Research Article

Received: 25 July 2023 / Accepted: 13 October 2023 / Published: 5 November 2023

Teacher Preferences for Formative Assessment: Leveraging Findings for Future Professional Development Resources

Leticja Gusho[1*]

Cleti Cervoni[2]

Allan Shwedel[2]

[1]*Pedagogy-Psychology Department,*
Tirana University,
Tirana, Albania
[2]*Secondary and Higher Education,*
Salem State University, Salem,
MA, United States
** Corresponding Author*

DOI: https://doi.org/10.36941/jesr-2023-0145

Abstract

Classroom-based formative assessment is invoked as an essential component of effective teaching, but there is no widely agreed-upon approach for targeted professional development. This survey of secondary school teachers examined 3 issues: (a) can McTighe and Ferrara's assessment framework provide a psychometrically tenable way to classify formative assessment techniques, (b) does their framework offer a viable approach to professional development; and (c) could their framework be applied in strengths-based professional development. Analyses of the teachers' self-reported skills showed (a) high levels of internal consistency across categories of formative assessment, (b) STEM and Humanities teachers could be grouped via discriminant analysis based on reported levels of skill in 3 categories of formative assessment, and (c) scenarios are presented to demonstrating a strength-based approach. The findings indicate that McTighe and Ferrara's framework provides a psychometrically tenable way to categorize formative assessment techniques and provide professional development. Future research should include K-6 teachers.

Keywords: *professional development, statistical methods, formative assessment*

1. Introduction

"Formative assessment is based on the assumption that the outcome of the instruction cannot be predicted" (Anderson, 2020, p. 76).

Teaching would be a very different profession if Andersson, a Special Education and Mathematics researcher from Sweden, was wrong about a teacher's ability to create lessons or even utilize the latest computer-based lessons so that all students would master all the material all the time. Fortunately, given the indeterminacy of outcomes from pre-planned instruction, there is a myriad array of tools, techniques, and tests that teachers and students can use to monitor learning to

make real-time adjustments in the teaching/learning process. "But how is a teacher supposed to decide what to use and how to use it effectively?" The classroom assessment encompasses various methods, ranging from traditional grading to standardized tests. Among these methods, formative assessment (FA) has garnered significant attention from teachers due to its multiple benefits. Since 1998, Black and Wiliam have been emphasizing the importance of formative assessment, including high-quality feedback, to help students improve and accelerate the learning process (Black & Wiliam, 1998). Stiggins et al. (2006) have aptly described formative assessment (FA) as an "assessment for learning", in contrast with the summative assessment which is an "assessment of learning" (Stiggins, Arter, Chppuis, & Chappuis, 2006). Popham (2014), provides a more specific definition of FA, i.e., "formative assessment is a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics" (p. 290). Furthermore, FA is a vital process that not only motivates students (Leenknecht, et al., 2020), but also focuses on enhancing their learning outcomes (Angelo & Cross, 2012; Irons & Elkington, 2021; Moyo, Combrinck, & Staden, 2022). It serves as a valuable feedback mechanism, enabling teachers to adjust their instructional methods and improve student achievement (Schneider & Johnson, 2018). The successful implementation of FA supports students' learning outcomes and provides them with constructive feedback (Cauley & McMillan, 2010).

While many studies have examined different FA practices in recent years, there is still a lack of robust research on how to effectively support teachers in implementing such practices (McMillan, Venable, & Varier, 2013; Andersson & Palm, 2018). Moreover, the concept of FA itself is not yet well-defined (Bennett, 2011). However, it is important to recognize that teachers who possess strong competencies in employing FA practices can greatly help their students. By involving students as active participants in the learning process, FA promotes self-regulated (Beesley, Clark, Dempsey, & Tweed, 2018; Pajares & Graham, 1999). Through FA, students are empowered to assess themselves and receive timely feedback to enhance their progress. Frequent feedback has been shown to significantly improve learning outcomes (Lee, Chung, Zhang, Abedi, & Warschauer, 2020). By the same token, multiple studies emphasize the value of integrating FA into daily instructional learning skills and improving the quality of classroom interactions (Nicol & Macfarlane-Dick, 2006). It has been found that FA when combined with motivational factors, influences students' performance in specific subjects like mathematics, recognizing it as a valuable approach that benefits both teachers and students. A recent publication by the U. S. Council of Chief State School Officers (CCSSO) presents a strong case for the importance of implementing inclusive formative assessments for students with Special Needs (Brookhart & Lazarus, 2017). In this light, formative assessment plays a pivotal role in the classroom by motivating students, enhancing their learning experience (Weurlander, Söderberg, Scheja, Hult, & Wernerson, 2012), and enabling teachers to adjust their instructional strategies across a wide range of subject areas and student needs (Darrow, 2015; Dudec, Reddy, Lekwa, Hua, & Fabiano, 2019; Frey & Fisher, 2013; Quigley, 2020; Wylie & Lyon, Developing Formative Assessment Protocol to Support Professional Growth, 2020).

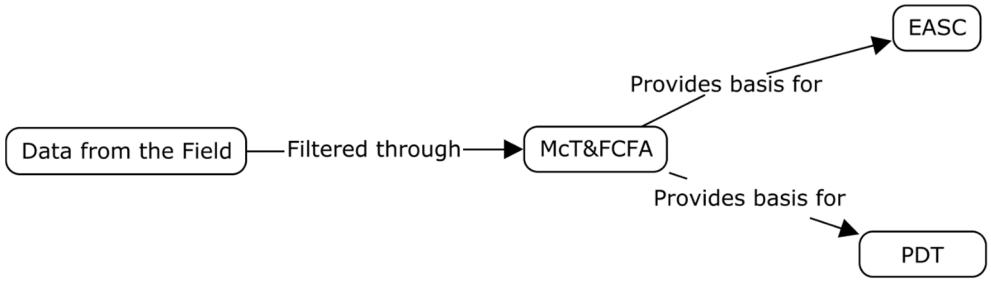
Based on the research conducted for this article, the existing literature highlights various benefits of formative assessment (FA) in the classroom. However, despite the extensive research on FA, there remains a gap in understanding how continuing professional development (CPD) can effectively support teachers in successfully implementing various forms of formative assessment. While the literature provides ample evidence of the positive impact of FA, there is limited research exploring the specific role of CPD in facilitating the successful application of formative assessment strategies in the classroom.

This research study examines the feasibility of categorizing assessment techniques and embedding those categories within a structured approach to strength-based professional development (He, 2009; Zwart, Korthagen, & Attema-Noordewier, 2014) to advance our understanding of ways to support teachers in the acquisition and use of FA skills. In this light we explore the feasibility and usefulness of applying McTighe and Ferrara's (2020) conceptual framework for assessment as the context and structure for strength-based formative assessment professional

development. The authors have identified an array of 46 tasks and activities that can be used for various assessment purposes including formative and summative assessment. Each of these potential assessment activities and tasks has been assigned to one of five categories which are: (1) 4 Selected Responses e.g., multiple-choice items, matching, (2) 13 Brief constructed responses, e.g., short answers and label a diagram, concept map, (3) 10 Products, e.g., art exhibit or essay, (4) 10 Performances, e.g., oral presentation, debate, and (5) 9 Process-focused assessments, e.g., oral questioning, student-self-assessment, conference (excerpted from Figure 3.1 McTighe & Ferrara, 2020). It's important to note that this research was not intended to serve as a validation study for any particular conceptual framework. However, the data does indicate that there is substantial support for McTighe and Ferrara's framework beyond mere face validity.

Figure 1 outlines the logical relationships that have been explored in this study regarding the identification of an effective formative assessment system in the classroom and a potentially valuable approach to professional development in this field. The approach is primarily quantitative, and the research questions are:

1. Does McTighe and Ferrara's Conceptual Framework for assessment provide a psychometrically tenable tool for classifying the FA activities in the classroom?
2. Do McTighe and Ferrara's categories make sense as a tool for providing professional development to groups of teachers?
3. In what ways is McTighe and Ferrara's framework helpful for thinking through ways to provide useful professional training regarding teachers in schools?



Notes on abbreviations: McT&FCFA: McTighe and Ferrara's Conceptual Framework for Assessment
EASC: Effective Assessment System in the Classroom; PDT: Professional Development Training

Figure 1: Research Conceptual Framework

By addressing the research gap concerning the role of CPD in supporting teachers' successful application of formative assessment, educators and policymakers can make informed decisions to develop comprehensive professional development programs that facilitate the effective implementation of FA strategies.

2. Review of Literature and Research Papers

It is important to note that numerous researchers have proposed frameworks for categorizing formative assessment techniques and strategies including computer adaptive assessment (McTighe & Ferrara, 2021; Stiggins, Arter, Chappuis, & Chappuis, 2006; McMillan, 2017; Berger, Verschoor, Eggen, & Moser, 2019). If it is examined in more detail, McTighe and Ferrara have assembled a comprehensive yet concise approach to professional development around classroom assessment which spans diagnostic, formative, and evaluative (summative) assessment. Their approach includes a sequential set of frameworks for planning assessment, selecting assessment technique(s), evaluating student work, and communicating results. In terms of assessment techniques, McTighe and Ferrara

have identified a non-exhaustive list of different techniques, 36 in their 1998 book and 46 in their most recent 2021 book, that could be used for one or more types of assessment, i.e., diagnostic, formative, summative, or evaluative, based on the type of information needed for a specific purpose or audience. These techniques are grouped into five different areas: selected responses, constructed responses, products, performances, and process focused. McMillan's (2017) category system is like McTighe and Ferrara's in that it contains 49 techniques grouped within eight different categories (McMillan, "Classroom Assessment: Principles and Practice for Effective Standards-Based Instruction", 2017). Perhaps the most significant difference between the two category systems is that McMillan includes distinct categories for teacher observation and student self-assessment while McTighe and Ferrara weave many of these techniques into a single process-focused category. Since the purpose of this initial study is to explore the potential usefulness of a set of categories to guide formative assessment professional assessment, rather than identifying a potentially optimal system, we have opted to examine McTighe and Ferrara's system since it contains fewer categories, five rather than McMillan's eight. Another reason for using McTighe and Ferrara's system for this initial study is that their 2021 book includes sections where the strengths and weaknesses of each broad category are discussed. Thus, it is important to note that while the individual techniques within each category bear important structural and/or operational similarities, they are not seen as interchangeable especially since the selection of a technique depends on the purpose. While both McTighe and Ferrara's and McMillan's conceptual arrangements of assessment techniques appear to have a high degree of face validity to the authors of this paper, the core question to be examined is whether McTighe and Ferrara's framework offers a useful and appropriate way to categorize approaches geared specifically to classroom-based formative assessment.

It should be noted that our approach is somewhat in contrast to the typical, or classical, survey development process where the goal is to create the most parsimonious set of items that accurately reflect some level of performance or opinion of the underlying or latent construct(s) (Morgado, Meireles, Neves, Amaral, & Ferreira, 2018). Each of the items on our survey represents a potentially useful form of classroom assessment. Thus, from a professional development perspective, each item on the survey provides useful information for planning training. If everyone in the group knows a particular form of assessment, it can either be ignored or from our viewpoint, potentially serve as the basis for strength-based professional development.

2.1 Professional Development Targeted for Formative Assessment

In terms of approaches to effective professional development, McTighe and Ferrara (2021) provide a series of 10 guidelines for school leaders (see Table 1) that are very much in keeping with the most recent research on effective PD practices (Darling-Hammond, Hyler, & Gardner, 2017; Hill & Papay, 2022). In their extensive review of research on "professional learning" Hill and Papay have identified effective formats and important areas of focus:

For the how of instructional delivery, research suggests the following PL formats can be particularly effective at producing changes in instructional effectiveness: (1) built-in time for teacher-to-teacher collaboration around instructional improvement; (2) one-to-one coaching, where coaches work to observe and offer feedback on teachers' practice; and (3) follow-up meetings to address teachers' questions and fine-tune implementation. For the what, there is growing evidence that PL may be more productive when it focuses on (1) building subject-specific instructional practices rather than building content knowledge alone; (2) supporting teachers' instruction with concrete instructional materials like curricula or formative assessment items rather than focusing only on general principles, and (3) explicitly attending to teachers' relationships with students. (p. 2)

Table 1: McTighe and Ferrara's "10 Tips for School Leaders"

1. Assess the Staff
2. Conduct a School Assessment Audit
3. Offer Targeted Professional Development
a. Principles of quality assessments
b. Effective use of diagnostic (pre-)assessments
c. Effective use of formative assessments
d. Strengths and limitations of various methods for evaluative assessment
e. Designing and using performance-based assessments
f. Designing and using criterion lists and rubrics
g. Giving effective feedback to students
h. Sound grading practices
4. Share and Discuss Assessment Related Articles and Books
5. Develop or Adopt Principles of Assessment and Grading
6. Share Successful Practices
7. Schedule Visitations
8. Design Assessments Collaboratively
9. Examine Assessment Results in Teams
10. Discourage Excessive Test Prep

Note: Adapted from: McTighe and Ferrara, (2021) (pp. 72 - 79).

While extensive professional development material has been developed to provide teachers with classroom assessment skills (McTighe & Ferrara, 2021; Stiggins, Arter, Chappuis, & Chappuis, 2006), research on actual classroom practices related to formative assessments has suggested that teachers have difficulty implementing these techniques (Alonzo, 2018); (Gulikers, Biemans, Wesselink, & Van der Wel, 2013). Some of the barriers that teachers face in using FA are related to the habits, expectations, and dispositions of students, and in recent decades, the pressure that teachers must carry out the entire curriculum to better prepare students for the end of the year, high-stakes exams (Box, Skoog, & Dabbs, 2015). Gulikers et al., have also provided evidence that teachers tend to talk about their FA as a product but with no reflection on its purpose (Gulikers, Biemans, Wesselink, & Van der Wel, 2013). An evaluation study by Weinbaum (2009) of a 10-state initiative to foster the use of formative assessment based on an extensive set of materials developed by the Assessment Training Institute, indicated that there was only limited positive impact on teachers' effective use of formative assessment. A more recent large-scale study by Randel et al. also reported a limited positive impact of formative assessment professional development (Randel, Apthorp, Beesley, Clark, & Wang, 2016). The situation is not hopeless in that there have been some reports of the positive impact of professional development in this area (Furtak, et al., 2016; Lyon, Nabors, & Wylie, 2019). Of relevance to the study reported here, Wylie and Lyon note that many of the teachers in their study commented on the need for more resources about specific FA techniques and practices. One of the teachers in their 2015 study commented:

> *"Teachers need a list or resource guide showing the different ways to assess students. Professional development on formative assessment and how to use the assessment to guide instruction. How to use the formative assessment in their planning. (5th grade, Generalist teacher, North Carolina)" (Wylie & Lyon, 2015).*

There also have been numerous studies that have documented the usefulness of FA professional development on teachers' work with Special Needs students across a wide range of subject areas (Wylie & Lyon, Developing Formative Assessment Protocol to Support Professional Growth, 2020; Andersson & Granberg, 2022)

However, it should be noted that in the case of Furtak et al. (2016), their work involved four

years of PD with science teachers. Is there a more efficient way to provide effective FA-PD for teachers? Perhaps rather than approaching FA-PD from a deficit perspective, a way is needed to quickly identify areas of a group of teachers' strengths and then build short-term FA-PD around it. In his conclusion to the Council of Chief State School Officers (CCSSO) report, Weinbaum (2009) says that:

> *"While many of the ideas were new to participating teachers, according to their reports, there were opportunities to make connections to their pre-existing knowledge (although such connections were not often made explicit). As teachers began to discover these connections for themselves, their confidence and interest in the material grew." (Weinbaum, 2009) (p.43)*

From Weinbaum's perspective to significantly impact teachers' behavior, professional development programs should build on their cognitions and practices (Van den Bergh, Ros, & Beijaard, 2014). The study reported here is a first step to determine if McTighe and Ferrara's framework can be used to identify teachers' FA strengths in different categories of formative assessment techniques.

3. Method

3.1 Population and Sample

The population for this study was secondary school teachers in Northeastern Massachusetts. The target sample consisted of two groups of secondary school teachers. The first group was teachers at three schools, a public high school, a public middle school, and a grade 6-12 charter school (target N =317). The second group was secondary school teachers who were enrolled in graduate education summer school courses at the authors' university (target N = 36). Thus, the total target population was 353 secondary school teachers. In total 121 teachers agreed to participate, or 38% of the target sample. This working sample included 90 teachers at the three public schools (28.3%) and 32 of the teachers attending summer school (88.8%).

3.2 Instruments

The survey that was developed for this study was based on two sources: Zhang and Burry-Stock (2003) and McMillan (2017). Zhang and Burry-Stock's "Assessment Practice Inventory" had been designed to capture elementary through high school teachers' perceived skill levels and usage across a broad range of classroom assessment practices. These practices included topics such as communicating assessment results, ensuring the reliability and validity of assessment tools, using standardized tests, and grading practices. With permission from the first author, Zhang, we used the inventory's item stems, the Likert scales, and 6 of the specific in-class assessment items. The item stem for skills was: "In my classroom for formative assessment practices when I…." and the Likert scale is described below.

To focus our survey on a broad range of FA tools, as opposed to assessment practices in general, we used items that had been identified by McMillan (2017) in his classroom assessment textbook that was designed for pre-service teachers. His items ranged from selected-response and constructed-response assessments through teacher observation and student self-assessment.

Thus, our survey incorporated 21 of McMillan's items along with the six in-class items from Zhang's inventory. Our 27-item survey was piloted with 26 School of Education faculty in May 2018. The pilot study indicated that the instrument had a strong level of internal consistency with a Cronbach alpha of 0.81. Thus, no changes were made to the survey items. A copy of the survey is in Appendix A.

Of the 27 items on the original survey, 25 served as the basis of analysis for this study. The two items that were eliminated were specifically related to writing exam items. Table 2 contains the list of included items. Participants rated each item on a 5-point rating scale, ranging from "1 - I am not at all skilled" to "5 –

I am very skilled". The surveys were anonymous but there were four basic demographic questions included in the survey, i.e., gender, years of teaching experience, teaching subject area, and recency of FA training. (We also asked teachers to rate their frequency of use for each item, but those data are not relevant to this study.) The survey took about 15 minutes to complete.

To address the first research question, i.e., does McTighe and Ferrara's Conceptual Framework for assessment provide a useful tool for classifying the FA activities in the classroom, 25 of the 27 items in our original survey were divided into five subscales based on McTighe and Ferrara's framework (see Table 2). There were seven items on the original survey that did not explicitly map onto their framework. One, "write test items for higher cognitive levels", was discarded before analysis because several participants asked for clarification about the item. Independently, the researchers classified the other six items and then met to compare their categorizations. We agreed on five of the six items and assigned each to the agreed-upon subscale (see Table 2). One item "paper and pencil tests" could not be confidently assigned to a category and thus was discarded for the remaining analyses. The wording of the remaining 19 items was either worded the same or similarly to the wording in McTighe and Ferrara's 1996 classification table.

Table 2: Descriptive statistics - Self-reported skill levels by McTighe and Ferrara's conceptual framework

Self-Reported skill levels by McTighe & Ferrara's formative assessment categories	Teacher's skill rating			
	Mean	SD	Skewness	n
Selected Responses (3 items)	3.94	0.82	-.468	91
(a) Multiple-choice questions	4.08	0.79		91
(b) Matching questions	3.92	0.98		91
(c) True-false questions	3.82	0.98		91
Constructed Responses (7 items)	3.60	0.72	-.278	91
(a) Fill-in-the-blank questions	4.09	0.89		91
(b) Short answer questions	4.13	0.83		91
(c) Essay items with restricted response*	3.51	1.21		91
(d) Essay items with extended response	3.68	1.06		91
(e) A task such as a graph	2.96	1.30		91
(f) A task such as illustration	3.26	1.29		91
(g) Oral questioning such as examinations*	3.57	1.18		91
Products (7 items)	3.59	0.91	-.608	91
(a) A task such as a project	3.89	0.90		91
(b) A task such as a portfolio	3.12	1.21		91
(c) A task such as a video	2.68	1.10		91
(d) A task such as a spreadsheet	2.54	1.19		91
(e) A task such as an exhibition	3.09	1.17		91
(f) A task such as a journal	3.54	1.14		91
(g) Tasks such as web page*	2.43	1.11		91
Performances (4 items)	3.04	0.77	0.15	91
(a) A task such as speech or talking in class	3.73	1.11		91
(b) A task such as demonstration	3.89	0.99		91
(c) A task such as debate	3.09	1.25		91
(d) A task such as readings*	3.65	1.25		91
Process-Focused (4 items)	3.68	0.91	-.663	91
(a) A task such as reflection	3.79	1.02		91
(b) Informal questioning*	4.03	0.97		91
(c) Oral questioning such as conferences	3.55	1.26		91
(d) Oral questioning such as interviews	3.33	1.28		91

Note: Items designated with an asterisk were assigned to a dimension through consensus discussion among the researchers. The remaining 19 items were identical or worded similarly to items in McTighe and Ferrara's 1996 framework.

3.3 *Data Collection*

The data collection occurred in two distinct phases in the Spring and Summer of 2018. Surveys were collected from the three public secondary schools near the end of the school year. Each school received a packet of material that included multiple individual envelopes with each one containing a survey and a disclosure form with instructions. Researchers for this study deliberately chose to collect the data using paper and pencil format because they wanted participating teachers to feel confident that the data, they were providing was anonymous. Teachers picked up and returned the surveys to the main office. They were instructed to seal their completed survey in the envelope before they returned it. The timing of survey distribution and collection was left to the discretion of each school's principal. Due to the relatively low survey return rate, about 28%, it was decided to collect survey data from secondary school teachers who were attending two graduate education courses during the summer. With the approval of the professors who taught these two courses, one of the researchers visited the classes to explain the study and administer the survey to willing participants. Surveys were distributed to everyone, and the graduate students could return either a completed anonymous survey or a blank survey in a sealed envelope.

While a total of 121 teachers participated, during data analysis it was noted that there were many instances of missing data although there were few missing items per teacher. To determine if there were systematic differences in assessment skills related to those respondents with missing data versus respondents without missing data, individual one-way ANOVAs were run for each of the 25 formative assessment skill items. Three of the items showed significant t-value differences between the two groups ($p < .05$) with the missing data respondents scoring lower than respondents without missing data. None of the differences were statistically significant based on post-hoc Bonferroni correction ($p > .002$) (Dunn, 1961). Given the overall lack of a pattern between the missing and non-missing data groups in terms of individual items nor a difference between Humanities and STEM teachers in terms of percentages with missing data (17.1% and 15.2% respectively), it was decided to ensure consistency across all the reported analyses by using listwise deletion based on Teaching Field and the McTighe and Ferrara framework. Thus, the working sample for this report is n = 91. Background characteristics regarding the subject area taught, years of teaching experience, and recency of professional development related to assessment are presented in Table 3.

Table 3: Frequency of teaching experience and recency of training by field of teaching

		Recency of professional development training		
Field of teaching	Years of teaching	Yes, recent	No, not recent	Total
STEM	< 5 yrs.	7 (70.0%)	3 (30.0%)	10
	6 to 10 yrs.	3 (60.0%)	2 (40.0%)	5
	> 10 yrs.	9 (69.2%)	4 (30.8%)	13
	STEM Total	19 (67.9%)	9 (32.1%)	28
Humanities	< 5 yrs.	16 (64.0%)	9 (36.0%)	25
	6 to 10 yrs.	8 (61.5%)	5 (38.5%)	13
	> 10 yrs.	19 (76.0%)	6 (24.0%)	25
	Humanities Total	43 (68.3%)	20 (31.7%)	63
Combined STEM & Humanities	< 5 yrs.	23 (65.7%)	12 (34.3%)	35
	6 to 10 yrs.	11 (61.1%)	7 (26.3%)	18
	> 10 yrs.	28 (73.7%)	10 (26.3%)	38
	Overall Total	62 (68.1%)	29 (31.9%)	91

Note: Based on Listwise deletion of missing data, n = 91

3.4 *Data Analyses*

Once the researchers classified the survey items into the 5 categories of McTighe and Ferrara's framework, there were three main phases to the data analysis. IBM SPSS Statistics Version 26, and

Amos Graphics Version 26 were used for all analyses (IBM Corp., 2021). The first phase involved an examination of the background characteristics of the obtained set of teachers via frequency counts (see Table 3). Most of the participants taught humanity subjects (68.3%), had training within the past five years (68.1%), and were either new teachers, i.e., fewer than 5 years, 38.5%, or experienced teachers with more than 10 years of experience, 41.8%.

The second phase related to the research question regarding the applicability of McTighe and Ferrara's Conceptual Framework to the FA skill items included in this survey. This examination of the data involved a sequence of four sets of analyses. First, there was an exploratory factor analysis (EFA) to determine if a latent structure could be identified from the survey data. Five factors were identified. This was followed by confirmatory factor analysis (CFA) using Amos to determine if this latent structure was parsimonious. Then Amos was used to conduct a CFA based on factors derived from McTighe and Ferrara's (2021) classification scheme. Then the two sets of CFA results, i.e., from the five factors identified through EFA and McTighe & Ferrara's five dimensions/factors were compared. Finally, Pearson Correlation coefficients were calculated to assess the internal consistency of each of the McTighe and Ferrara sub-scales.

The third and final phase of the data analysis involved a one-way MANOVA to address the second and third research questions regarding the potential usefulness of McTighe and Ferrara's framework for providing professional development to in-service teachers based on their field of teaching, STEM, or Humanities. This division among respondents was based on professional development that has shown the value of affinity group and discipline-focused approaches (Noonan, 2019; Viskupic, et al., 2019).

In preparation for the MANOVA, the data were examined to determine the degree to which they met the assumptions for this statistical procedure. A description of these tests is contained in Appendix A along with tables showing the results of the tests for each assumption. See Tables A1 through A5 in Appendix A. There were scattered violations of the assumptions for the MAOVA analysis. However, follow-up analyses and a review of research related to violating assumptions for MANOVA, indicated that while the violations may have limited the power of the MANOVA (Tabachnick & Fidell, 2007), the researchers felt that the MANOVA could still possibly reveal useful information about this sample of teachers' skills based on McTighe and Ferrara's conceptual framework for FAs. See Table A6 find the appendix or a summary of the tests of MANOVA assumptions and potential impact on subsequent analyses.

3.5 Ethical Considerations

The design for this study was submitted and approved by the university's Institutional Review Board (IRB) before data collection. Participants were informed of the goals of the study and the nature of their voluntary participation via a disclosure form and verbally for those participants who were enrolled in the summer school classes. The surveys were anonymous and did not pose any risk to the participants.

4. Results

4.1 Research Question #1: "Does McTighe and Ferrara's Conceptual Framework for assessment Provide a Psychometrically Tenable Tool for Classifying the FA Activities in the Classroom?"

The research question about the potential psychometric feasibility of using McTighe and Ferrara's Conceptual Framework for assessment to classify teachers' self-reported FA skills was addressed in two ways. First, we examined the survey data via EFA to determine if there was a way to classify the FA skills other than McTighe and Ferrara's framework. The EFA was conducted with the 24 survey items based on a Principal Components analysis via Oblimin Rotation with Kaiser Normalization. Five factors were identified with eigen values greater than 1. 0.. The factors with their item loadings

are presented in Tables 4 and 5. Given our relatively small sample size, we were cautious about ascribing items with somewhat low factor ratings. Thus, we applied guidelines for samples less than 100 so that we only selected items with loadings greater than .6 (MacCallum, Widaman, Zhang, & Hong, 1999). The 19 items with factor loadings greater than .6 were transferred to Amos Graphics to run a CFA. The results from the CFA indicated that the five factors met the minimum criteria for an acceptable fit for the data, CMIN/DF = 2.046 (Cucos, 2023) However, other indicators of this five-factor model suggest that it does not provide a very useful model for simplifying the data. The results for the RMSEA test, 0.108, do not meet general standards for good fitting parsimonious models where the RMSEA should be ≤ 0.06 (Cucos, 2023). Given the indeterminacy of the EFA 5-factor model, we next conducted a CFA by assigning the 24 items to McTighe and Ferrara's 5 dimensions/factors. Here too, the results indicated that their 5 factors were an acceptable fit for the data, CMIN/DF = 2.785 but not a very parsimonious model, RMEA = 0.141. Since both models were minimally acceptable but not parsimonious, we next compared the two models visually. From Figure 2 it can be seen that a large portion of the items are similarly classified in both models, 12 out of 19. Furthermore, 5 survey items did not load on any of the EFA factors. Thus, while there was no expectation that McTighe and Ferrara's assessment dimensions would parsimoniously capture teachers' current level of formative assessment skills, this comparative CFA analysis does suggest that McTighe and Ferrara's classification system can be a useful tool for professional development in formative assessment. Indeed, based on research by Sato and colleagues (Sato, et al., 2006; Sato, Wei, & Darling-Hammond, 2008) we would assume that before formal professional development, teachers' skill levels would cluster around those summative assessment skills that they acquired in their teacher preparation program or were emphasized in their school district. We then asked if there was other evidence for the coherence of McTighe and Ferrara's dimensions by examining the degree of internal consistency within each of the factors/dimensions. As can be seen in Table 6, the Cronbach alphas were all strong to moderately strong (Tavakol & Dennick, 2011), ranging from .865 for the Selected Response subscale to .764 for the Constructed Response subscale. Table 7 shows the change in alpha levels with individual items removed from each subscale. Not surprisingly, removing items from the 3 and 4-item scales tends to have more of an impact than removing items from the scales with 7 items. The strong and moderately strong measures of internal consistency within each subscale suggest that if a teacher's self-reported skill level of an FA technique is high, then some aspect of that skill can be applied to mastering other FA techniques that are within the same subscale. Thus, while there may be many statistically viable ways to organize these self-reported FA skill items, the use of McTighe and Ferrara's Conceptual Framework as a basis for FA professional development is supported by this analysis.

Table 4: Exploratory factor analysis - Eigen values and variance accounted for via oblimin rotation with Kaiser normalization.

Total Variance Explained			
Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	5.049	20.194	20.194
2	3.843	15.371	35.565
3	3.682	14.728	50.293
4	2.723	10.893	61.186
5	1.827	7.307	68.493

Table 5: Exploratory factor analysis pattern matrix item loadings based on oblimin rotation with Kaiser normalization.

Pattern Matrix					
	Component				
	1	2	3	4	5
Skills to use tasks such as oral questions at conferences	.877				
Skills to use tasks such as interviews	.786				
Skills to use tasks such as readings	.771				
Skills to use tasks such as debate	.753				
Skills to use tasks such as oral examinations	.730				
Skills to use tasks such as journal	.570		.497		
Skills to write essay items with extended response	.568	.414			
Skills to write essay items with restricted response	.518		.414		
Skills to write matching questions		.905			
Skills to write true-false questions		.880			
Skills to write fill-in-the-blank questions		.808			
Skills to write multiple-choice questions		.798			
Skills to write short answer questions		.715			
Skills to use tasks such as portfolio			.840		.368
Skills to use tasks such as web page			.710		
Skills to use tasks such as video			.681		
Skills to use tasks such as exhibition			.676		-.303
Skills to use tasks such as project			.653		
Skills to use tasks such as reflection			.486		-.364
Skills to use tasks such as graph				.873	
Skills to use tasks such as spreadsheet				.784	
Skills to use tasks such as demonstration					-.727
Skills to use tasks such as speech or talking in class	.443				-.643
Skills to use tasks such as informal questioning	.467				-.587
Skills to use tasks such as illustration			.306	.316	-.384

Notes: Extraction Method: Principal Component Analysis. Rotation converged in 14 iterations.

Figure 2: Comparison of McTighe & Ferrara's framework with EFA factor loadings

Table 6: Cronbach's Alpha Internal Consistency Estimates for Each of the McTighe and Ferrara's Dimensions

Formative assessment contexts	Cronbach's Alpha	# of Items
1. Selected Responses (3 items)	.865	3
2. Constructed Responses (7 items)	.764	7
3. Products (7 items)	.812	7
4. Performances (4 items)	.797	4
5. Process-Focused (4 items)	.807	4

Note: n = 91 (listwise deletion)

Table 7: Impact on Cronbach Alpha if Individual Items are Removed from McTighe and Ferrara's Subscales

Formative assessment items by McTighe and Ferrara's categories	Impact on the scale if an item deleted		
	Mean	SD	Cronbach's Alpha
Selected Responses (3 items)			
Multiple-choice questions	7.75	3.46	.884
Matching questions	7.90	2.47	.707
True-false questions	8.00	2.67	.809
Constructed Responses (7 items)			
Fill-in-the-blank questions	21.11	20.67	.736
Short answer questions	21.07	20.24	.719
Essay items with restricted response	21.69	18.15	.714
Essay items with extended response	21.52	19.92	.738
A task such as a graph	22.24	21.27	.799
A task such as an illustration	21.93	17.91	.722
Oral questioning such as examinations	21.63	18.08	.708
Performances (4 items)			
A task such as speech or talking in class	10.63	7.90	.718
A task such as a demonstration	10.46	9.63	.821
A task such as a debate	11.26	7.02	.693
A task such as readings	10.70	7.39	.732
Products (N = 7 items)			
A task such as a project	17.40	23.22	.788
A task such as a portfolio	18.16	21.03	.782
A task such as a video	18.60	21.53	.778
A task such as a spreadsheet	18.75	23.24	.819
A task such as an exhibition	18.86	21.06	.769
A task such as a journal	18.20	21.01	.776
Tasks such as web page	17.75	22.30	.797
Process-Focused (4 items)			
A task such as a reflection	10.91	9.24	.822
Informal questioning	10.67	8.87	.799
Oral questioning such as conferences	11.15	6.64	.682
Oral questioning such as interviews	11.37	6.84	.719

4.2 Research Question #2: "Do McTighe and Ferrara's Categories Make Sense as a Tool for Providing Training to Groups of Teachers?"

Tables A2 and A3 in the appendix show the mean self-reported FA skill levels by the teacher's field of teaching across each of McTighe and Ferrara's subscales. Given the acceptable levels of internal consistency among the subscales, the second research question was investigated via MANOVA where

the independent variable was the teacher's field of training, i.e., STEM or Humanities and the dependent repeated measures variables were McTighe and Ferrara's 5 subscales. While STEM teachers reported higher skill levels for Selected Responses and Constructed Responses FA than Humanities teachers, the pattern was reversed for Products, Performances, and process-focused FAs. Given these average score differences between groups, MANOVA was used to determine if there was a statistically significant difference between these two groups of teachers on the five subscales. Indeed, there was a statistically significant difference between the two groups of teachers on the combined set of subscales, $F(5, 85) = 5.681$, $p < .01$; Wilks' $\Lambda = .750$, partial $\eta^2 = .250$. (See Table 8, online supplemental materials.) This finding suggests that their conceptual framework could provide a way to scaffold FA training based on teachers' instructional responsibilities.

Given the statistically significant overall multivariate difference between the two groups of teachers, Descriptive Discriminant Analysis (Warne, 2014), was used to determine which if any of the subscales were contributing to differences between the teachers' field of teaching. As can be seen from the discriminant analysis in Table 4 nearly three-quarters of the teachers were correctly identified. The discriminant analysis was slightly more accurate in its classification of Humanities teachers. While discriminant analysis provides both standardized canonical discriminant function coefficients and structure matrix coefficients, structure matrix coefficients are not affected by collinearity (IBM, 2021). Thus, given that there were some relatively high correlations between subscales particularly for STEM teachers (See Tables A2 and A3), it was decided to rely on the Structure Matrix coefficients to interpret the cluster of subscales that were discriminating between STEM and Humanities teachers. The three subscales that contributed meaningfully to discriminating between teachers' fields of teaching were skill ratings in Performances, Products, and Process-Focused subscales. As can be seen in Table 9, each of their coefficients was greater than .3 (Stella, 2019). Based on this descriptive discriminant analysis, it appears that for these three areas of FA, it would be appropriate to utilize the teachers' fields of teaching as a starting point to customize their professional development.

Table 8: Multivariate Test of Differences Between STEM and Humanities Teachers on Self-Reported Formative Skill Levels

Effect	Value	F	Hypothesis df	Error df	Sig	Partial Eta Squared
Intercept	.030	540.471	5	85	< .01	.970
Field2	.750	5.681	5	85	< .01	.250

Table 9: Descriptive Discriminant Analysis – Influence of McTighe and Ferrara's Subscales on Predicting Field of Teaching

Subscale	Structure Matrix Coefficient	Standardized Canonical Discriminant Function Coefficient
Selected Responses (3 items)	-.100	.383
Constructed Responses (7 items)	-.111	-1.716
Performances (4 items)	.469	.946
Products (7 items)	.354	.715
Process Focused (5 items)	.333	.456

4.3 Research Question #3: "In What Ways are McTighe and Ferrara's Framework Helpful for Thinking Through Approaches to Useful Professional Training for Teachers?"

We approach this question from the perspective of strength-based professional development (He, 2009; Zwart, Korthagen, & Attema-Noordewier, 2014) Is there a way to identify a teacher's strongest formative assessment skill(s) that would serve as the starting point for targeted and efficient

professional development? Looking at the mean skill ratings in Table 2, the specific area where teachers report the most skill is in the use of short answer questions. While a strength-based professional development initiative could start with short answer questions, it would be more efficient, to begin with a formative assessment skill that is operationally related to the type of new techniques that teachers have identified as important to master. Thus, for example, if a teacher wants to acquire skills in using debate for formative assessment, McTighe and Ferrara's framework would suggest that the appropriate category to examine the teacher's existing skill would be "Performance". For this sample of teachers, the strongest self-reported formative assessment skill in the Performance category is "Demonstrations" (M = 3.89). Of course, the operational dynamics of an FA-PD would differ across school districts, but by using McTighe and Ferrara's framework, the starting point would be to identify the target FA skills to be acquired and then to identify teachers' area of strength within the relevant area of the framework.

5. Discussion

Teachers' self-assessment of their formative assessment skills varies both among groups of teachers and across specific assessment techniques used by each teacher. The data collected for this study suggests that McTighe and Ferrara's framework can serve as a useful scaffolding for structuring FA-PD. since specific formative assessment techniques cluster very well within their five broad assessment categories (see Research Question #1). These categories can be used to identify sets of techniques that groups of teachers, in this case, Humanities and STEM teachers, differ in terms of their self-reported skill levels (see Research Question #2). It is important to note that in other school settings, the groups of teachers' reported skill levels may not cluster around the teaching area but perhaps around grade level, class composition, teaching experience, etc. The key point is that a simple survey like the one used in this study can be used as a starting point for identifying specific formative assessment techniques within categories of McTighe and Ferrara's framework where teachers feel that are highly or at least adequately skilled. Then learning communities or other kinds of groups can be formed using this information (see Research Question #3). Instead of gearing FA-PD around a deficit model, teachers could be actively engaged in exploring ways to transfer their existing FA skills to assessment techniques in the targeted area of McTighe and Ferrara's framework. While there are countless ways this could unfold in a school setting, two scenarios are described below. Scenario A presents a situation in which the target FA skill is viewed as complementary to the teacher's highly skilled FA technique and the second is where the target skill is viewed as sharing various components with the teacher's highly skilled FA technique. Allowing teachers to self-assess their formative assessment skills regarding specific types of classroom activities within a clear conceptual framework sets the stage for a motivationally enabled professional development initiative which Weinbaum (2009) noted was of central importance in his conclusion to the evaluation of the 10-state collaborative assessment initiative.

5.1 Scenario #1: Professional Development Based on Complementary FA Techniques

For example, if it were decided by teachers to address formative assessment observation skills, which is a process-focused technique in McTighe and Ferrara's framework, teachers from our sample, would begin by discussing their successes using informal questioning (M = 4.023) as a formative assessment technique with students since this was the technique with the highest average rating (see Table 2). Then the conversation(s) could shift to the usefulness of informal questioning among their students with Special Needs and English Language Learners. At this point, if needed or decided, teachers would go back to their classrooms to pay close attention to the benefits and limitations of their use of informal questioning with their students with Special Needs. From there, these discussions could segue into similarities and differences between the observing and the informal questioning techniques with particular attention to the characteristics of their students with Special Needs.

Teachers would ponder the following question: how can observation help inform the practice of informal questioning? Thus, teachers would not be asked to drop one form of formative assessment while increasing their skills with another but rather to explore and practice ways that another type of process-focused technique, in this case, observation, can serve to enhance the information gleaned from informal questioning. The key here in terms of engaging teachers in enhancing their observation skills is that it would be framed in the context of improving the usefulness of FA information. As a concrete example, perhaps teachers point out that students with ADHD are comfortable with informal questioning about their reading, but it is often difficult to pose questions that get at a specific student's understanding of something read in class. By observing a student with ADHD during reading, the teacher can identify those portions of the text that the student may have been attending to and then generate informal questions that relate to those passages of text. If the teacher finds that the student understood that passage, perhaps in the future the teacher could help the student by literally highlighting smaller sections of text where they should put their attention during in-class reading. On the other hand, if the student's responses to observation-based informal questioning suggest that the student did not fully comprehend the text then a different intervention strategy may be needed.

5.2 *Scenario #2: FA-PD Based on Parallels Between FA Techniques*

For example, if the humanities faculty in our sample decided that they wanted their students to be highly competent and confident in their ability to present and defend their opinions and beliefs, they might begin to implement debating as a recurring activity in their teaching, but they would also realize that their FA skills are somewhat limited in this area (M = 3.09). Since debates are one type of "Performance' in McTighe and Ferrara's framework, based on our sample, these teachers would begin by examining their FA practices related to demonstrations (M = 3.89) to identify components that they could apply to debates. This could include elements such as asking probing questions about information sources being utilized and giving feedback during rehearsals. Of course, the probing questions and aspects of "performance" to focus on during rehearsals would differ between demonstrations and debates but the core skills that they have mastered in crafting useful probing questions and giving informative yet supportive feedback for demonstrations could be transferred to working with students as they prepare debates.

6. Limitations of the Study

As noted earlier, the relatively small sample size has constrained some of the analyses. This can be addressed in future studies across multiple types of school settings and grade levels. A second limitation is that these analyses are based on self-reported skill levels which may not reflect accurately the teacher's skills using the formative assessment tool. Thus, while the survey is a reasonable starting point for instrument development, both the planning and the impact of PD should also be assessed via classroom observation. Another possible limitation relates to the way that some participants may have responded to survey items. While the instructions and wording of survey items were focused on FA, some respondents may have thought about their skill levels more generally to encompass both formative and summative assessment. Even if this were the case, it probably would not have affected the analyses related to the three research questions, but it may have affected the mean scores on the various categories of McTighe and Ferrara's framework, especially the selected response and constructed response scales. Thus, those who are responsible for professional development should be cautious about the wording of such a survey and cautious about interpreting the means when planning formative assessment professional development.

7. Future Research

Future research should proceed along two different paths. One path relates to establishing the structural validity of McTighe and Ferrara's framework in terms of formative assessment. One way to do this would be to administer this survey to two groups of teachers, those who have had professional development that utilized their framework and another group of similarly experienced teachers who have not been formally exposed to McTighe and Ferrara's framework. Support for their framework would be assessed via confirmatory factor analysis that compared indicators such as AIC for the two groups. The second research path would focus on strength-based professional development. Are there any meaningful differences in terms of the effectiveness of strength-based PD compared to the more typical deficit-based approaches? The strengths-based approach could use either or both scenarios described above, i.e., having teachers enhance their assessment skills by focusing on complementary FA tasks and activities and or focusing on parallel aspects of FA tasks and activities.

8. Implications and Conclusions

Our review of prior research brought out two important yet out-of-synchrony points regarding formative evaluation in classrooms. On the one hand, there is considerable data that points to the value of FA in terms of enhancing students' learning, e.g., Fukuda, et al. (2020), and helping teachers improve their instruction, e.g., Sondergeld, et al. (2010). On the other hand, while teachers are aware of the importance of FA, teachers report many barriers to effectively using these FA tools, e.g., Alonzo (2008), and Gulikers, et al. (2013). Given the wide range and multitude of FA techniques, a conceptual framework that provides a meaningful structure to categorize these techniques can go a long way towards helping groups of teachers select useful techniques and subsequently acquire the skills to utilize these techniques effectively in their teaching. In this small-scale study of secondary school teachers, we examined McTighe and Ferrara's conceptual framework to see if it could serve as a guide to professional development. The findings across the three research questions indicate that their framework offers a logical and psychometrically defensible way to categorize broad groups of FA techniques as well as providing a basis to identify clusters of teachers across broad subject areas for targeted strength-based professional development.

At least in terms of teachers' self-reported skill levels, the framework appears to possess adequate levels of construct validity (Matthay & Glymour, 2020) This can and should be examined in future large-scale studies that include elementary school teachers.

Thus, this study's survey instrument could serve as the basis for more site-specific data collection instruments and perhaps serve as a straightforward pre-post assessment of the effectiveness of training in specific FA techniques.

9. Acknowledgment

The first author wishes to thank the Fulbright Scholar Program and the International Institute of Education, for their awarding of grants, which have enabled a study in the realm of formative assessment. In the framework of this paper, only the primary data that have been collected in some schools in the Northeast of Massachusetts USA, have been used. The authors wish to thank all the teachers who have participated, and a special thanks go to the Professor. C. Julie Whitlow, Vice Provost for Academic Affairs, at Salem State University, provided essential aid in creating opportunities for the realization of data collection.

References

Airasian, P. W. (2001). Classroom Assessment: Concepts and Applications. (4th ed). New York, NY: McGraw-Hill.

Alonzo, A. C. (2018). An argument for formative assessment with science learning progression. Applied Measurement in Education, 31(2), 104-112. doi:10.1080/08957347.2017.1408630

Anderson, C. (2020). Formative assessment – from the view of special education teachers in mathematics. Nordic Studies in Mathematics Education, 25(3-4), 73-83. http://ncm.gu.se/wp-content/uploads/1605/83/25_34_073094_andersson.pdf.

Andersson, C., & Granberg, C. (2022). From Passenger to Pilot-Using Formative Assessment to Support Students With Intellectual Disabilities to Become Self-Regulated Learners. Frontier Education: Special Educational Needs, 7. doi:https://doi.org/10.3389/feduc.2022.845208

Andersson, C., & Palm, T. (2018). Reasons for teachers' successful development of a formative assessment practice through professional development - a motivation perspective. Assessment in Education: Principles, Policy & Practice, 25(6), 576 -597. doi:10.1080/0969594X.2018.1430685

Angelo, T. A., & Cross, P. K. (2012). Classroom Assessment Techniques: A handbook for College Teachers, 2nd Ed. San Francisco, CA: Jossey-Bass.

Beesley, A. D., Clark, T. F., Dempsey, K., & Tweed, A. (2018). Enhancing formative assessment practices and encouraging middle school mathematics engagement and persistence. School Science and Mathematics, 118(1-2), 4-16. doi:10.1111/ssm.12255

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

Berger, S., Verschoor, A. J., Eggen, T. J., & Moser, U. (2019). Development and Validation of a Vertical Scale for Formative Assessment in Mathematics. Frontier Education: Assessment, Testing and Applied Measurement, 4. doi: https://doi.org/10.3389/feduc.2019.00103

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

Black, P., & Wiliam, D. (1998b). Inside the Black Box: Raising Standards Through Classroom Assessment. Phi Delta Kappan, 80(2), 139-144. doi:10.1177%2F003172171009200119

Box, C., Skoog, G., & Dabbs, J. M. (2015). A case study of teacher personal practice assessment theories and complexities of implementing formative assessment. American Educational Research Journal, 52(5), 956-986. doi:10.3102/0002831215587754

Brookhart, S. M., & Lazarus, S. (2017). Formative assessment for students with disabilities. Council of Chief State School Officers State Collaboratives on Assessing Special Education and Formative Assessment. Retrieved from https://ccsso.org/sites/default/files/2017-12/Formative_Assessment_for_Students_with_Disabilities.pdf

Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 83(1), 1-6. doi:https://doi.org/10.1080/00098650903267784

Chappuis, S., & Stiggins, R. J. (2002). Classroom Assessment for Learning. Educational Leadership, 60(1), 40-43.

Cucos, L. (2023). How to Interpret Model Fit Results in AMOS. Retrieved 6 2023, from Uedufy: https://uedufy.com/how-to-interpret-model-fit-results-in-amos/

Darling-Hammond, K., Hyler, M., & Gardner, M. (2017, June). Effective Teacher Professional Development. Retrieved from Learning Policy Institute - Policy Brief: https://learningpolicyinstitute.org/product/effective-teacher-professional-development-report

Darrow, A. (2015). Differentiated instruction for students with disabilities: Using DI in the music classroom. General Music Today, 28(2), 29-32.

Dudec, C. M., Reddy, L. A., Lekwa, A., Hua, A. N., & Fabiano, G. A. (2019). Improving Universal Classroom Practices Through Teacher Formative Assessment and Coaching. Assessment for Effective Intervention, 44(2), 91-94. doi:https://doi.org/10.1177/1534508418772919

Dunn, O. J. (1961). Multiple Comparisons among Means. Journal of the American Statistical Association, 56(263), 52-64.

Frey, N., & Fisher, D. (2013). A formative assessment system for writing improvement. The English Journal, 103(1), 66-71.

Fukuda, S. T., Lander, B. W., & Pope, C. (July 2020). Formative assessment for learning how to learn: Exploring university student learning experiences. RELC Journal. doi:10.11772F0033688220925927

Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., Leon, V., Morrison, D., & Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to student learning: Findings from a four-year intervention study. Instructional Science, 44, 267-291. doi:https://doi.org/10.1007/s11251-016-9371-3

Gulikers, J. T., Biemans, H. J., Wesselink, R., & Van der Wel, M. (2013). Aligning formative and summative assessments: A collaborative action research challenging teacher conceptions. Studies in Educational Evaluation, 39(2), 116-124. doi:10.1016/j.stueduc.2013.03.001

He, Y. (2009). Strength-based mentoring in pre-service teacher education: A literature review. Mentoring & Tutoring: Partnership in Learning, 17(3), 263-275.

Hill, H., & Papay, J. P. (2022, October 25). Building Better PL: How to Strengthen Teacher Learning. Retrieved 6 12, 2023, from Research Partnership for Professional Learning: https://annenberg.brown.edu/sites/default/files/rppl-building-better-pl.pdf

IBM. (2021, March 2). Structure Matrix. Retrieved August 9, 2021, from SPSS Statistics Documentation: https://www.ibm.com/docs/en/spss-statistics/26.0.0?topic=predictors-structure-matrix

IBM Corp. (2021). SPSS Statistics. Armonk, NY: (RRID: SCR_016479).

Irons, A., & Elkington, S. (2021). Enhancing Learning through Formative Assessment and Feedback. London Imprint Routledge. doi:https://doi.org/10.4324/9781138610514

Laerd Statistics. (2015). Retrieved from One-way MANOVA using SPSS Statistics. Statistical tutorials and software guides: https://statistics.laerd.com/

Laerd Statistics. (2015). One-way MANOVA using SPSS Statistics. Retrieved from Laerd Statistics: Statistical tutorials and software guides: https://statistics.laerd.com/

Lee, H., Chung, H., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The Effectiveness and Features of Formative Assessment in US K-12 Education: A Systematic Review. Applied Measurement in Education, 33(2), 124-140. doi:https://doi.org/10.1080/08957347.2020.1732383

Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2020). Formative assessment as practice: the role of students' motivation. Assessment & Evaluation in Higher Education, 46(2), 236-255. doi:https://doi.org/10.1080/02602938.2020.1765228

Lyon, C. J., Nabors, O. L., & Wylie, E. (2019). Working toward integrated practice: Understand the interaction among formative assessment strategies. The Journal of Educational Research, 112(3), 301-314. doi:10.1080/00220671.2018.1514359

MacCallum, R., Widaman, K., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. Psychological Methods, 4(1), 4-99.

Matthay, E. C., & Glymour, M. M. (2020). A Graphical Catalog of Threats to Validity: Linking Social Science with Epidemiology. Epidemiology, 31(3), 376-384. doi:10.1097/EDE.0000000000001161

McMillan, J. H. (2017). "Classroom Assessment: Principles and Practice for Effective Standards-Based Instruction". Boston, MA: Pearson.

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). Retrieved from https://eric.ed.gov/?id=EJ1005135

McTighe, J., & Ferrara, S. (1998). Assessing Learning in the Classroom. Student Assessment Series. Washington, DC: National Education Association.

McTighe, J., & Ferrara, S. (2021). Assessing Student Learning by Design: Principles and Practices for Teachers and School Leaders. New York, NY: Teachers College Press.

Morgado, F., Meireles, J., Neves, C., Amaral, A., & Ferreira, M. (2018). Scale development: ten main limitations and recommendations to improve future research practices. Psicologia: Reflexão e Crítica, 30(3).

Moyo, S., Combrinck, C., & Staden, S. V. (2022). Evaluating the Impact of Formative Assessment Intervention and Experiences of the Standard 4 Teachers in Teaching Higher-Order-Thinking Skills in Mathematics. Frontier Education: STEM Education, 7. doi: https://doi.org/10.3389/feduc.2022.771437

NCES. (2022). Table 204.60. Retrieved March 15, 2022, from Digest of Education Statistics: https://nces.ed.gov/programs/digest/d20/tables/dt20_204.60.asp

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199-218. doi:10.1080/03075070600572090

NIST/SEMATECH. (2013). Retrieved from e-Handbook of Statistical Methods: https://www.itl.nist.gov/div898/handbook/

Noonan, J. (2019). An Affinity for Learning: Teacher Identity and Powerful Professional Development. Journal of Teacher Education, 70(5), 526-537.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. Contemporary Educational Psychology, 24(2), 124 -139. doi:10.1006/ceps.1998.0991

Popham, W. J. (2014). Classroom Assessment: What Teachers Need to Know. Upper Saddle River, NJ: Pearson Education Inc.

Quigley, M. (2020). The value of assessment for the inclusive general music classroom. [Master's thesis, Liberty University].
Randel, B., Apthorp, H., Beesley, A. D., Clark, T. F., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. Journal of Educational Research, 109(5), 491-502. doi:10.1080/00220671.2014.992581
Sato, M., Baker, V., Fong, E., Gilbertson, J., Liebig, T., & Schwartzfab, N. (2006). Changing Mindsets About Classroom Assessment. Science Educator, 15(1), 21-28.
Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of National Board Certification. American Educational Research Journal, 45(3), 669-700 DOI:. doi:10.3102/0002831208316955
Schneider, C., & Johnson, R. L. (2018). Using Formative Assessment to Support Student Learning Objectives. New York, NY: Routledge. doi:10.4324/9781315625805
Siegle, D. (2021). Educational Research Basics. Retrieved from UCONN: https://researchbasics.education.uconn.edu/r_critical_value_table/#
Sondergeld, T., Bell, C., & Leusner, D. (Spring 2010). Understanding how teachers engage in formative assessment. Teaching and Learning, 24(2), 72-86.
Stella, O. (2019). Discriminant Analysis: An Analysis of Its Predictship Function. Journal of Education and Practice, 10(5), 50-57.
Stiggins, R., Arter, J., Chppuis, J., & Chappuis, S. (2006). Classroom Assessment for Student Learning. Doing it right - Using it well. New York, NY: Merill Prentice Hall, Educational Testing Service.
Tabachnick, B., & Fidell, L. (2007). Using multivariate statistics (5th ed.). Upper Saddle River, NJ: Pearson.
Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International journal of medical education, 2, 53-55. doi:10.5116/ijme.4dfb.8dfd
Van den Bergh, L., Ros, A., & Beijaard, D. (2014). Improving teacher feedback during active learning: effects of a professional development program. American Educational Research Journal, 51(4), 772-809. doi:10.3102/0002831214531322
Viskupic, K., Ryker, K., Teasdale, R., Manduca, C., Iverson, E., Farthing, D., . . . McFadden, R. (2019). Classroom Observations Indicate the Positive Impacts of Discipline- Based Professional Development. Journal for STEM Education Research, 2(2), 201-228.
Warne, Z. R. (2014). A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. Practical Assessment, Research & Evaluation, 19(17), 1-10. doi:10.1.1.667.5616&rep=rep1&type=pdf
Weinbaum, E. H. (2009). Learning about assessment: An evaluation of a ten-state effort to build assessment capacity,(#RR-61). Consortium for Policy Research in Education., 43. Retrieved from https://repository.upenn.edu/cpre_researchreports/52/
Weurlander, M., Söderberg, M., Scheja, M., Hult, H., & Wernerson, A. (2012). Exploring Formative Assessment as a tool for learning: Students' experiences of Different methods of Formative Assessment. Assessment & Evaluation in Higher Education, 37(6), 747-760. doi:https://doi.org/10.1080/02602938.2011.572153
Wylie, E. C., & Lyon, C. J. (2015). The fidelity of formative assessment implementation: Issues of breadth and quality. Assessment in Education: Principles, Policy & Practice, 22(1), 140-160. doi:DOI: 10.1080/0969594X.2014.990416
Wylie, E. C., & Lyon, C. J. (2020). Developing Formative Assessment Protocol to Support Professional Growth. Educational Assessment, 25(4), 314-330. doi:https://doi.org/10.1080/10627197.2020.1766956
Yoo, W., Mayberry, R., Bae, S., Singh, K., Peter He, Q., & Lillard, J. W. (2014). A study of effects of multicollinearity in the multivariable analysis. International Journal of Applied Science and Technology, 4(5), 9-19. Retrieved from https://pubmed.ncbi.nlm.nih.gov/25664257/
Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. Applied Measurement in Education, 16(4), 323 - 342. doi:10.1207/S15324818AME1604_4
Zwart, R. C., Korthagen, F. A., & Attema-Noordewier, S. (2014). A strength-based approach to teacher professional development. Professional Development in Education, 41(3), 579-596.

Appendix A: Survey Instrument

Salem State University Institutional Review Board (IRB)

Disclosure Statement

Formative Assessment in the Classroom

Introduction: This research will ask you questions about how you use formative assessment practices in your classroom and how you would rate your skills in formative assessment. This study will:

- identify the perceptions of middle & high school teachers who are using formative assessment practices in their daily work.
- analyze teachers' self-perceived assessment skills according to formative assessment practices.
- examine the relationships among different variables, such as teachers' perceptions about formative assessment practices, years of teaching, and content areas.

Participation: Taking part in this survey is completely voluntary. You may stop your participation at any time. You are free to decline to answer any question you do not wish to answer. There are no right or wrong answers. All answers will remain completely anonymous and confidential. The results will only be used for the research proposed.

Risks: There are no foreseeable risks involved in participating in this study other than those minimal risks encountered in day-to-day life

Benefits: The benefits of your participation in this survey are self-assessment of how you use formative assessment practice in your classroom and how you can assess your skills.

Anonymity/Confidentiality: Your name or identity will not be used in reports or presentations of the findings of this research. Information provided to the researchers will be kept anonymous and confidential. This research project has been approved by the Institutional Review Board at Salem State University. ***Thank you for your help.***

For concerns about your treatment as a research participant, please contact:

Institutional Review Board (IRB) Sponsored Programs and Research Administration Salem State University

352 Lafayette Street
Salem, MA 01970
(978) 542-7556 or (978) 542-7177 or irb@salemstate.edu

Directions: These two scales contain 54 items that address issues in classroom formative assessment. For each item, please use the following rating scales to indicate (1) how frequently you use the formative assessment practice described by the item and (2) how skilled you are in using that assessment practice.

1. The scale of teachers' perceptions of using formative assessment practice (TPFAP Scale)

1. Not at all used; 2. Seldom used; 3. Used occasionally; 4. Used often; 5. Used very often.

Item In my classroom for formative assessment practices, I;	1. Not at all used	2. Seldom used	3. Used occasionally	4. Used often	5. Used very often.
1. Use paper-pencil test					
2. Use multiple-choice questions					
3. Use matching questions					
4. Use true-false questions					
5. Use fill-in-the-blank questions					
6. Use short answer questions					
7. Use essay items with restricted response					
8. Use essay items with extended response					
9. Use product performance tasks such as project					
10. Use product performance tasks such as portfolio					
11. Use product performance tasks such as video					
12. Use product performance tasks such as spreadsheet					
13. Use product performance tasks such as a web page					
14. Use product performance tasks such as exhibition					
15. Use product performance tasks such as reflection					
16. Use product performance tasks such as journal					
17. Use product performance tasks such as graph					
18. Use product performance tasks such as illustrations					
19. Use skills performance tasks such as speech					
20. Use skills performance tasks such as demonstration					
21. Use skills in performance tasks such as debate					
22. Use skills performance tasks such as readings					
23. Use oral questioning such as informal questioning					
24. Use oral questioning such as examinations					
25. Use oral questioning such as conferences					
26. Use oral questioning such as interviews					
27. Use test items for higher cognitive levels					

2. The scale of teachers' self-perceived assessment skills according to formative assessment practices. (TSPFAP Scale)

1.Not at all skilled; 2. A little skilled; 3. Somewhat skilled; 4. Skilled; 5. Very skilled

Item In my classroom for formative assessment practices, when I;	1. I am not at all skilled	2. I am a little skilled	3. I am somewhat skilled	4. I am skilled	5. I am very skilled
28. write a paper-pencil test,					
29. write multiple-choice questions,					
30. write matching questions,					
31. write true-false questions,					
32. write fill-in-the-blank questions,					
33. write short answer questions,					
34. write essay items with restricted responses,					
35. write essay items with extended responses,					
36. use product performance tasks such as projects,					
37. use product performance tasks such as portfolio,					
38. use product performance tasks such as video, I am					
39. use product performance tasks such as spreadsheets,					
40. use product performance tasks such as web page					
41. use product performance tasks such as exhibition					
42. use product performance tasks such as reflection					
43. use product performance tasks such as journal					
44. use product performance tasks such as graph					
45. use product performance tasks such as illustrations					
46. use skills performance tasks such as speech					
47. use skills performance tasks such as demonstration					
48. use skills in performance tasks such as debate					
49. Use skills performance tasks such as readings					
50. use oral questioning such as informal questioning					
51. use oral questioning such as examinations					
52. use oral questioning such as conferences					
53. use oral questioning such as interviews.					
54. write test items for higher cognitive levels					

3. Demographic information

1. What is the highest degree you have?
_____Bachelor's degree (BA, BS, etc.)
_____ Master's degree (MA, MS, M. Ed, etc.)
_____Doctoral Degree (Ph.D., Ed. D, etc.)
2. What is your field of teaching? _____

3. How long have you been teaching? _____years
4. In the last five years did you receive any training regarding formative assessment? _____Yes _____No

Thank you for your time!
For any comment, question, or suggestion, please feel free to contact the researcher.

Appendix B: Tests of MANOVA Assumptions

The first assumption that was tested was whether there were univariate or multivariate outliers in the dataset. Scores on each of McTighe's categories were examined using the SPSS Boxplot procedure (see Figure A1). Overall, there were only four outliers on the univariate outcome variables. There was one outlier from the STEM group on the Selected Responses Subscale and two outliers on this subscale from the Humanities group. There was one outlier from the STEM group on the Constructed Response Subscale and one on the Products Subscale. There were no other outliers for the Humanities group. The multivariate examination of outliers was based on Mahalanobis distance values generated via one of the components of the SPSS Regression procedure. The largest Mahalanobis value was 15.41. Using a commonly agreed-upon cutoff level of 20.52, there were no multivariate outliers (Laerd Statistics, 2015). Given the combination of a few outliers and a relatively small sample size, it was decided not to delete any of these outliers. A follow-up MANOVA was done with these outliers excluded listwise to check for any possible impact on the results. The data show that excluding these outliers had no meaningful impact on any aspect of the results.

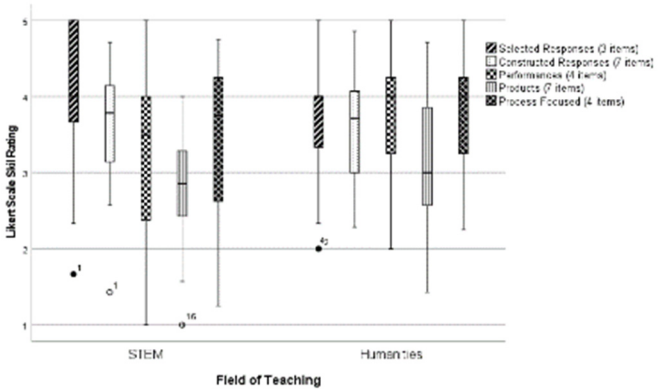


Figure A1: Box and Whiskers plot of teachers' skill ratings on McTighe's subscales by field of teaching

The next MANOVA assumption checked was for evidence of multivariate normality. Using the SPSS procedure to Test Normality, the Shapiro-Wilk results suggest that the normal distribution was violated for most of the subscales by field of teaching, 3 of the 5 subscales for STEM teachers, and 4 of the 5 scales for Humanities teachers (see Table A1). This is not surprising since most of the distributions were negatively skewed. (See skewness values in Tables A2, A3, & A4). Subsequent visual inspection of the SPSS Q-Q plots did not reveal any major distortions of the data. Thus, it was decided to move on with the planned MANOVA analyses even with the significant Shapiro-Wilk results.

Table A1: Shapiro-Wilks (S-W) Test of Normality for each of McTighe and Ferrara's subscales

	STEM			Humanities		
Subscale	S-W Statistic	df	Sig.	S-W Statistic	df	Sig.
Selected Responses (3 items)	.875	28	.003	.898	63	<.001
Constructed Responses (7 items)	.932	28	.067	.959	63	.035
Performances (4 items)	.947	28	.165	.962	63	.050
Products (7 items)	.968	28	.537	.976	63	.255
Process Focused (4 items)	.894	28	.008	.947	63	.008

Table A2: Check for linearity among STEM teachers' skill ratings across McTighe and Ferrara's subscales – Pearson correlations

	STEM teachers' skill ratings							
Subscale	M	SD	Skewness	1	2	3	4	5
Selected Responses (3 items)	4.01	0.94	-0.82	-	.601**	.257	.358	.318
Constructed Responses (7 items)	3.67	.076	-0.89		-	.751**	.608**	.749**
Performances (4 items)	3.23	1.11	-0.47			-	.700**	.849**
Products (7 items)	2.81	0.69	-0.61				-	.632**
Process Focused (4 items)	3.42	1.07	-0.80					

Note: * p < .05 ** p <. 01 (two-tailed) n = 28 (listwise deletion)

Table A3: Check for linearity among Humanities teachers' skill ratings across McTighe and Ferrara's subscales – Pearson correlations

	Humanities teachers' skill ratings							
Subscale	M	SD	Skewness	1	2	3	4	5
Selected Responses (3 items)	3.91	0.76	-0.27	-	.541**	.155	.190	.228
Constructed Responses (7 items)	3.67	0.76	0.12		-	.631**	.579**	.737**
Performances (4 items)	3.59	9.91	-0.20			-	.187	.676**
Products (7 items)	3.14	.078	0.11				-	.386**
Process Focused (4 items)	3.79	0.81	-.027					-

Note: * p < .05 ** p <. 01 (two-tailed) N = 63 (listwise deletion)

Table A4: Check for multicollinearity among teachers' skill ratings across McTighe and Ferrara's subscales – Pearson correlations

Subscale	M	SD	Skewness	1	2	3	4	5
Selected Responses (3 items)	3.94	0.82	-0.468	-	.564**	.177	.224*	.249*
Constructed Responses (7 items)	3.60	0.72	-0.278		-	.631**	.559**	.771**
Performances (4 items)	3.59	0.91	-0.608			-	.393**	.764**
Products (7 items)	3.04	0.77	0.150				-	.481**
Process Focused (4 items)	3.68	0.91	-0.663					-

The third MANOVA assumption that was tested was for evidence of multicollinearity. Using the SPSS Bivariate procedure, the Pearson correlation coefficients ranged from a minimum of .177 to a maximum of .771 (See Table A4). Thus, there were no extreme instances of multicollinearity, i.e., none of the correlations between subscales were above .90, the typical level at which multicollinearity would impact MANOVA analyses. However, there were 2 Pearson correlations involving the Process Focused subscale that were .70 or higher which could have impacted the MANOVA (Yoo, et al., 2014).

Next, the data were examined for evidence of a linear relationship between the dependent variables, i.e., McTighe's Subscales for the independent variable, i.e., Field of Teaching. Visual examination of SPSS Scatterplots was followed by calculating Pearson Correlation coefficients among McTighe's subscales for the STEM group and the Humanities group. Among STEM teachers, linear relationships were seen among all pairs of subscales except between the Selected Responses subscale and three other subscales, i.e., Performances, Products, and Process Focused subscales. All the other pairings showed statistically significant correlations, i.e., df = 26, p >.374. (See Tables A2 and A3.) A somewhat similar pattern was found for Humanities teachers with a lack of linear relationships between the Selected Responses subscale and three other subscales, i.e., Performances, Products, and Process Focused subscales. However, among Humanities teachers there was also a non-linear relationship between the Performances subscale and the Products subscale. For purposes of the MANOVA, these limitations due to instances of non-linearity were noted and are discussed below in the limitations section.

The homogeneity of variance-covariance matrices assumption was tested using the Box Test of Equality of Covariance Matrices. The observed p-value with df =15 was .035 which is larger than the traditional probability cutoff of p < .001 (Laerd Statistics, 2015). Thus, this assumption was not violated.

The final assumption for the MANOVA relates to the homogeneity of variances. To test this assumption, Levene's test of medians was used (NIST/SEMATECH, 2013). None of the subscale median variances were statistically significant and thus did not violate the assumption of homogeneity. (See Table A5.)

Table A5: Levene's Test of Equality of Error Variances based on median.

Subscale	Levene statistic	df	p
1. Selected Responses (3 items)	1.519	1, 89	.270
2. Constructed Responses (7 items)	0.184	1, 89	.732
3. Performances (4 items)	3.151	1, 89	.079
4. Products (7 items)	1.249	1, 89	.267
5. Process Focused (5 items)	1.713	1, 89	.194

As noted above, there were a limited number of MANOVA assumptions that were violated. Follow-up analyses and a review of research related to violating these assumptions indicated that while the violations may have limited the power of the MANOVA (Tabachnick & Fidell, 2007), the researchers felt that the MANOVA could still possibly reveal useful information about this sample of teachers' skills based on McTighe and Ferrara's conceptual framework for FAs. Table A6 contains a summary of the tests of MANOVA assumptions and potential impact on subsequent analyses.

Table A6: Summary of Results of Testing Assumptions for MANOVA relations

Assumption	Type of violation if any	Follow-up
No Outliers	4 univariate outliers No multivariate outliers	A MANOVA was done without these outliers with similar results.
Multivariate normality	Most subscales were not normal	Q-Q plots did not reveal any major distortions.
No Multicollinearity	No extreme instances of multicollinearity but there were 2 instances of moderate Pearson r on the process-focused subscale	None needed
Linear Relationships Between DVs by Field of Teaching	Quite a few non-linear relationships	Tabachnick & Fidell (2007), the "failure of linearity of residuals in regression does not invalidate an analysis as much as weaken it" (p. 127)
Homogeneity of Variance-Covariance Matrices	Not violated	None needed
Homogeneity of Variances.	Not violated.	None needed