factor models were plausible for both of our EFA samples, we compared the models using a replicability analysis (Osborne 2014). This involved comparing the standardized factor loadings of each model across our two samples in two ways. The first was analyzing whether the strongest factor loading for each item was on the same factor across samples, which is referred to as basic factor structure replicability. Second, we calculated the squared differences between each item that had basic factor structure replicability to assess the extent that the factor loading changed from one sample to the next. This is referred to as strong factor structure replicability, where values above 0.04 are considered problematic.

Finally, we conducted a CFA to statistically test which of the models discovered in the replicability analysis best fit our data. CFA differs from EFA in that the researchers must specify the factor model prior to conducting the analysis, whereas an EFA is determined without prior model specification based on the correlations between variables. Unlike EFA, CFA can be used to test whether a proposed model is a significantly better fit of the data compared to other models. Therefore, researchers can be more confident in claims regarding which measurement model is more appropriate for modeling their data (Hair Jr. et al. 2010).

3 Results

3.1 Exploratory factor analysis

We conducted an EFA for each of our two EFA samples specifying the extraction of five factors. The correlation matrices for EFA1 sample (determinant = 6.805*10^{-12}; KMO = 0.953; \( \chi^2 = 4556, p \leq 0.001 \)) and EFA2 sample (determinant = 5.137*10^{-12}; KMO = 0.958; \( \chi^2 = 7881, p \leq 0.001 \)) were determined to be appropriate for factoring. In EFA1 sample, the first five factors explained 53%, 6%, 6%, 3%, and 2% of the shared variance among the items, respectively. In EFA2 sample, the first five factors explained 54%, 8%, 4%, 3%, and 2% of the shared variance among the items, respectively. The scree plot for both solutions (Fig. 1) provides clear evidence for a 1-factor model. In the 5-point scale sample, there was also an abnormality in the scree plot around the third factor, and in the 6-point scale sample, there was an abnormality around the second factor. A parallel analysis (Fig. 1) confirmed that both the second and third factors of the EFA1 sample explained more shared variance among the 27 PAL items than among 27 random variables, but only the second and third factors explained more shared variance among the items in the EFA2 sample. Therefore, this plot could also be providing evidence for a 2-factor model of correlated factors. Based
on these statistics, we decided to examine the factor matrices for each of our samples using 1-factor, 2-factor, 3-factor, and 5-factor models.

We examined the factor matrices simultaneously across samples using basic and strong replicability analyses (Osborne 2014). We found the 1-factor model to have basic replicability since all items loaded highly on that factor in the 5-point sample (EFA1) and 6-point sample (EFA2). For the 3-factor model, two items did not pass the test of basic replicability (alda2 and alda3), while for the 5-factor model, five items did not pass the test of basic replicability (alda1, alda2, alf3, alv2, and alv5). The 2-factor model stood out as having failed the test of basic replicability because the entire Reliability subscale did not pass the test, as well as two additional items from the Validity subscale (alv4 and alv5). After assessing the basic replicability of the factor loading matrices, we assessed the strong replicability of those matrices, excluding those items in each model that did not pass the basic replicability test. The results were that the 2-factor and 3-factor models did not pass the test of strong replicability. In the 2-factor model, item alda5 had the highest change in its rotated factor loading (1.039 to 0.785), followed by alv1 (0.616 to 0.380), and then alda3 (0.584 to 0.787). The 3-factor model also had three items that did not pass the test of strong replicability: items alsvc6 (1.026 to 0.762), alv4 (0.484 to 0.742), and alr6 (0.888).

Thus, it appeared from this analysis that the 1-factor model was the most plausible. Even using the replicability methods, it was difficult to compare the plausibility of the 3-factor and 5-factor model. On the one hand, the 3-factor model had a higher percentage of items that passed the basic replicability test than the 5-factor model. On the other hand, the 5-factor model had a higher percentage of items that passed the strong replicability test than the 3-factor model. The 2-factor model was the least replicable across samples. Therefore, we chose to proceed by studying the fit of the 1-factor, 3-factor, and 5-factor models in a confirmatory factor analysis.
3.2 Confirmatory factor analysis

We created three CFA models based on the 1-factor, 3-factor hierarchical, and 5-factor hierarchical models discovered in the EFA. For the 1-factor model, every item was specified to load onto that factor. For the 3-factor model, the items were divided based on which factor they loaded the highest on in the replicability analysis. The two items that did not pass the basic test of replicability on the 3-factor model were specified to load onto the data analysis literacy factor since they were both qualitative about analyzing student data. A higher-order factor was also specified called PAL. For the 5-factor model, items were specified to load onto the five factors that they were designed to load onto. A higher-order factor was also specified called PAL.

Table 3 summarizes the overall model fit statistics for each of the three CFA models. As shown, the 5-factor model has the highest CFI statistic, and the lowest AIC, BIC, RMSEA, and SRMR statistics of the three models. This provides evidence that the 5-factor hierarchical model has the best fit of all three models explored in the EFA replicability analysis. Therefore, we determined that the answer to our first research question was that the best model for performance literacy was a 5-factor hierarchical model. For this solution, the items loaded as we hypothesized, so we named the five factors student voice and choice, reliability, data analysis, validity, and fairness, respectively.

3.3 Item fit

Although our research determined that the 5-factor model was the best fitting of the three tested, it still did not fit our data as well as we would have hoped. Therefore, we attempted to improve the fit by identifying potentially problematic items and removing them from the model. To identify these items, we considered a variety of statistics, including the item–total correlations, the item–subscales correlations, alpha-if-deleted values, and highest factor loadings. We also considered statistics specific to our EFA replicability analyses, including the extracted communalities, whether the item passed the basic replicability test, and the squared difference of the factor loadings used in the assessment of strong replicability.

The results are summarized in Table 4. As shown, most of these statistics for the items were very high. However, one item stood out as being particularly problematic, which was item alda1. Across samples, this item had a lower alpha-if-deleted value than its Cronbach’s alpha value, indicating that removing the item would make the data

| Table 3 | Model fit statistics for 1-, 3-, and 5-factor CFA models |
|---------|----------------------------------------------------------|
| Fit statistic | 1-factor model | 3-factor model | 5-factor model |
| CFI | 0.71 | 0.81 | 0.89 |
| AIC | 20,563 | 19,665 | 18,999 |
| BIC | 20,868 | 19,982 | 19,323 |
| RMSEA | 0.16 | 0.13 | 0.10 |
| SRMR | 0.08 | 0.09 | 0.07 |
analysis subscale more reliable. Furthermore, it had low factor loadings in the two EFA samples and did not pass the test of basic replicability. Therefore, we removed the item and re-ran the overall model fit statistics. The CFI statistic was at an acceptable standard after removing this item (0.91), and the AIC (18204), BIC (18517), RMSEA (0.09), and SRMR (0.05) statistics all decreased. Both findings point to better model fit after removing alda1, which in turn affirms that alda1 was a problematic item. Our final model, without item alda1, is illustrated with path estimates in Fig. 2.

4 Discussion

The recent growth in the use of performance assessments aims to improve assessment systems by using multiple assessment types to inform decisions (Wylie and Lyon 2017). However, it is important to ensure that every assessment type in these systems is of high quality. First, it is important for policymakers, researchers, and practitioners to have a shared understanding around what “high-quality” assessment means. Then, it is important to create systems where assessment stakeholders are trained on how high-quality assessments can be validly designed, used, and interpreted, thereby improving the assessment literacy of assessment stakeholders. This study adds to the existing literature on assessment literacy as it specifically relates to performance assessments by providing a tool for the self-assessment of educational practitioners’ confidence in designing, using, and interpreting performance assessments. This measure is based on the QPA framework’s definition of high-quality performance assessment. Teachers, coaches, and educational administrators often lack access to research-based tools, particularly tools that are both rooted in an explicit theory of assessment and also created with practitioner input. The PAL scale of this study can be added to the toolbox of practitioners interested in improving five areas of the PAL of their teams: valid design, reliable scoring, data analysis, fair assessment, and student voice and choice. Although the PAL needs further refinement and research to support its validity, this study takes the first steps in establishing a psychometrically evaluated measure that supports the provision of feedback to educational practitioners on their PAL. This feedback is a key ingredient in improving the quality of these practitioners’ performance assessments. For example, suppose a coach is looking to improve the performance assessment knowledge of their staff. This coach could begin by using the PAL scale to assess the staff on where they are most and least confident in their PAL, then design a professional development curriculum around the areas where staff have the lowest confidence according to the scale. After providing training on those areas, the coach could reassess teachers’ PAL using the scale. This creates an opportunity for a focused, individualized discussion between coach and teacher on the same five areas that the teacher was assessed on originally regarding where the teacher has grown and where they still have room to improve post-intervention. While the conversation between coach and teacher is likely the most valuable aspect of the intervention, the PAL scale’s five dimensions of assessment literacy can provide the basis and data for informing those conversations. Over time, coaches could use this data in aggregate to explore trends in teachers’ self-reported PAL or estimate the impact of certain PAL interventions.
| Corrected item–total R | Corrected item–subscale R | Subscale alpha-if-deleted | Highest factor loading | Communaliities |
|------------------------|--------------------------|---------------------------|------------------------|---------------|
| Sample                 | Sample                   | Sample                    | Sample                 | Sample        |
| 1                      | 2                        | 3                         | 1                      | 2             | 3             |
| alv1                   | 0.605                    | 0.625                     | 0.562                  | 0.651         | 0.666         | 0.647         | 0.857         | 0.842         | 0.877         | 0.754         | 0.752         | 0.684         | 0.546         | 0.576         | 0.468         |
| alv2                   | 0.705                    | 0.739                     | 0.732                  | 0.725         | 0.691         | 0.746         | 0.839         | 0.836         | 0.853         | 0.628         | 0.420         | 0.823         | 0.618         | 0.611         | 0.677         |
| alv3                   | 0.797                    | 0.724                     | 0.749                  | 0.829         | 0.721         | 0.802         | 0.812         | 0.828         | 0.841         | 0.797         | 0.642         | 0.872         | 0.803         | 0.702         | 0.760         |
| alv4                   | 0.668                    | 0.680                     | 0.670                  | 0.604         | 0.711         | 0.714         | 0.867         | 0.832         | 0.861         | 0.412         | 0.659         | 0.754         | 0.508         | 0.655         | 0.569         |
| alv5                   | 0.686                    | 0.603                     | 0.657                  | 0.702         | 0.651         | 0.705         | 0.845         | 0.846         | 0.863         | 0.654         | 0.597         | 0.745         | 0.588         | 0.471         | 0.555         |
| alr1                   | 0.577                    | 0.651                     | 0.720                  | 0.671         | 0.767         | 0.826         | 0.889         | 0.909         | 0.931         | 0.814         | 0.885         | 0.850         | 0.555         | 0.685         | 0.723         |
| alr2                   | 0.566                    | 0.685                     | 0.687                  | 0.684         | 0.801         | 0.825         | 0.888         | 0.905         | 0.931         | 0.793         | 0.807         | 0.840         | 0.609         | 0.729         | 0.706         |
| alr3                   | 0.700                    | 0.723                     | 0.743                  | 0.774         | 0.727         | 0.813         | 0.872         | 0.914         | 0.932         | 0.656         | 0.541         | 0.836         | 0.690         | 0.618         | 0.699         |
| alr4                   | 0.735                    | 0.714                     | 0.752                  | 0.787         | 0.766         | 0.804         | 0.871         | 0.909         | 0.934         | 0.647         | 0.727         | 0.842         | 0.695         | 0.648         | 0.709         |
| alr5                   | 0.669                    | 0.721                     | 0.749                  | 0.760         | 0.813         | 0.881         | 0.875         | 0.903         | 0.924         | 0.724         | 0.779         | 0.903         | 0.634         | 0.721         | 0.815         |
| alr6                   | 0.683                    | 0.707                     | 0.750                  | 0.697         | 0.791         | 0.806         | 0.884         | 0.906         | 0.933         | 0.720         | 0.765         | 0.834         | 0.628         | 0.750         | 0.696         |
| alda1                  | 0.817                    | 0.806                     | 0.834                  | 0.654         | 0.645         | 0.697         | 0.927         | 0.926         | 0.947         | 0.431         | 0.464         | 0.716         | 0.713         | 0.703         | 0.513         |
| alda2                  | 0.763                    | 0.744                     | 0.774                  | 0.799         | 0.789         | 0.824         | 0.905         | 0.906         | 0.931         | 0.531         | 0.488         | 0.830         | 0.699         | 0.688         | 0.689         |
| alda3                  | 0.718                    | 0.720                     | 0.760                  | 0.765         | 0.789         | 0.853         | 0.910         | 0.906         | 0.928         | 0.510         | 0.486         | 0.864         | 0.641         | 0.678         | 0.746         |
| alda4                  | 0.663                    | 0.711                     | 0.795                  | 0.851         | 0.845         | 0.887         | 0.898         | 0.899         | 0.924         | 0.969         | 0.834         | 0.929         | 0.880         | 0.814         | 0.863         |
| alda5                  | 0.650                    | 0.705                     | 0.732                  | 0.844         | 0.837         | 0.868         | 0.899         | 0.900         | 0.925         | 0.967         | 0.957         | 0.917         | 0.867         | 0.907         | 0.841         |
| alda6                  | 0.662                    | 0.693                     | 0.710                  | 0.777         | 0.768         | 0.827         | 0.908         | 0.909         | 0.931         | 0.804         | 0.762         | 0.884         | 0.768         | 0.740         | 0.781         |
| alf1                   | 0.793                    | 0.772                     | 0.790                  | 0.749         | 0.690         | 0.773         | 0.827         | 0.788         | 0.799         | 0.450         | 0.498         | 0.871         | 0.718         | 0.725         | 0.759         |
| alf2                   | 0.663                    | 0.703                     | 0.727                  | 0.709         | 0.760         | 0.743         | 0.844         | 0.756         | 0.810         | 0.492         | 0.625         | 0.819         | 0.590         | 0.757         | 0.671         |
| alf3                   | 0.716                    | 0.591                     | 0.683                  | 0.722         | 0.563         | 0.638         | 0.838         | 0.841         | 0.852         | 0.607         | 0.547         | 0.726         | 0.709         | 0.538         | 0.527         |
|          | Corrected item–total R | Corrected item–subscale R | Subscale alpha-if-deleted | Highest factor loading | Communalities |
|----------|------------------------|--------------------------|---------------------------|-----------------------|---------------|
|          | Sample                 | Sample                   | Sample                    | Sample                | Sample        |
| 1        | 2                      | 3                        | 1                         | 2                     | 3             |
| alf4     | 0.660                  | 0.613                    | 0.660                     | 0.835                 | 0.793         | 0.833         | 0.642 | 0.663 | 0.729 | 0.598 | 0.604 | 0.531 |
| alsvc1   | 0.754                  | 0.757                    | 0.796                     | 0.876                 | 0.934         | 0.938         | 0.942 | 0.817 | 0.877 | 0.937 | 0.824 | 0.812 | 0.878 |
| alsvc2   | 0.824                  | 0.789                    | 0.815                     | 0.933                 | 0.937         | 0.947         | 0.820 | 0.983 | 0.896 | 0.813 | 0.844 | 0.803 |
| alsvc3   | 0.811                  | 0.757                    | 0.762                     | 0.886                 | 0.885         | 0.871         | 0.976 | 0.985 | 0.910 | 0.835 | 0.843 | 0.828 |
| alsvc4   | 0.792                  | 0.784                    | 0.837                     | 0.868                 | 0.895         | 0.880         | 0.964 | 0.878 | 0.858 | 0.785 | 0.766 | 0.736 |
| alsvc5   | 0.749                  | 0.763                    | 0.774                     | 0.833                 | 0.836         | 0.840         | 0.721 | 0.753 | 0.836 | 0.778 | 0.707 | 0.699 |
| alsvc6   | 0.763                  | 0.771                    | 0.761                     | 0.830                 | 0.813         | 0.816         | 0.939 | 0.946 | 0.953 | 0.711 | 0.696 | 0.689 |

*Alpha-if-deleted value exceeds subscale Cronbach’s alpha value
Importantly, this scale does not have to be used inside the coaching context. For example, suppose that a teacher is interested in developing their performance assessment skills. They might begin their search for professional development opportunities related to performance assessment by responding to the PAL survey scale, then by examining their responses to the scale items to identify specific skills in which they could use the most improvement. They might also choose to look at their results in aggregate across the five sets of skills based on the five dimensions of PAL explored and confirmed by this study to identify a general area of PAL where they could most improve rather than a specific PAL skill.

In addition to providing a tool for practitioners, this study provides rigorous quantitative research support of a PAL measure based on a widely used professional development framework. Educators at the school, district, and state levels can adopt the QPA framework and accompanying processes, protocols, and tools with greater confidence of the soundness of the alignment between the theory and this survey measure. Since student assessment continues to be at the center of education policy debates, teacher assessment literacy will also continue to be an important element of professional development and learning. This study has provided one answer to the various
calls for a more rigorous approach for scale development as it relates to assessment literacy (DeLuca et al. 2016; Gotch and French 2014; Looney et al. 2017).

4.1 Limitations and future directions

While this study provides important advancements to understanding assessment literacy specifically in relation to performance assessments, there were limitations associated with this research. The first limitation is that although the teachers in our sample were relatively similar in their demographics as the teacher population in the USA, recruiting a more diverse sample of teachers in terms of gender, race, and educational background could provide different item fit statistics for the items of the scale. In particular, our sample underrepresented teachers with a Bachelor’s degree or less. These teachers would have taken fewer education courses than those teachers with Master’s degrees. Therefore, teachers with a Bachelor’s degree or less may have received less training around classroom assessment use and design in general, and may be less familiar with the language used in our scale items. The lack of familiarity could worsen the psychometric properties of the scale by an unknown magnitude if more teachers with a Bachelor’s degree or less had been included. A second limitation of this study is how we identified duplicate records using participants’ email addresses. For example, not all participants provided an email address in the EFA1 sample, which could mean that duplicate records still existed in that sample. Although the fit of the items was comparable between the EFA1 sample with potential duplicates and the EFA2 sample where all participants provided email addresses, we cannot be sure that there are no duplicates in the sample. A third limitation is that the type of school a teacher works in might influence their confidence in performance assessments. For example, some schools might have structures that are more rigid, have less distributed leadership, or experience other factors (e.g., title I, charter status) that influence how teachers think about assessments or the flexibility they have in developing them. Though our sample includes a wide range of schools, we could not test for differences by schools as that data was not collected, and future studies should explore whether these differences influence participant responses to the PAL scale items.

While our study provided evidence for the dimensionality of educational practitioners’ self-perceived assessment literacy skills, future research is needed on corroborating those self-perceptions with direct evidence measures of PAL. For example, a future study might develop a rubric measure for PAL that could be used to rate performance assessments developed during the academic year or during professional development workshops using the five dimensions of PAL that were confirmed by the current study as rubric criteria. Alternatively, test items could be developed to measure teachers’ knowledge and understanding of how to design, use, or interpret performance assessments based on the five dimensions of PAL confirmed by this study. These alternative types of measures could be used in conjunction with our survey measure to validate whether teachers’ confidence in their performance assessment abilities relates to evaluations of those abilities.

Focusing on the limitations specific to self-perception measures of practitioners’ literacy in performance assessments, two areas of research are particularly important in terms of enhancing evidence around the PAL scale’s validity. The first is locating other measures associated with assessment literacy and correlating those measures with PAL.
scale scores. Measures that theoretically should be highly related to PAL should yield high correlations with our scale, and thereby provide evidence of convergent validity. For example, a variable measuring the number of assessment-specific professional development opportunities that a practitioner has attended could be used to explore the convergent validity of our scale with the theory that teachers who attend more professional development in assessment should be more confident in their PAL. A second direction of future research specific to self-perceptions of practitioners’ assessment literacy could explore our PAL scale data again but using item–response theory (IRT) methods (e.g., Rasch analysis). These methods would provide a second framework to evaluate the fit of the scale items.

References

Adie, L. E., Willis, J., & Van der Kleij, F. M. (2018). Diverse perspectives on student agency in classroom assessment. *Australian Educational Researcher, 45*(1), 1–12. https://doi.org/10.1007/s13384-018-0262-2.

Arter, J. (2005). Teaching about performance assessment. *Educational Measurement: Issues and Practice, 18*(2), 30–44. https://doi.org/10.1111/j.1745-3992.1999.tb00012.x.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice, 5*(1), 7–74.

Borko, H. (1997). New forms of classroom assessment: implications for staff development. *Theory Into Practice, 36*(4), 231–238. https://doi.org/10.1080/00405849709543773.

Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers’ developing ideas and practices about mathematics performance assessment: successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education, 13*(3), 259–278. https://doi.org/10.1016/s0742-051x(96)00024-8.

Brookhart, S. (2011). Educational assessment knowledge and skills for teachers: Spring 2011. *Educational Measurement: Issues and Practice, 30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x.

Brown, C. & Mednick, A. (2013). Quality performance assessment: a guide for schools and districts. Boston, MA: The Center for Collaborative Education.

Brown, C. & Mevs, P. (2012). Quality performance assessment: harnessing the power of teacher and student learning. Boston, MA: The Center for Collaborative Education.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276. https://doi.org/10.1207/s15327906mbvr0102_10.

Darling-Hammond, L., Newton, S. P., & Wei, R. C. (2013). Developing and assessing beginning teacher effectiveness: the potential of performance assessments. *Educational Assessment, Evaluation and Accountability, 25*, 179–204. https://doi.org/10.1007/s11092-013-9163-0.

Darling-Hammond, L., & Pecheone, R. (2009). Reframing accountability: using performance assessments to focus learning on higher-order skills. In *Meaningful measurement: the role of assessments in improving high school education in the twenty-first century*. Washington DC: Alliance for Excellent Education.

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/10.1080/10654915.2015.1052712.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303. https://doi.org/10.1207/s15324818ame0404_3.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430–457. https://doi.org/10.1207/s15323880sem0803_5.

Every Student Succeeds Act of 2015, Pub. L. 114-195.

Frey, B. B., & Schmitt, V. L. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics, 18*(3), 402–423. https://doi.org/10.4219/jaa-2007-495.

Gagnon, L. (2010). Ready for the future: the role of performance assessments in shaping graduates’ academic, professional, and personal lives. Boston: The Center for Collaborative Education.
Gallant, D. J. (2009). Predictive validity evidence for an assessment program based on the Work Sampling System in mathematics and language and literacy. *Early Child Research Quarterly, 24*(2), 133–141. https://doi.org/10.1016/j.ecresq.2009.03.003.

Goldschmidt, P., Martinez, J. F., Niemi, D., & Baker, E. L. (2007). Relationships among measures as empirical evidence of validity: incorporating multiple indicators of achievement and school context. *Educational Assessment, 12*(3–4), 239–266.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice, 33*(2), 14–18. https://doi.org/10.1111/emip.12030.

Hair Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *SEM: an introduction. Multivariate data analysis: a global perspective* (pp. 629–686).

Hess, K. (2009). Local assessment toolkit: assessment task validation tools-criteria for high quality performance assessment.

Hoffman, P., Goodwin, B., & Kahl, S. (2015). *Re-balancing assessment: placing formative and performance assessment at the heart of learning and accountability*. Denver, CO: MeREL International.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). Assessing performance: designing, scoring, and validating performance tasks. New York: Guilford Press.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141–151. https://doi.org/10.1177/001649086002000116.

Koh, K. H. (2011). Improving teachers’ assessment literacy through professional development. *Teaching Education, 22*(3), 255–276. https://doi.org/10.1080/10476210.2011.593164.

Koloi-Kealikite, S. (2016). Assessment training: a precondition for teachers’ competencies and use of classroom assessment practices. *International Journal of Training and Development, 20*(2), 107–123. https://doi.org/10.1111/ijtd.12072.

Levy-Vered, A., & Nasser-Abu Alhija, F. (2015). Modelling beginning teachers’ assessment literacy: the contribution of training, self-efficacy, and conceptions of assessment. *Educational Research and Evaluation, 21*(5–6), 378–406. https://doi.org/10.1080/13803611.2015.1117980.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Research, 20*(8), 15. https://doi.org/10.2307/1176232.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202. https://doi.org/10.2307/2290157.

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. (2017). Reconceptualising the role of teachers as assessors: teacher assessment identity. *Assessment in Education: Principles, Policy & Practice, 25*(5), 442–467. https://doi.org/10.1080/0969594X.2016.1268090.

Meisels, S. J., Dorfman, A., & Steele, D. (1995). Equity and excellence in group-administered and performance-based assessments. In M. T. Nettles & A. L. Nettles (Eds.), *Equity and excellence in performance-based assessments. In M. T. Nettles & A. L. Nettles (Eds.),* *Equity and excellence in educational testing and assessment* (pp. 221–242). Boston: Kluwer Academic Publishers.

Mertler, S. C. (2009). Teachers’ assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(2), 101–113. https://doi.org/10.1111/1365480209105575.

National Center for Educational Statistics (2013). Number and percentage distribution of teachers in public and private elementary and secondary schools, by selected teacher characteristics: selected years, 1987-88 through 2011-12 [Table]. *Digest of Education Statistics.*

Osborne, J. W. (2014). *Best practices in exploratory factor analysis*. South Carolina: Createspace Publishing.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: the science and design of educational assessment*. Washington DC: National Academy Press.

Popham, W. J. (2004). *America’s “failing” schools: how parents and teachers can cope with No Child Left Behind*. New York: Routledge Falmer.

Popham, W. J. (2009). Assessment literacy for teachers: faddish or fundamental? *Theory Into Practice, 48*(1), 4–11. https://doi.org/10.1080/00405840802577536.

Popham, W. J. (2011). Assessment literacy overlooked: a teacher educator’s confession. *Teaching Education, 46*(4), 265–273. https://doi.org/10.1080/08878730.2011.605048.

Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., Jarvin, L., & Sharpe, K. (2009). Using the theory of successful intelligence as a framework for developing assessments in AP physics. *Contemporary Educational Psychology, 34*(3), 195–209. https://doi.org/10.1016/j.cedpsych.2009.04.001.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33–42. https://doi.org/10.1111/j.1745-3992.1987.tb00507.x.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534–539.
Stiggins, R. J., & Duke, D. (2008). Effective instructional leadership requires assessment leadership. *Phi Delta Kappan, 90*(4), 285–291.

Tung, R. (2010). Including performance assessments in accountability systems: a review of scale-up efforts. Boston: The Center for Collaborative Education.

Watson, J. C. (2017). Establishing evidence for internal structure using exploratory factor analysis. *Measurement and Evaluation in Counseling and Development, 50*, 232–238. https://doi.org/10.1080/07481756.2017.1336931.

Williams, J. C. (2015). “Assessing without levels”: preliminary research on assessment literacy in one primary school. *Educational Studies, 41*(3), 341–346. https://doi.org/10.1080/03055698.2015.1007926.

Wylie, C., & Lyon, C. (2017). Supporting teacher assessment literacy: a proposed sequence of learning. Teachers College Record. https://www.ets.org/s/winsight/pdf/balanced-assessment-article-tcr-ada.pdf

Zane, T. W. (2009). Performance assessment design principles gleaned from constructivist learning theory (part 1). *TechTrends, 53*(1), 81–90. https://doi.org/10.1007/s11528-009-0242-5.

Zhang, Z., & Burry-Stock, J.A. (1997). Assessment practices inventory: a multivariate analysis of teachers’ perceived assessment competency. In Annual meeting of the American Educational Research Association, Chicago.

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers’ self-perceived assessment skills. *Applied Measurement in Education, 16*, 323–342. https://doi.org/10.1207/S15324818AME1604_4.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.