International Journal of STEM Education

RESEARCH

Open Access

Measures of success: characterizing teaching and teaching change with segmented and holistic observation data

Timothy J. Weston[1*], Sandra L. Laursen[2] and Charles N. Hayward[2]

Abstract

Background Numerous studies show that active and engaging classrooms help students learn and persist in college, but adoption of new teaching practices has been slow. Professional development programs encourage instructors to implement new teaching methods and change the status quo in STEM undergraduate teaching, and structured observations of classrooms can be used in multiple ways to describe and assess this instruction. We addressed the challenge of measuring instructional change with observational protocols, data that often do not lend themselves easily to statistical comparisons. Challenges using observational data in comparative research designs include lack of descriptive utility for holistic measures and problems related to construct representation, non-normal distributions and Type-I error inflation for segmented measures.

Results We grouped 790 mathematics classes from 74 instructors using Latent Profile Analysis (a statistical clustering technique) and found four reliable categories of classes. Based on this grouping we proposed a simple proportional measure we called Proportion Non-Didactic Lecture (PND). The measure aggregated the proportions of interactive to lecture classes for each instructor. We tested the PND and a measure derived from the Reformed Teaching Observation Protocol (RTOP) with data from a professional development study. The PND worked in simple hypothesis tests but lacked some statistical power due to possible ceiling effects. However, the PND provided effective descriptions of changes in instructional approaches from pre to post. In tandem with examining the proportional measure, we also examined the RTOP-Sum, an existing outcome measure used in comparison studies. The measure is based on the aggregated items in a holistic observational protocol. As an aggregate measure we found it to be highly reliable, correlated highly with the PND, and had more statistical power than the PND. However, the RTOP measure did not provide the thick descriptions of teaching afforded by the PND.

Conclusions Findings suggest that useful dependent measures can be derived from both segmented and holistic observational measures. Both have strengths and weaknesses: measures from segmented data are best at describing changes in teaching, while measures derived from the RTOP have more statistical power. Determining the validity of these measures is important for future use of observational data in comparative studies.

Keywords Structured Observation, Undergraduate STEM Education, Teaching Practices, Educational Measurement, Science Teaching Reform

*Correspondence:
Timothy J. Weston
westont@colorado.edu
Full list of author information is available at the end of the article

Introduction

Numerous studies show that active, engaging and collaborative classrooms help students learn and persist in college STEM courses and degree programs, but instructors' adoption of new teaching practices has been slow (American Association for the Advancement of Science (AAAS), 2013; Laursen et al., 2019; Matz et al., 2018). In a recent study, observations of 2008 STEM classes at 24 institutions found that most courses were primarily lecture-based, with only a small proportion of classes incorporating significant amounts of student-centered learning (Stains et al., 2018). Professional development programs are an important lever intended to help instructors implement new teaching methods and change the status quo in STEM undergraduate teaching (Laursen et al., 2019; Manduca et al., 2017). However, learning whether these programs change teaching practices is challenging, because typical means of measuring teaching, such as surveys, student testing, and classroom observations, all have methodological shortcomings and may be difficult to implement (AAAS, 2013; Ebert-May et al., 2011; Weston et al., 2021). Without reliable evidence of the efficacy of interventions it is difficult to confidently implement and improve professional development.

While observation data are often perceived as more objective than self-report data from surveys or interviews (AAAS, 2013), data derived from observational studies pose particular challenges when used in statistical tests, thus complicating the ability to make claims about the efficacy of professional development and other interventions (Bell et al., 2012). Some observational systems also may lack clarity in their descriptions of teacher and student activities, making it difficult to learn how instruction has changed over time and what exactly changed in the teaching practices of participants (Lund et al., 2015). Because observation is resource-intensive, investigators often observe only a small number of sessions, which may not provide a representative sample of teaching practices across an entire course (Weston et al., 2021). This in turn limits the utility of observations to characterize instructional practice or to inform efforts to improve instruction.

The current study addresses these analytical challenges of using classroom observation data to characterize teaching, particularly with the intent to measure change in teaching and to describe those changes. The study focuses on structured observational protocols and their statistical and descriptive characteristics in assessing instructional change in undergraduate STEM teaching. We examine one descriptive and segmented protocol, the TAMI-OP (Toolkit for Assessing Mathematical Instruction-Observation Protocol), and one evaluative and holistic protocol, the RTOP (Reformed Teaching Observation Protocol). While the study results apply strictly to these protocols alone, each serves as an exemplar of its type, and the analysis highlights issues that scholars should consider in using structured observations for evaluation and research on teaching.

The statistical and descriptive utility of observational protocol measures

Many structured observation protocols are available to capture what occurs in undergraduate STEM classrooms. Observational protocols are either holistic or segmented (Hora & Ferrare, 2013), with segmented protocols asking observers to mark the presence of specific teaching and learning activities, usually in 2-min intervals. In contrast, when using holistic protocols like the RTOP and MCOP^2, observers take structured notes throughout the class and at the end of the class, rate the class on Likert-scaled survey items (Gleason et al., 2017; Sawada et al., 2002). Some observational protocols include the TDOP (Hora et al., 2013), RTOP (Piburn et al., 2000), COPUS (Smith et al., 2013), MCOP^2 (Gleason et al., 2017), CLASS-S (Hamre & Pianta, 2005), the 3D-LOP (Bain et al., 2020), and the TAMI-OP (Hayward et al., 2018), among others. Protocols may be customized for different disciplines such as mathematics (Hayward et al., 2018) or for alternative classroom contexts such as course-based undergraduate research experiences (CUREs) (Esparza et al., 2020).

Observational systems are by their nature descriptive but are also utilized to evaluate instruction and learning throughout all levels of STEM education. Observational data are used in the individual evaluation of instructors (Whitehurst et al., 2014), the evaluation of professional development efforts (Stains et al., 2015), and validity studies of surveys (Ebert-May et al., 2011). Researchers also have used observations to assess interventions such as teacher–scientist collaboration (Campbell et al., 2012), co-teaching (Beach et al., 2008), changes to teaching style (Budd et al., 2013), and the efficacy of pedagogies in encouraging student engagement (Lane & Harris, 2015). All these studies examined reform-based teaching and learning activities in the classroom, usually in STEM disciplines.

While a growing number of educational researchers use observational data for outcome studies, there are two primary concerns about these methods that we explore in this paper. These concerns have a direct impact on decisions about the efficacy of interventions such as professional development or the implementation of new curricula in STEM. The first is the psychometric qualities of measures derived from observational data and their suitability for use in comparative studies involving statistical hypothesis testing. These are basic validity concerns shared with any use of quantitative measures and are

important, because claims for the efficacy of programs and interventions are made with these data (Kane, 2012). These concerns focus on the reliability of a measure, potential bias, and the validity of aggregate measures or subscales. For statistical comparisons it is also important to learn a measure's distributional characteristics, and if it can be used in common statistical tests with enough statistical power to provide valid inferences (Glass & Hopkins, 1996).

The second important component of a valid observational measure is an aspect of score validity called *extrapolation*, "the degree to which scores from the observational protocol are related to a broader conception of teaching quality" (Bell et al., 2012, p. 68). Observational protocols—and any measures derived from them—ideally should provide clear descriptions of how an instructor is teaching and the activities students are engaged in during class (Cash et al., 2012; Hora & Ferrare, 2014). For use in professional development, it is vital that the measure offers diagnostic feedback to program facilitators on the effects of their training while providing information to inform changes to workshop design or lesson plans that could help improve training (Egert et al., 2018). Critical for the purposes of the current study, we wanted to know if observations contain sufficient descriptive information, so that those responsible for implementing professional development can understand what specific teaching and learning activities changed from pre to post, and if teaching style changed for the better.

A dependent measure from the observation protocol should also provide an accurate evaluation of teaching quality (Bell et al., 2012). Some observational protocols have quality assessments baked into the design of measures that embrace a set of criteria and standards for what counts as good instruction (Cash et al., 2012; Gleason et al., 2017). For program development in undergraduate mathematics education, desired instruction includes teaching that is non-didactic, engaging, active, inquiry-based, and authentic. In classrooms this commonly translates to practices, where the teacher is a facilitator who fosters more group work, more dialogue, and an emphasis on conceptual learning (National Research Council, 2012).

Some holistic protocols, like the RTOP and MCOP[2], embed evaluation criteria such as those found in national standards (National Council of Teachers of Mathematics, 2022) into the design of their items, asking observers to make assessments of the quality of teaching practices, while they observe. For instance, an RTOP item asks the observer to rate the instructor to determine if "The lesson was designed to engage students as members of a learning community" (Sawada et al., 2002, p. 253). These types of holistic ratings can be guided by rubrics that attempt to anchor judgements in observable frequencies of behavior ("over half of the students…") or proportion of class time devoted to an activity ("students spend two-thirds or more of the lesson…") (Gleason et al., 2017, p. 124). Asking observers to make these assessments can be problematic given that observer's beliefs, previous experience, and expertise can all bias observations, in turn making it more difficult for observers to reliably agree on ratings (Cash et al., 2012). Hora and Ferrare (2013) commented on how judgments of teaching quality are built into the RTOP while noting its poor descriptive qualities:

> *Because the Reformed Teaching Observation Protocol is based on underlying scales of instructional practice (e.g., classroom culture) and a priori determinations of instructional quality, …the resulting data do not provide descriptive accounts of teaching but instead prejudge which practices are effective and which are not. (p. 218).*

In contrast, for most segmented observations such as the COPUS, the "raw" observation data is meant to be (mostly) value-free with determinations of quality coming *after* the observation. Researchers use the data and impose a model of observed behaviors thought to represent good teaching (Smith et al., 2013; Stains et al., 2018). Observers code each 2-min time-period for activities, such as lecture and group work, and then assess if the resulting description of teaching and learning fits an inquiry-based, authentic, or active model of instruction. Such protocols emphasize capturing behaviors rather than evaluating quality. This can be a drawback for segmented protocols; even if the surface elements of good instruction are present, it is not guaranteed that activities were implemented effectively and that students are learning (Borda et al., 2020).

While descriptive observational protocols can provide thick descriptions of teaching, it can be difficult to summarize and characterize teaching styles with only the raw data. The primary way of organizing and categorizing segmented observational data about teaching and learning characteristics are statistical clustering or latent profile/class methods (Spurk et al., 2020). For analyses of teaching, clustering methods identify classes that fit different categories, each of which represents a set of underlying teaching and learning methods. The largest survey so far of STEM courses using observational methods, conducted by Stains et al. (2018), used Latent Profile Analysis (LPA) to classify instructional styles for 2008 classes at 24 universities using the COPUS protocol. The observational survey found three broad types of teaching styles that the researchers characterized as didactic, interactive lecture, and student-centered classes, and

they described how the proportions of these styles varied across class size, level, discipline, and physical layout. Lund et al. (2015) classified teaching styles using RTOP and COPUS data with cluster analysis for 269 chemistry classes. These researchers found 10 clusters for differing types of lectures, Socratic, peer, and group work instruction, which were then classified into three broad categories for mostly lecture, emergence of group work, or extensive group work. Denaro et al. (2021) compared clustering techniques on 250 classes. Across all three studies, didactic and (to a lesser degree) interactive lecture were found to be the dominant teaching styles. These results shed light on the general state of teaching reform in STEM education but also reflect an important way to organize observational data about teaching.

Statistical and descriptive components of the RTOP

Researchers and evaluators have used both segmented and holistic measures to categorize classroom teaching styles and in other research about teaching (Budd et al., 2018), but they are also utilized as outcome variables to assess the impact of interventions, such as workshops and other professional development efforts. Many studies that employ observational data to assess teaching change in response to interventions use the Research Teaching Observation Protocol (RTOP), a holistic observational measure (Sawada et al., 2002). An overview reveals a mixed picture for the validity of the RTOP for some uses. While it has high internal reliability and some criterion validity, the measure seems to lack structural score validity in that its proposed sub-scales did not form separate factors in the original validity study (Piburn et al., 2000). When this occurs, composites formed from subscales may be highly correlated and are essentially the same measures with different names (Marsh et al., 2019). This is problematic, because some studies have used RTOP subscales to make research claims (Budd et al., 2018; Emery et al., 2020). However, a unitary summed measure using all RTOP items was found to correlate with other measures of inquiry-based or active learning and to function psychometrically as a reliable continuous measure (Sawada et al., 2002).

Those using the measure also seem limited in their ability to extrapolate from scores to descriptions of teaching. These descriptions are important to understanding what changes did or did not occur after an intervention. Holistic protocols like the RTOP, by their nature, are less able to support researchers in developing detailed descriptions of teaching and learning activities (Hora & Ferrare, 2013). This lack of descriptive utility for the RTOP was discussed by Lund et al. (2015), who noted that the same score ranges can describe classes with very different instructional practices and varied even more widely from study to study. Descriptions of score ranges were provided by Sawada and coauthors (2003) and were subsequently adapted by Ebert-May et al. (2011). These score range categories are described as:

> *…0–30 Straight lecture, 31–45 Lecture with some demonstration and minor student participation, 46–60 Significant student engagement with some minds-on as well as hands-on involvement, 61–75 Active student participation in the critique as well as the carrying out of experiments, 76–100 Active student involvement in open-ended inquiry, resulting in alternative hypotheses, several explanations, and critical reflection. (p. 555)*

While these categories provide some descriptions of teacher and student activity scoring standards for each score range (especially for lecture), it seems difficult to fit some of the descriptions such as "significant student engagement with some minds-on as well as hands-on involvement" to observed teacher and student behaviors. In our current study we also examine the relationship between RTOP score categories and segmented observations.

Some of the challenges of using the RTOP to describe the outcomes of comparative studies are illustrated by Adamson et al. (2003), who assessed the effects of attending summer professional development classes for pre-service teachers when they later became secondary school teachers. These researchers used the RTOP to compare participating and comparison group teachers and found significant statistical differences between groups, favoring the intervention; however, it was unclear which of the teachers' activities were responsible for the higher scores. While the authors described the program in detail and its desired outcomes, there were few details about which reformed teaching practices were adopted.

The same lack of detailed description is seen in a recent study about long-term effects of STEM teaching professional development (Emery et al., 2020). The researchers used RTOP scores (alongside other measures) to compare those participating in a postdoctoral professional development program with instructors in a comparison group, tracking both groups of study participants into faculty jobs after 6–10 years beyond the professional development. In addition, those in the participating group were compared with paired faculty in their departments. Gains from the professional development program persisted over time, and participating faculty had higher RTOP scores than the paired members in their academic departments. In the faculty comparison, the authors describe the difference between groups as those in the participating group with mainly level-3 scores ("significant student engagement") with the comparison

group who taught at level-2 ("primarily lecture with some demonstration"), a remarkable finding about the efficacy of the program. However, it was not clear from the RTOP (or the ATI survey also administered) which specific behavioral and observable practices (e.g., group work) were being implemented in the classrooms studied—only that the professional development group was, overall, more interactive than the paired comparison group.

Statistical and descriptive components of segmented observations

While the RTOP is a highly reliable measure that can be readily used in statistical procedures, it appears to be limited in its ability to describe in detail what instructors and students are doing in their classrooms. In contrast, segmented protocols would appear to be better at providing material that supports thicker descriptions of teaching practices but can be awkward to work with statistically, complicating inferences about the efficacy of interventions.

Like the RTOP and other holistic protocols, segmented protocols such as the COPUS are also employed in comparative research designs but can pose measurement challenges. Difficulties arise in using segmented observational protocols for several reasons. The use of single observation codes (such as the proportion of class time devoted to lecture) can result in poor and incomplete representation of the complex underlying instructional styles occurring in the classroom (Bell et al., 2012). This is partly remediated using composite, aggregate, or collapsed measures that join two or three codes together, as are used in the COPUS (Smith et al., 2013). While aggregates are better representations of underlying teaching styles, single codes (like single survey or test items) tend to have low or no psychometric reliability, and collapsed codes created from two or three items generally have very low internal reliabilities (Nunnally & Bernstein, 1994). Segmented data may also have unwieldy distributional characteristics. The distributions of many relatively low-frequency codes are dramatically skewed, with high numbers of zero observations for any given classroom, and skewed distributions are also common when aggregated over multiple classrooms and instructors (Tomkin et al., 2017). While non-parametric tests are often preferable to inferential statistical tests for many reasons (e.g., when sample sizes are small or data are drawn from convenience samples), the distributional properties of segmented observational data may necessitate the use of non-parametric tests, which in turn cause possible loss of statistical power (Dwivedi et al., 2017).

Another concern is a result of the high number of codes generated by segmented protocols compared to a holistic protocol's single composite score or few sub-scale scores. When multiple hypothesis tests (e.g., multiple t tests) are made in the same study, the true probability of making Type-I errors (saying there is a difference when one does not exist) increases substantially (Abdi, 2007). Abdi provides an example of this phenomena with repeated trials of 20-coin tosses. The probability of seeing a rare (but still random) combination of 14 heads and 6 tails is 5% with one trial, but if the experiment of 20-coin tosses is attempted 10 times this rare event becomes much more likely at 40%. This is analogous to hypothesis testing in that the outcome occurred in the absence of any systematic manipulation of the coin; in hypothesis testing the p value is the probability that a large mean difference is observed when in fact there really is no difference in the population. This hidden inflation applies to multiple comparisons of single activity codes such as comparing the frequency of lecture, group work, and student presentation in the same set or family of analyses. It is possible to adjust for inflated p values using the Bonferroni or Sidak corrections, but this makes it much more difficult to find significant differences.

Some of these statistical problems can be seen in research studies using segmented protocols. A quasi-experimental study of learning communities conducted by Tomkin et al. (2019) used the COPUS to compare 25 instructors who took part in learning communities and 35 instructors who did not participate. Metrics derived from the observations collapsed categories of codes to consider the amount of time devoted in each group for teachers presenting and guiding, and students receiving, talking, and working. The researchers used a battery of specialized statistical techniques to make the comparison, including non-parametric tests, Poisson and Zip regression, Kruskal–Wallis tests, and Shapiro–Wilk tests to check for normality. While these tests were appropriate for the data, their use added a layer of complexity to the analysis complicating the ability to draw clear inferences from the study. The researchers also used seven comparisons of collapsed codes and twelve comparisons of single codes from the COPUS, and while some significant differences were found between groups, the comparison of multiple single and collapsed codes left open the possibility of inflating Type-I error. In addition, the use of individual activity codes lacked psychometric reliability.

In comparison with the holistic protocols, segmented observations while providing the details needed for thick descriptions of teaching practices, seem to be difficult to work with statistically. This can obscure inferences made about whether and how teaching changed in response to an intervention.

Rationale for the study

In the current study, we consider two observation protocols, TAMI-OP and RTOP, evaluating their characteristics as measures on their own merits while also recognizing them as typical examples of segmented and holistic protocols. These protocols are also distinguished by their descriptive and evaluative approaches and are used often in research and evaluation about STEM teaching and learning. In our current study, we worked from a large dataset that included observations scored with both the TAMI-OP and the RTOP. We asked if a simplified measure formed from a segmented observational protocol, TAMI-OP, could be used with common statistical tests and avoid multiple comparisons while maintaining score validity. We also reexamined the structural validity of the RTOP (a popular example of a holistic and evaluative protocol) and assessed how the RTOP's measures worked in statistical tests with actual pre/post data.

The following questions are addressed in this study:

1) What are the characteristics of profile groups for classes that can be derived from our TAMI-OP observational dataset of mathematics instructors?
2) What dependent measures can be derived from the TAMI-OP?
3) What is the validity of the RTOP sub-scales and its validity when used as an aggregate measure?
4) How do two dependent measures derived from observation data—the RTOP-Sum and the measure based on segmented TAMI-OP observations—function with statistical tests in terms of statistical power and the distributional characteristics of the measures?
5) How can the RTOP aggregate dependent measure and the segmented TAMI-OP dependent measure be extrapolated to provide descriptions of teaching and teaching change?

To address these questions, we coded a large data set of observations from college mathematics courses with both protocols, including multiple observations per course. We first describe the instruments and the data set, then the analyses conducted and the results responding to each of the five research questions.

Methods

Instruments

Our observational protocol began as part of a broader study matching survey responses to observational data. After reviewing various observation protocols, we started with the COPUS (Smith et al., 2013), which draws heavily from the TDOP protocol (Hora et al., 2013). We modified these protocols to reflect teaching practices common in undergraduate mathematics classrooms, but kept the TDOP's segmented, descriptive approach. The resulting protocol is the Toolkit for Assessing Mathematics Instruction-Observation Protocol (TAMI-OP) (Hayward et al., 2018). At 2-min intervals during the class, observers coded for the presence (yes/no) of 11 student behaviors and 9 instructor behaviors. We called these categories *activity codes* or more generally, observation *items*. In addition, observers counted the frequencies of student and instructor questions and answers (for details of these and other activity codes see Hayward et al., 2018). We also completed the RTOP for a subset of 484 of the same classes in the study.

We used a pool of 11 activity codes in our analysis which represented the activities seen in the classrooms we studied. We also used three indicators of global teaching style called *Number of Activities*, *Later Lecture*, and *Between Activity Variance*. These additional measures quantify the number, balance, and sequence of teaching and learning activities and were derived to quantify persistent patterns we observed among classrooms. For *Number of Activities,* we counted the number of teaching and learning activity codes present in a class. The *Between-Activity Variance* described the balance in duration of activities within a class; a high variance represented a lopsided balance (usually a lot of lecture and very little other activities), and a low variance represented a class with similar amounts of time devoted to each activity. (In analyses, we used the inverse of this variance). The *Lecture Later* variable quantified when instructors did activities during class time, with instructors either lecturing during the first quartile of class time or starting later. Again, these variables seemed to represent observed patterns of teaching that distinguished groups of classes. Tables 1 and 2 describe activity codes and global indicators used either directly in our Latent Profile Analysis (LPA), or as descriptors of each group generated by LPA.

Sample

Our full dataset contained 790 observations of full classes by 74 teachers, gathered from three different research studies related to professional development in mathematics teaching. One of the studies was conducted to develop tools for evaluating outcomes of professional development programs representing 297 observations of full classes by 34 teachers. A second study examined teaching of early career mathematics faculty who participated in teaching-related professional development; the observation sample from this study represents 215 observations of 24 teachers, prior to their program participation.

Table 1 Activity codes used in latent profile analysis

Activity code	Description
Lecture	Instructor lectures about novel content
Reviewing content	Instructor reviews students' previous work (e.g., homework, group activity)
Student presentation	Students present problems at the board or whiteboard
Real time writing by instructor	Instructor writes on board, overhead or whiteboard
Moderating and inviting participation	Instructor solicits student comments, manages student discussions
Moving and guiding	Instructor works with students in groups
Student question	Students ask questions of teacher
Instructor asks informational question	Instructor question asking for specific information or answer
Instructor asks for reasoning	Instructor question asks for students to explain an answer to a mathematics problem

Table 2 Global indicators used as profile descriptors

Global indicator	Description
Number of activities	The number of activity codes present during one class
Between-activity variance	Summed deviations of combinations of activities in one class averaged by number of activities. One minus this term is used to characterize classes as having more balance between activities. Single activity classes are given a value of 0.
Later Lecture	Indicates if lecture started after the first quartile of class time
EOC factor score	Composite factor score of 13 end-of-class items used to characterize classes. Survey was designed by the study team

Table 3 Types of courses

Type of course	Number of courses	Percentage
Algebra	7	10
Calculus 1	11	16
Calculus 2	9	13
Upper division	7	10
Discrete math	5	7
Geometry	6	9
General education math	14	21
Pre-calculus	3	4
Statistics	6	9
Total	68	100

Six courses missing descriptions

Finally, the third study examined changes in teaching among instructors who participated in a professional development workshop. The observation sample from this study includes 15 instructors who taught 278 classes, some pre- and some post-intervention. The results for these instructors are used as an example of how these measures characterize teaching change but are not meant to offer a formal assessment of that program.

The instructors in the combined data set taught a range of mathematics courses at different undergraduate levels. Classes included calculus 1 and 2, geometry, general education mathematics, statistics, and upper division courses for math majors (see Table 3 for full description). Class sizes ranged from 30 or less (65%), 31 to 75 (25%) to over 100 (10%). The instructors included women and men, experienced and early career instructors; they taught at a variety of types of institutions distributed across the US and used a variety of teaching practices. While we do not claim the sample is generally representative of mathematics instruction in US higher education, we do believe the sample captures the range and variation of such instruction.

Latent profile analysis

Latent Profile Analysis (LPA) is a statistical classification technique that identifies subpopulations or groups within a population based on a set of continuous variables (Spurk et al., 2020). In many cases the technique is used to classify people but can be extended as it was in Stains et al. (2018) to group together classes. LPA is similar but preferable to traditional cluster analysis, because it offers the ability to assess the ideal number of groups in a solution and generate probabilities of group membership, which provide estimates of how close any given case is to a profile exemplar (Ferguson et al., 2020).

The software R-Studio 3.5.0 was used to conduct LPA for the 790 classes in our database. Packages used in R included *mclust* and *tidyLPA* (both versions 3.5.3). The component variables for analysis all used

class-level proportions of activity codes and global indicators. While these variables are continuous, most did not form normal univariate distributions. We used a Maximum Likelihood Estimation (MLE), and tested models with different constraints on variance and covariance. Best fitting models used estimation with equal variances and covariance equal to zero. No outliers were found or removed from the data, and there were no missing data.

We sought to follow best practices in choosing the number of factors based on fit statistics and logical coherence of resulting groups (Spurk et al., 2020). To specify models, we worked iteratively to reach a coherent solution starting with only frequently used activity codes, then adding less frequently used codes and global indicators. We assessed the model each time to examine the solutions for number of groups, fit statistics, the relative influence of individual variables, and coherence. The resulting model included only activity codes shown in Table 1 but not the global indicators found in Table 2. Fit statistics and the resulting groupings are presented in the Results section. We also validated group membership by randomly selecting 30 classes (blind to actual profile) and classifying each based on average scores of activity codes and global indicators; 27 out of 30 were correctly identified.

Reliability of observations

Observers established high interrater reliability through training for each project. Interrater agreement was high across all TAMI-OP items (> 95%). We also examined the reliability of scores using Generalizability Theory given the number of nested classes for each teacher (Weston et al., 2021). We observed between four and twelve classes for each teacher. Overall generalizability for the sample was greater than $G = 0.8$, indicating that enough classes were sampled during a semester course for a reliable measure. Interrater reliability was also high, again with Generalizability coefficients higher than $G = 0.8$ for both TAMI-OP and RTOP. Reliability for the RTOP was also established through assessment of the internal reliability coefficient alpha as well as inter-rater agreement.

Results

Characteristics of LPA profile groups

We found four reliable profiles that characterize the 790 mathematics classes in our sample. We determined the ideal number of profiles through a balance of quantitative fit indexes and the logical coherence of the resulting groupings. Standards for good fit included statistics for Entropy greater than 0.9, BLRT probabilities greater than 0.8, and lower AIC and BIC values when adding additional profiles (Spurk et al., 2020). While some indices improved slightly in five- and six-cluster solutions, examination of these solutions did not show obviously different characteristics among new groupings, indicating that adding groups did not substantially differentiate classes based on the variables in the model or other descriptors. Table 4 presents model fit statistics for our LPA.

We named profiles for the variables that best differentiated between groups, resulting in profiles named *Didactic Lecture, Student Presentation and Review, Interactive Lecture,* and *Group Work*. These are described in Table 5 and profiles for each are shown in Figs. 1 and 2. Global indicators were also entered into the LPA models, but they did not improve fit; these global variables were used as descriptors that illustrated the difference between profiles.

Characterization of each profile given the means on component variables are presented in Figs. 1 and 2, and the predominance of component variables can be seen in the same profiles. Profiles for *Didactic Lecture* and *Group Work* are almost exclusively dominated by their eponymous activities, while the profile for *Student Presentation and Review* shows higher student presentation, but also has lecture occurring later in the class (if at all), and higher proportions of teacher review, students writing on the board, and instructors asking questions calling for conceptual (versus informational) answers. The fourth profile for *Interactive Lecture and Review*, is the most mixed and does not have a dominant class activity. It is marked by more group work (usually more integrated

Table 4 Latent profile analysis fit statistics for 791 Classes

Number of profiles in solution	Log likelihood	AIC	BIC	Entropy	BLRT (value)	BLRT (p)
1	− 11,204.6	22,449.2	22,542.6	1	–	–
2	− 10,228.7	20,519.4	20,664.2	0.932	1951.8	0.009
3	− 9584.8	19,253.6	19,449.8	0.930	1287.7	0.009
4	− 9253.7	18,613.3	18,860.8	0.929	662.2	0.009

Table 5 Characterization of four latent profiles

Profile	Description	N
Didactic Lecture	Classes have an average of 80% lecture accompanied by the teacher writing on the board. Some question and answer occurred, although most questioning by teacher asked students to provide specific information and not conceptual reasoning.	354 (45%)
Student Presentation	While the prevalence of student presentation is somewhat low (30%), this profile is the only one in which this activity occurs other than very minimally. This profile also has more review of work than others (53%), instructors moderating discussions (38%) and students writing on the board (16%). We considered this to be the most interactive profile.	113 (14%)
Interactive Lecture and Review	This profile is the most mixed in terms of activities. It has more review by the instructor, more lecture, and more group work integrated into classes, versus group work that takes up the entire class time.	206 (26%)
Group Work	Classes were generally devoted to students working in groups an average of 75% of class time. In most group sessions, the teacher interacts with students. We saw very little presence of other activities in this profile with some occurrence of teacher review (8%) and lecture (8%).	117 (15%)

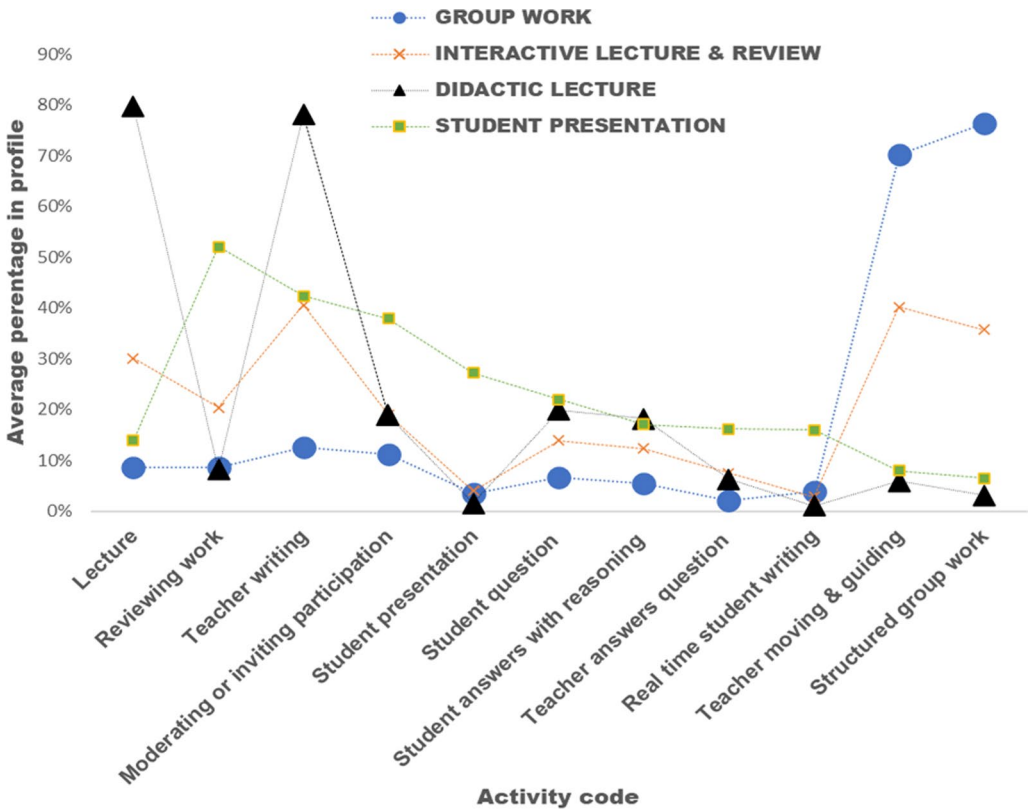

Fig. 1 Average proportions of observational activity codes for four profiles

into class time), some lecture mixed with other activities, and more activities in balance.

Thus, to answer Research Question 1, we found that a coherent four-group solution was possible with our data. While any solution is both sample- and model-dependent, the groups appear to represent distinct styles of instruction.

Dependent measures derived from the TAMI-OP

While it is important to be able to create a taxonomy of class types, we also wanted to know if we could characterize and assess change in teaching style using a single dependent measure.

The first outcome measure we attempted was based on factor analysis, an approach used in analyzing data from

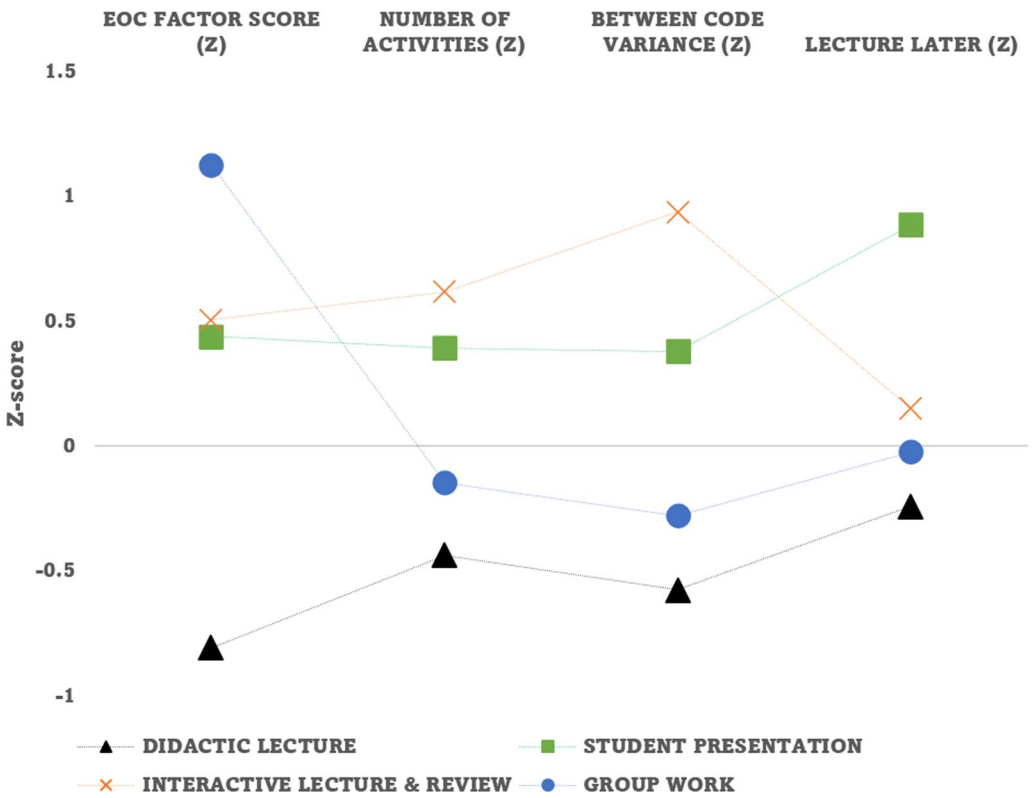

Fig. 2 Profiles of four LPA groups on global measures

the MCOP[2] observational protocol (Gleason et al., 2017). In attempting to conduct factor analysis on the TAMI-OP activity codes, we encountered problems with the solution for the factor variables, indicating that they should not be used in analysis. Most problems were related to the underlying correlations between variables; some pairs of variables lacked independence (e.g., were mutually exclusive), or had highly skewed and non-normal distributions, making it difficult to have confidence in the solution. However, the rotated solution did produce two factors explaining 54.1% of variance, and the solution was mostly orthogonal. The first scale described a continuum with the Lecture activity code loading highly at one end and the code for Reviewing Work at the other. The second spanned structured group work at one end of the scale to lecture at the other. Both scales showed very low internal reliability at $\alpha = 0.39$ and $\alpha = 0.25$, making their use inadvisable.

The second outcome measure attempted was the simple proportion of non-didactic lecture classes by a teacher as a dependent variable: *Proportion Non-Didactic Lecture* (PND). For example, the observation data set for a particular teacher may have six out of eight classes that fit the profile for the *Didactic Lecture* profile and two that do not, resulting in a proportion of non-didactic classes of PND = 0.25. If after an intervention, this proportion increases to 0.75, this could be considered a meaningful change in teaching practice.

The assumptions for the PND measure are that (1) enough classes are observed for each teacher to create a reliable measure and proportion (Brennan, 2001; Weston et al., 2021), and (2) profiles derived from LPA reflect both traditional and reform teaching practices. In addition, any pre/post or group comparisons based on LPA results must be created at the same time with the same sample and models.

To answer Research Question 2, we found that it was possible to derive a measure representing the proportion of non-didactic lecture classes for each instructor. Attempts to create factor variables with our data did not succeed because of the low reliabilities of resulting composite variables.

Validity of the RTOP

A number of studies cited above use the RTOP as a dependent measure to assess the effects of interventions such as workshops or other professional development. The unitary RTOP scale is continuous but is also meant to be criterion-referenced in that it classifies different

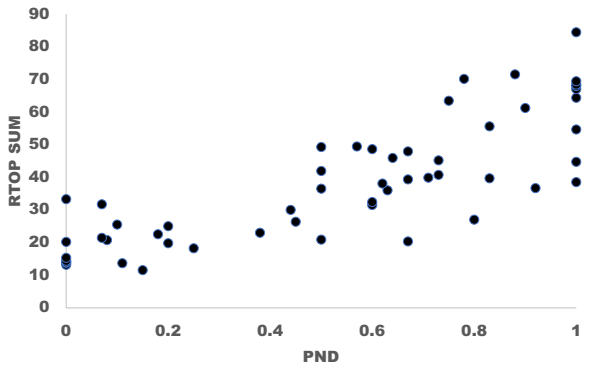

Fig. 3 Scatter of PND and RTOP-Sum. Correlation of $r = 0.68$ for $n = 54$ instructors

Table 6 Correlation matrix for RTOP sub-scale variables

	LDI	PK	PRK	CI
Lesson Design and Implementation (LDI)				
Propositional Knowledge (PK)	0.610**			
Procedural Knowledge (PRK)	0.921**	0.702**		
Communicative Interactions (CI)	0.936**	0.633**	0.916**	
Student Teacher Relationships	0.926**	0.645**	0.917**	0.952**

All correlations have $n = 484$, all correlations statistically significant at $p < 0.01$**

scale points as reflecting more or less interactive instruction (Piburn et al., 2000). The correlation between the PND and RTOP measures in our full dataset was high at $r = 0.68$, suggesting that the two measures both capture the same underlying construct. Figure 3 shows the scatter of PND and the RTOP-Sum.

The RTOP contains sub-scales for design, concepts, procedures, communication, and student participation. While these sub-scales, with the exception of the propositional scale, did not form separate factors in the original RTOP validity study (Sawada et al., 2003), we wanted to know if the same factor structure was also evident in our data.

Correlations between the summed sub-scale composites were calculated and plotted (see Table 6). These showed that the sub-scale variables were collinear with all correlations higher than $r = 0.6$, with six of ten correlations greater than $r = 0.9$. These latter correlations indicated the variables are essentially identical when measurement error is taken into account.

The Exploratory Factor Analysis (EFA) also strongly indicated that RTOP sub-scales in our data were measuring the same construct with a unified scale. The results of the factor analysis using Maximum Likelihood Estimation, an Oblique Rotation, and Parallel Analysis for retention of factors, found one dominant factor that encompassed all but one item with an Eigenvalue of 16.2. A second weaker factor with an Eigenvalue of 1.48 had six items loading with values greater than 0.3 on the factor; however, this factor did not adhere to any single sub-scale and was not orthogonal to the dominant factor. The secondary factor was different from the primary factor in that two items related to student interaction gave negative factor loadings, suggesting that this factor represented classes with little or no student-to-student discussion. The correlation between factor scores from the primary general factor and the RTOP-Sum was r = 0.99, indicating that the two measures were essentially identical. Overall, the results of this factor analysis show that one generalized factor accounts for most of the variance among the items, the secondary factor is mostly redundant to the first, and that RTOP subscales do not form separate and independent variables in our data. Table 7 presents the Exploratory Factor Analysis.

We also conducted a Confirmatory Factor Analysis (CFA) using the five proposed RTOP subscales. CFA is used to test and confirm the factor structure found in the EFA. The factor solution showed a poor model fit with $\chi^2 = 1922$, $df = 265$, $\chi^2/df = 7.2$, $RMSEA = 0.114$, $PCFI = 0.78$. The model fit statistics indicated that subscales did not form separate factors, and that the correlations among factor variables were collinear, confirming the same result found with the EFA. However, the unitary scale formed from the sum of the item scores (a measure used in many studies) was highly reliable at $\alpha = 0.97$ forming a robust composite scale.

In sum, the results for Research Question 3 support the use of the RTOP-Sum as a unitary scale and dependent measure. However, the RTOP does not form reliable subscales in our data.

The RTOP-Sum and the PND use in statistical tests

The PND and RTOP-Sum both form reliable measures, but it is unclear if these measures can detect pre/post or group differences when used in common statistical tests. We first examined the PND and the RTOP-Sum in our full dataset to assess univariate distributions (see Fig. 4). The RTOP-Sum, while skewed, resembled a normal distribution except for a bimodal bump at the high end of the scale. In contrast, the PND distribution was rectangular, with a high number of instructors (20 out of 88) who taught all non-didactic courses (PND = 1). The high number of "1" values potentially created a ceiling effect for pre/post comparisons. For our pre/post data (a subset of the larger sample), means for the distributions were 38 (RTOP-Sum) and 0.57 (PND), both comfortably near the middle of each scale. Relative to the scale (100 and 1), variability was greater in the PND compared to the RTOP-Sum (38% v. 19%), again the result of the high

Table 7 Exploratory factor analysis of RTOP measures ($n = 484$) (factor loadings >|0.3| included)

Sub-scale	Item	FL(1)	FL(2)
Lesson Design & Implementation	The instructional strategies and activities respected students' prior knowledge and the preconceptions inherent therein	0.80	
	The lesson was designed to engage students as members of a learning community	0.90	-0.33
	In this lesson, student exploration preceded formal presentation	0.84	
	This lesson encouraged students to seek and value alternative modes of investigation or of problem solving	0.84	
	The focus and direction of the lesson was often determined by ideas originating with students	0.83	
Propositional Knowledge	The lesson involved fundamental concepts of the subject	0.51	0.30
	The lesson promoted strongly coherent conceptual understanding	0.69	0.39
	The teacher had a solid grasp of the subject matter content inherent in the lesson	0.56	0.35
	Elements of abstraction (i.e., symbolic representations, theory building) were encouraged when it was important to do so	0.70	0.37
	Connections with other content disciplines and/or real-world phenomena were explored and valued	–	
Procedural Knowledge	Students used a variety of means (models, drawings, graphs, symbols, concrete materials, manipulatives, etc.) to represent phenomena	0.81	
	Students made predictions, estimations and/or hypotheses and devised means for testing them	0.86	
	Students were actively engaged in thought-provoking activity that often involved the critical assessment of procedures	0.90	
	Students were reflective about their learning	0.82	
	Intellectual rigor, constructive criticism, and the challenging of ideas were valued	0.89	
Communicative Interactions	Students were involved in the communication of their ideas to others using a variety of means and media	0.91	
	The teacher's questions triggered divergent modes of thinking	0.83	
	There was a high proportion of student talk and a significant amount of it occurred between and among students	0.87	-0.40
	Student questions and comments often determined the focus and direction of classroom discourse	0.82	
	There was a climate of respect for what others had to say	0.89	
Student Teacher Relationships	Active participation of students was encouraged and valued	0.92	
	Students were encouraged to generate conjectures, alternative solution strategies, and/or different ways of interpreting evidence	0.76	0.31
	In general, the teacher was patient with students	0.90	
	The teacher acted as a resource person, working to support and enhance student investigations	0.89	
	The metaphor "teacher as listener" was very characteristic of this classroom	0.91	

number of 1's in the PND scale. Greater variability can also translate to less statistical power when conducting hypothesis tests. Figure 4 presents the univariate distributions of both RTOP-Sum and PND.

For our sample data of 15 teachers and 278 classes, which include both pre and post observations for the same group of teachers, separated by a professional development intervention, we conducted a parametric Paired-Sample t test and a non-parametric Marginal Homogeneity test comparing pre and post values for the PND and RTOP-Sum. We also calculated effect sizes for pre/post gain. The results presented in Table 8 shows statistically significant results for change in both the RTOP and PND measures.

Overall, addressing Research Question 4, both measures detect significant differences in a pre/post comparison study. Using the same data, the RTOP-Sum has a bigger effect size and lower p value than the PND, indicating that a larger pre–post effect was found with the RTOP.

Descriptive qualities of the RTOP and PND measures

We examined how each measure allows extrapolation to score scales and descriptions of score domains. This is an important, but often overlooked, aspect of measurement that examines how a measure or measurement system describes its content or skill domain and allows those interpreting a study to learn what exactly changed between pre and post, or what is different between participating and comparison groups.

The extrapolation of the RTOP-Sum is mainly limited to the scale descriptions used by Sawada (2003) and others (Ebert-May et al., 2011). As noted, these seem to lack reference to observable teaching behaviors, but instead

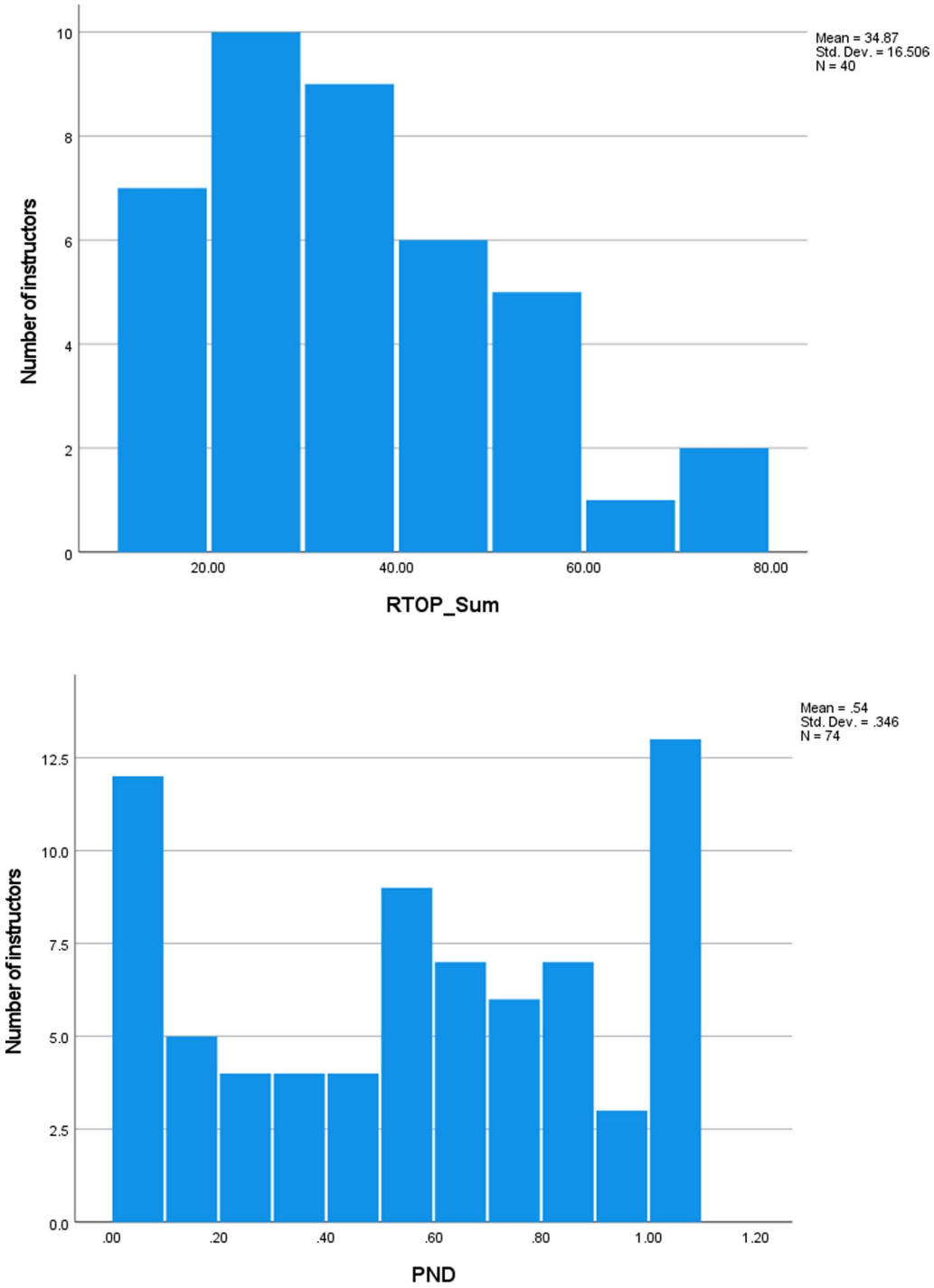

Fig. 4 Univariate Distributions of RTOP-Sum and PND Variables

provide more global assessments of instruction. One valuable descriptive outcome from this figure is the number of instructors who seemed to make categorical increases from pre to post. In our data set, eight of the fifteen instructors increased their scores to higher categories in this data. This could potentially provide valuable feedback to professional developers about the effects of their program, although, again, the exact nature of what occurred in the classroom would still be obscured by the language describing the categories.

Table 8 Test statistics for the RTOP-Sum and PND measures for pre/post comparison

Test	RTOP-Sum (Scale 0–100)	PND (Scale 0–1)
Paired *t* test	Mean difference = 17 SD = 14.5 Correlation pre/post = 0.63 Standard Error = 3.76 t = 4.56. df = 14, p < 0.001***	Mean difference = 0.22 SD = 0.27 Correlation pre/post = 0.58 Standard Error = 0.07 t = 3.1. df = 14, p = 0.004**
Related samples Wilcoxon Signed Rank Test	Test statistic = 119 N = 15 Standard Error = 17.6 t-statistic = 3.35 Asymptotic Sig. < 0.001***	Test statistic = 82 N = 15 Standard Error = 14.3 t-statistic = 2.5 Asymptotic Sig. = 0.01*
Effect size	Cohen's d = 1.17	Cohen's d = 0.81

$p < 0.05$*, $p < 0.01$ **, $p < 0.001$***. Both effect sizes are considered large

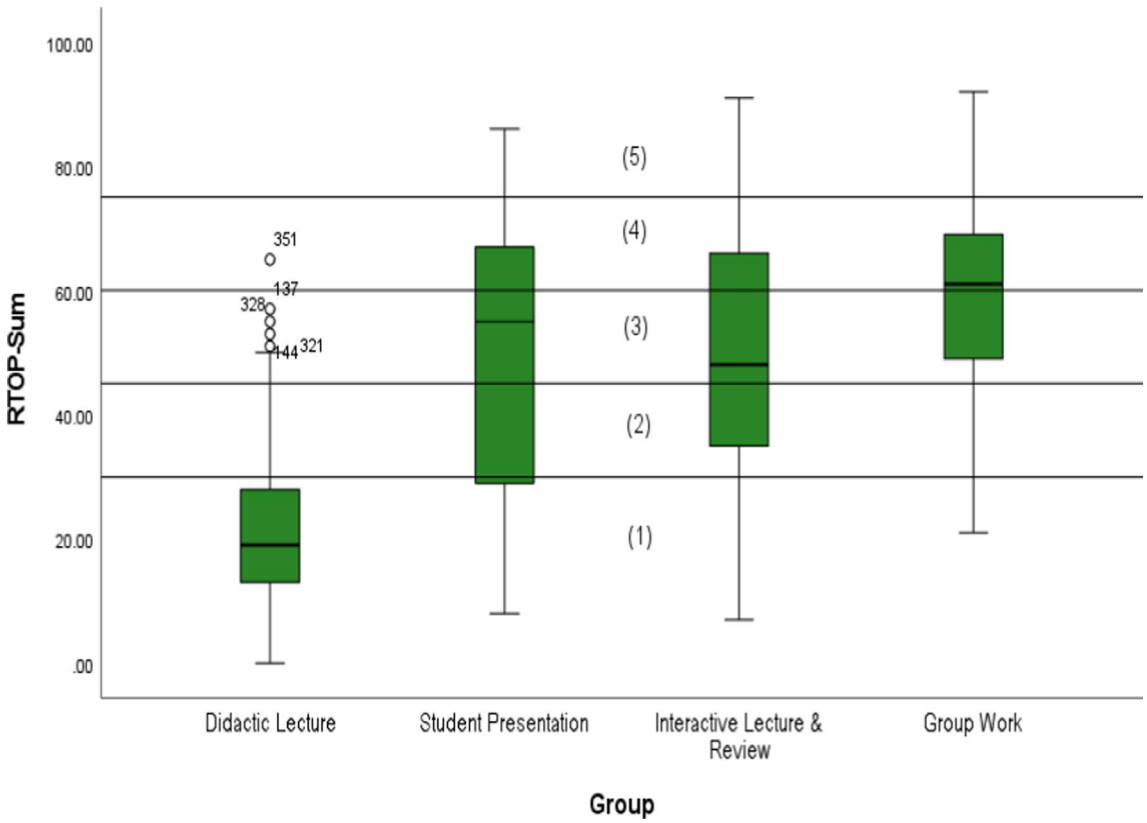

Fig. 5 LPA Group and Distribution of RTOP-Sum Score. Numbers in parentheses correspond to RTOP categories: (1) straight lecture, (2) lecture with some demonstration and minor student participation, (3) significant student engagement with some minds-on as well as hands-on involvement, (4) active student participation in the critique as well as carrying out of experiments, (5) active student involvement in open-ended inquiry, resulting in alternative hypotheses and critical reflection. Boxplot lines mark the mean RTOP score for each profile

We plotted the RTOP scores for each LPA group identified from TAMI-OP data. While the mean scores for each profile indicate that the average profile RTOP score for non-lecture groups become more interactive, comparing the two measures shows considerable overlap of RTOP-Sum scores for each group. In this comparison, the same score on the RTOP (in our data at least) often represent very different teaching styles in the LPA groups. Figure 5 presents the distribution of RTOP-Sum scores for each latent profile.

The descriptive utility of the PND is linked to its derivation from component Latent Profile Analysis groups. The separate activity codes and global variables used to form groups were graphed to learn which activities

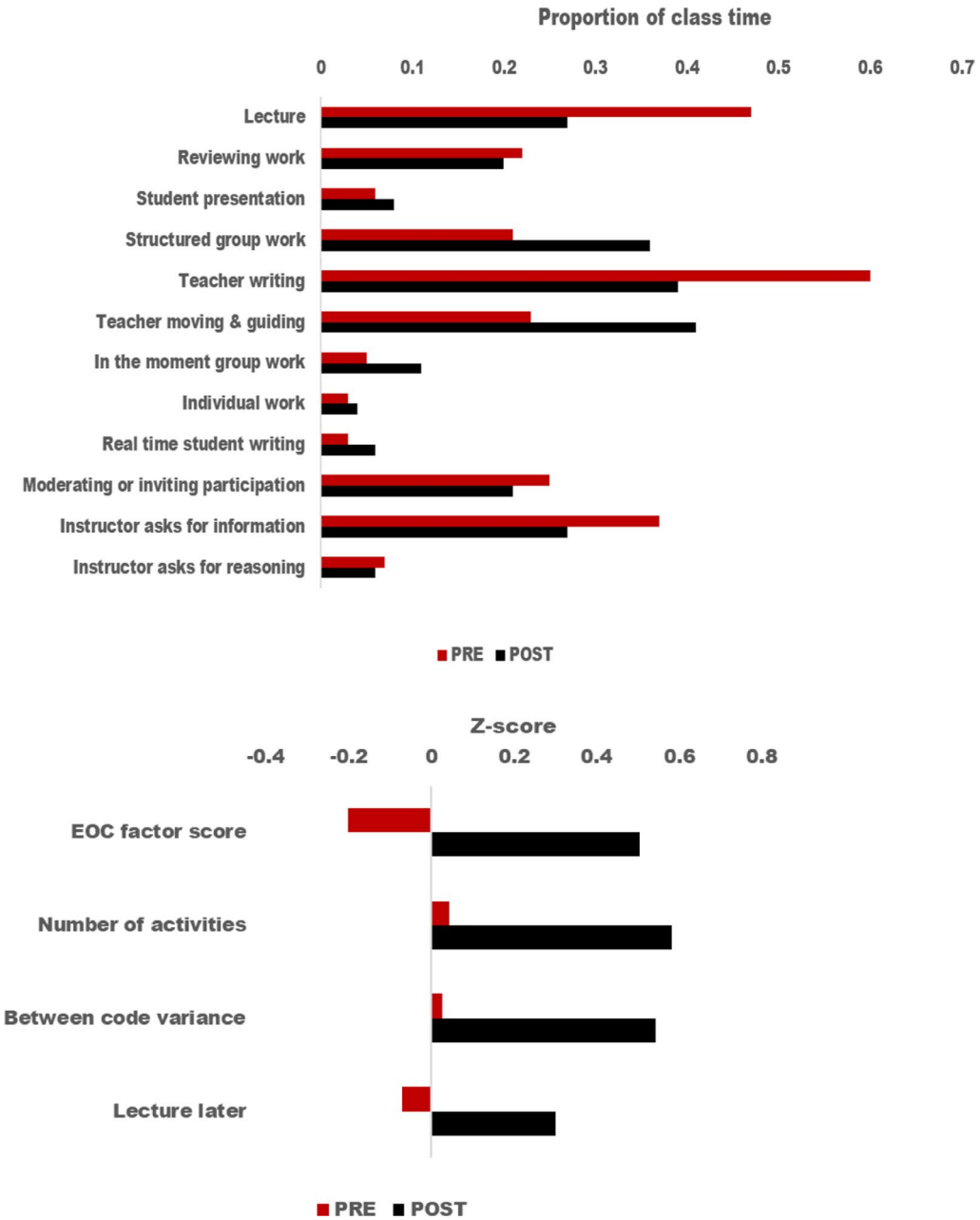

Fig. 6 Pre/post Change for Activity Codes and Global Variables

changed from pre to post. Most codes changed in ways consistent with the goals of the professional development in which they participated, with lecture and teacher writing decreasing and group work and student presentation increasing. The average number of activities and balance of activities also increased. Figure 6 presents pre/post change for activity codes and global variables.

While the PND can be used in statistical tests, finer grained movements between instructional categories can also be tracked pre/post and quantified. Shifts in proportions of classes across teachers' each grouping (e.g., from *Didactic Lecture* to *Group Work*) for our dataset are shown in Fig. 7. For these instructors, the proportion of lecture classes and student presentation decreased, while the proportion of interactive lecture

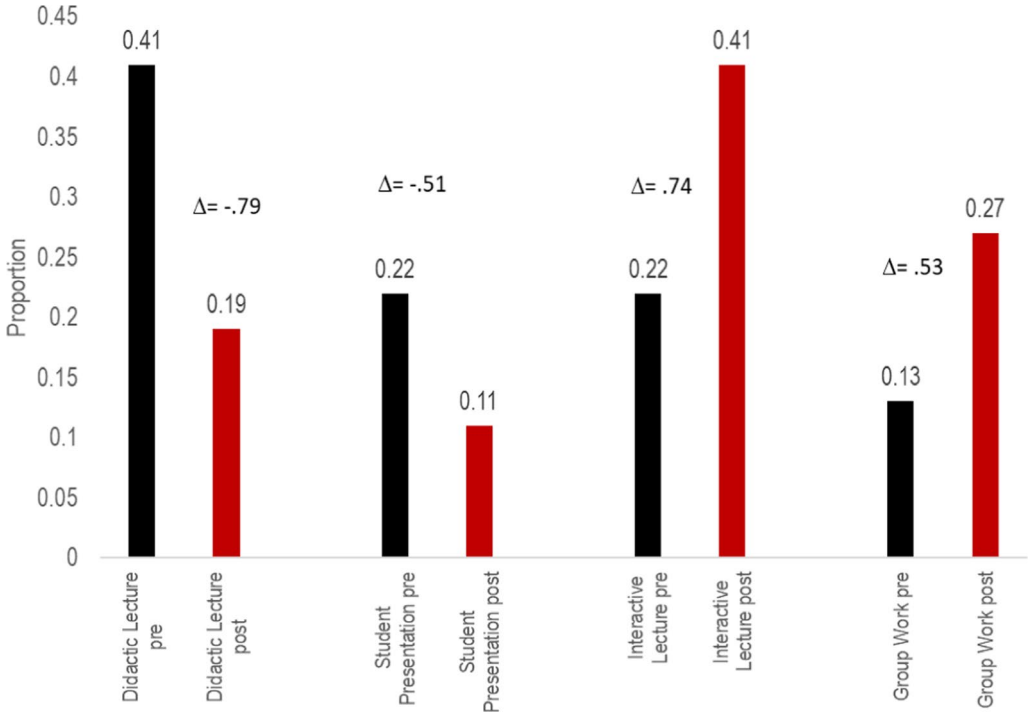

Fig. 7 Changes in Proportion of Latent Profiles from Pre to Post. Δ denotes effect size equal to the difference between pre post means divided by pooled standard deviation

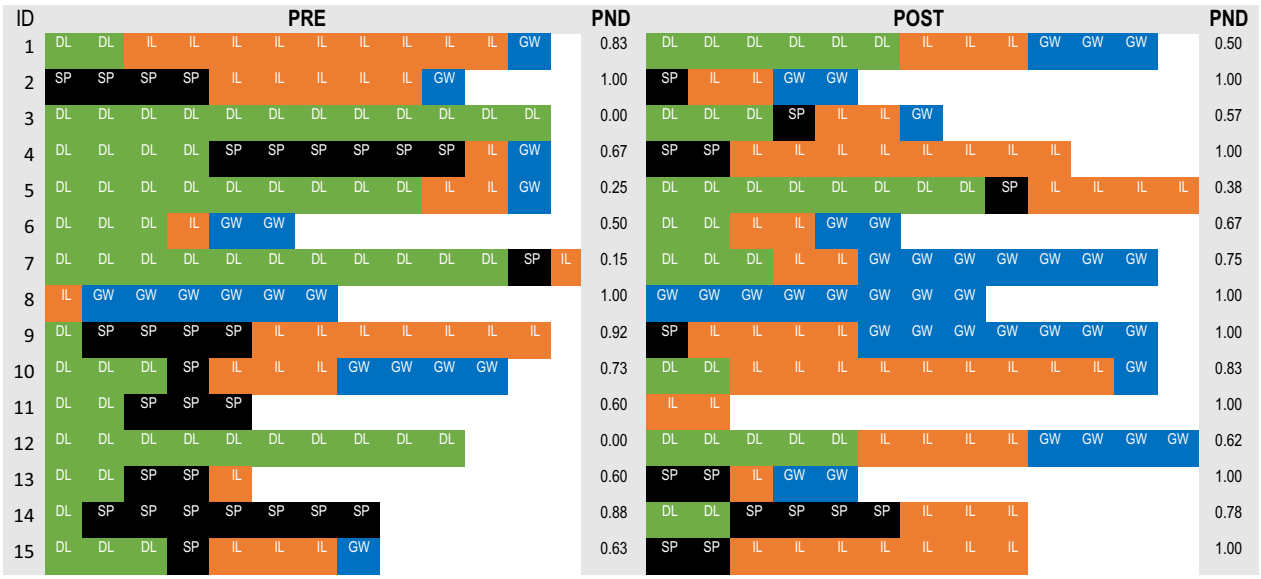

Fig. 8 Pre–Post Class Profiles for 15 Instructors. DL = Didactic-Lecture, SP = Student Presentation, IL = Interactive Lecture and Review, GW = Group Work

and group work increased, changes that can be quantified as effect sizes. The change in proportions of classes for each LPA grouping can also be seen in "mosaic" visualizations of the four LPA categories as they shift from pre to post. In Fig. 8, we can see the shift from Didactic Lecture (blue) to more varied forms of non-didactic teaching. Thus, in answer to Research Question 5, measures based on the segmented observation data offer richer ways to describe changes over time in instructional style.

Discussion

Profiles of classes created from Latent Profile Analysis provided four groups, which we labeled *Didactic Lecture*, *Interactive Lecture and Review*, *Student Presentation,* and *Group Work* (RQ1). The grouping method was reliable, and we believe these groups represent different underlying styles of teaching and learning present in our observations of 790 mathematics classrooms. In the *Didactic Lecture* group, instructors averaged 80% of their time lecturing, usually with little question and answer. This contrasted with the three non-lecture groups, where students participated in more interactive activities such as group work (usually working though problem sets), presenting problems on the board, or participating in more back-and-forth dialogue with the instructor during lecture and review. Instructors for classes in the three non-didactic lecture groups also engaged in more activities in their classrooms and tended to have more balance in time devoted to each activity.

The need for a usable dependent measure comes from the specific characteristics of segmented observational data, which are difficult to summarize in a way that adequately represents underlying constructs and lacks the usual statistical qualities of continuous data with highly skewed distributions (RQ2). Our first attempt at creating outcome scores with TAMI-OP data employed factor analysis to create continuous variables as was used in the MCOP^2 with holistic data (Gleason et al., 2017). Here, we determined that this approach was not advisable, because our data did not meet assumptions for factor analysis given highly skewed distributions within activity codes and the low frequencies of many observations (Schmidt, 2011). The resulting aggregate variables derived from the factor analysis had very low internal reliabilities, which would make many analyses flawed. It is possible that other research or evaluation projects using different observational protocols could overcome these challenges and use factor variables as continuous outcome measures. A resulting measure would then need to be extrapolated to describe the characteristics of score ranges.

From the LPA results we created a measure called the Proportion of Non-Didactic Lecture (PND) that represented the proportion of more interactive classes, contrasted to didactic lecture classes, for each instructor. The value of a measure lies in its ability to summarize data from multiple activity codes and other variables into one measure while avoiding the pitfalls of poor construct representation, strict reliance on non-parametric tests, and multiple comparisons found in many studies that use segmented data (Tomkin et al., 2019). We found that the PND measure had some shortcomings caused by its reliance on proportional frequency data. In our wider dataset the PND had a significant number of "1" values, which created the possibility of ceiling effects and lacked distributional normality. While statistical tests are robust to non-normality (Glass & Hopkins, 1996), comparisons made with small numbers like ours (i.e., the pre/post subset of 15 instructors) have less statistical power. In fact, the pre/post statistical comparison conducted with the measure showed less statistical power than did the comparison with the RTOP-Sum, but in our case provided similar statistical inferences as the RTOP about pre–post change.

We examined the RTOP-Sum observational measure to assess its validity and as a way of comparing outcomes with the PND (RQ3, RQ4). The unitary measure is a simple sum of item ratings and is used for many comparisons in studies about STEM education (see Emery et al., 2020). We found that the measure had very high internal reliability and worked well statistically as a continuous measure. It also is more statistically powerful than the PND, because it is a reliable continuous variable. The PND correlated highly with the RTOP-Sum, suggesting both measures tap into the same general construct of active learning.

However, we cannot make strong claims for the validity of the RTOP. Our findings do identify some concerns about the validity of the RTOP that supported the original findings of Sawada (2002). The subscales proposed by the designers of the survey do not significantly tap different underlying constructs, an important criterion for a usable dependent measure (Bell et al., 2012; Huppert, 2021). This was evidenced by very high correlations among composite sub-scale scores, a lack of simple representative structure in Exploratory Factor Analysis, and poor model fit for the proposed structure in Confirmatory Factor Analysis. This result has implications for those wishing to confidently use the RTOP in research studies, while the use of the summed total score would seem appropriate; in general it would seem unwise to make research claims based on the RTOP subscales.

It is our conclusion that it is not a good idea to exclusively use the RTOP observational protocol and its resulting measures to describe STEM teaching and teaching change. Problems with describing teaching are rooted in item wording, with items calling for expert judgment and inferences about the inner states of participants, all logical item writing errors (Fowler, 2009). Some of these challenges could possibly be overcome through training with a detailed rubric as is provided by the MCOP^2 (Gleason et al., 2017). While the RTOP-Sum as a dependent measure is responsive to changes in teaching, the categorical descriptions of teaching style linked to total score are poorly defined—especially in the middle of the scale, where many instructors' practices fall. We also saw that each LPA group

contained a wide range of RTOP scores, confirming the observations of Lund et al. (2015) that the same RTOP score may represent very different styles of teaching.

We believe that the PND derived from the TAMI-OP protocol provides much more detailed and actionable information to stakeholders such as those who implement and participate in professional development workshops (RQ5). This is seen in the rich descriptions of teaching style that emerge from LPA group definitions and the ability to represent changes in the overall representation of teaching styles with visualizations from pre to post. While the PND measure may suffer from less statistical power and possible ceiling effects, its derivation from segmented data provides a richer description of the activities of instructors and students than is possible with holistic protocols.

An obvious question would be why not just use both holistic and segmented measures? This is the ideal solution we used that takes advantage of the strengths of both measures. However, implementing observations is logistically daunting when compared to surveys or test data (Hill et al., 2012). Conducting a single classroom observation already requires substantial planning, time, resources, and coordination, all of which cost time and money. If not gathered by remote video, observers must get to a site, observe, and then enter data (Cash et al., 2012). Because of these logistical challenges, and the cognitive demands on the observer, it may not be possible to conduct both types of observations in real time. In these cases, it may be preferable to use a segmented observational protocol given its superior ability to describe teaching and teaching change. Alternately, video recording can allow for more observations and less time traveling, and more readily enables coding with multiple protocols. This is nonetheless expensive, as raters must still watch and code the videos (Madigan et al., 2017), and more time is required to apply multiple protocols.

The finding that the PND is usable statistically has implications for the role of observational measures in research and evaluation about STEM education. Use of a PND measure works best in providing diagnostic information for those conducting professional development (or some other intervention) by providing a map of instructional approaches before and after a program, or in contrast to a comparison group. This allows those assessing these interventions to learn not only if the intervention had an effect, but what changes truly occurred when instructors changed their teaching style. Conceivably, individual instructors could also benefit from the same diagnostic information. The PND can also be used to support claims about the efficacy of professional development in the literature, so that those implementing STEM professional development or other curricula can confidently assess their efficacy.

Limitations

There are several critical caveats to the use of a measure based on LPA or any other clustering technique. The final categorization of classes (or people in other analyses) is dependent on both the sample used and the variables included in the model. The ultimate category, where classes end up can vary depending on the characteristics of the initial pool of classes and the specification of the model (Williams & Kibowski, 2016). Any project also needs a relatively large pool of classes to make cluster or profile methods viable. In their overview of LPA studies, Spurk and coauthors (2020) found a median sample size near 500; in our study we were fortunate to have a collection of nearly 800 classes. It is possible to leverage the earlier work of others; for example, those using the COPUS can take advantage of the COPUS Analyzer (Harshman & Stains, 2020), an online method for profiling observational data. While it may seem obvious, pre and post or participant/comparison groupings (for any clustering technique) must be made at the same time and from the same model. In addition, the creation of an LPA model should be done independently from, and before any type of statistical comparison is made. Shopping for the model that creates the largest effect for a comparison would constitute a breach of research ethics.

Deriving a proportional measure from segmented observational data is also limited by several important assumptions. First, there must be enough classes observed for each teacher to form a reliable measure, a number that is usually higher than is found in most research studies (Weston et al., 2021), and observing enough classes for a reliable measure is resource intensive. Second, profiling or clustering solutions must conform to a continuum from didactic to interactive instruction. This seems to be a common finding for profile studies, where a large proportion of classes are didactic lecture (Denaro et al., 2021; Lund et al., 2015; Stains et al., 2018). Closer examination of the characteristics of the clusters in these studies, it becomes clear that the didactic lecture style described in our study is common. In the study by Stains et al. (2018), the average amount of lecturing was 80% with minimal question and answer, identical to our finding. Likewise, Lund et al. (2015) found three clusters with 87% to 94% time spent lecturing. Both studies also found clusters similar to the more interactive groups we found; both described distinct clusters emphasizing group work, and Lund et al. found a cluster defined by student presentation. Disciplinary differences in preferred styles of active learning may play a role; for example, student presentation is well-developed

as an active learning approach in mathematics and may be more common in math classes than in the sciences (Laursen et al., 2019). The main limiting factor for some studies may be the small number of truly interactive classes observed; in Stains et al. (2018) approximately 25% of classes were student-centered, although mathematics classes had the highest percentage of these courses (~ 35%).

A simple proportion such as the PND also inevitably glosses over distinctions along the continuum of interactive teaching. In the data of Stains et al. (2018), classes belong to three groupings that exist on a continuum representing didactic, interactive, and student-centered instruction. Imposing a simple proportion on this type of grouping would oversimplify this scale. However, it would still be possible to test proportions of didactic and interactive classes to student-centered as a separate comparison while taking advantage of the descriptive utility of describing the change in teaching style afforded by the profiling method. In contrast, Lund et al. (2015) found 10 groupings that were then simplified to four instructional styles. As with our data, three of the four clusters represented somewhat equal but different active instructional approaches including the clusters they termed Socratic, Peer Instruction, and Collaborative Learning. Again, any researcher finding similar groupings in a pre/post comparison would be able to map any changes to overall instructional style that are oversimplified in the proportional measure. The method described in our paper also glosses over differentiations of teaching quality within clusters. In Fig. 5, we see a large variation in RTOP scores within each cluster group, suggesting that instruction can differ substantially in the levels of inquiry tapped by the RTOP. However, it should also be noted that the majority of cases in the *Didactic Lecture* profile have RTOP scores lower than the main distribution of cases in the other groups. Further development of observational rubrics could add more refinement within each profile to better describe variations in teaching and learning. Ultimately, more reliable statistical characterizations of teaching will make possible future studies that seek to link observed teacher practices—and changes therein—to measures of students' learning experiences and outcomes.

Conclusion

Teaching observations can be used in multiple ways to describe and assess instruction. We addressed the challenge of measuring instructional change with segmented observational protocols, data that do not lend themselves easily to statistical comparisons. We proposed a simple proportional measure (PND) based on latent profiles of classroom teaching. The measure aggregated the proportions of interactive to lecture classes for each instructor. The PND worked in simple hypothesis tests but lacked some statistical power due to possible scaler ceiling effects. However, the PND provided effective descriptions of changes in instructional approaches from pre to post among participants in teaching-focused professional development.

In tandem with examining the proportional measure, we also examined the RTOP-Sum, an existing outcome measure used in comparison studies. The measure is based on the aggregated items in a holistic observational protocol. As an aggregate measure it is highly reliable, correlated highly with the PND, and had more statistical power than the PND. However, the measure suffered from poor scale descriptions and did not seem to provide the thick descriptions afforded by the measure based on latent profiles of instruction.

As mentioned, the main limiting factors in conducting observational research are logistical. However, the wide availability of cell phone and other affordable video cameras and increasingly large data storage capacities have made it easier for researchers to gather video of classroom teaching. This expanded access, coupled with the development of machine learning as an aid in coding qualitative data (Chen et al., 2018), could make conducting observational studies much easier in the future. Having analytical and measurement systems in place provides the means for adequately detecting and describing teaching change.

Acknowledgements

We would like to acknowledge the important contributions of Devan Daly and Kyra Gallion, both staff members of Ethnography and Evaluation Research at the University of Colorado. We also acknowledge the work of Holly Devaul in proofreading the manuscript, and Tim Archie for his contributions to discussions about research design and implementation.

Author contributions

TJW is a co-PI on two of the projects listed. He participated in the design of data collection instruments, analyzed data, and led the writing of the manuscript. SLL is PI of the funded projects in the study, participated in the development of instruments, helped with data analysis and write-up of the article. CNH participated in the design of data collection instruments, collected and coded data, participated in analysis of data and contributed to writing the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by the National Science Foundation under grant DUE grant numbers 0920126, 1225833, 1225820, 1225658, 1525058, and 1525077.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to privacy protection of the participants but are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details
[1]National Center for Women and Information Technology (NCWIT), University of Colorado, Boulder, 2343 Sandpiper Dr., Lafayette, CO 303-501-4636, USA. [2]Ethnography and Evaluation Research, University of Colorado, Boulder, CO 80309, USA.

Received: 11 October 2022 Accepted: 10 March 2023
Published online: 29 March 2023

References

American Association for the Advancement of Science (AAAS) (2013). *Describing and Measuring Undergraduate STEM Teaching Practices. A Report from a National Meeting on the Measurement of Undergraduate Science, AAAS: Washington, DC.*

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics, 3*, 103–107.

Adamson, S. L., Banks, D., Burtch, M., Cox, F., III., Judson, E., Turley, J. B., & Lawson, A. E. (2003). Reformed undergraduate instruction and its subsequent impact on secondary school teaching practice and student achievement. *Journal of Research in Science Teaching, 40*(10), 939–995.

Bain, K., Bender, L., Bergeron, P., Caballero, M. D., Carmel, J. H., Duffy, E. M., & Cooper, M. M. (2020). Characterizing college science instruction: The Three-Dimensional Learning Observation Protocol. *PLoS ONE, 15*(6), e0234640.

Beach, A. L., Henderson, C., & Famiano, M. (2008). 13: Co-teaching as a faculty development model. *To Improve the Academy, 26*(1), 199–216.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87.

Borda, E., Schumacher, E., Hanley, D., Geary, E., Warren, S., Ipsen, C., & Stredicke, L. (2020). Initial implementation of active learning strategies in large, lecture STEM courses: Lessons learned from a multi-institutional, interdisciplinary STEM faculty development program. *International Journal of STEM Education, 7*(1), 1–18.

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

Budd, D. A., Van der Hoeven Kraft, K. J., McConnell, D. A., & Vislova, T. (2013). Characterizing teaching in introductory geology courses: Measuring classroom practices. *Journal of Geoscience Education, 61*(4), 461–475.

Campbell, T., Der, J. P., Wolf, P. G., Packenham, E., & Abd-Hamid, N. H. (2012). Scientific Inquiry in the genetics laboratory: Biologists and university science teacher educators collaborating to increase engagement in science processes. *Journal of College Science Teaching, 41*(3), 74–81.

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529–542.

Chen, N. C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TIIS), 8*(2), 1–20.

Denaro, K., Sato, B., Harlow, A., Aebersold, A., & Verma, M. (2021). Comparison of cluster analysis methodologies for characterization of classroom observation protocol for undergraduate STEM (COPUS) data. *CBE Life Sciences Education, 20*(1), 3.

Dwivedi, A. K., Mallawaarachchi, I., & Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine, 36*(14), 2187–2205.

Ebert-May, D., Dertling, J., Momsen, J., Long, T., & Jardeleza, S. (2011). What we say is not what we do: Effective evaluation of faculty professional development programs. *BioScience, 61*, 550–558.

Egert, F., Fukkink, R. G., & Eckhardt, A. G. (2018). Impact of in-service professional development programs for early childhood teachers on quality ratings and child outcomes: A meta-analysis. *Review of Educational Research, 88*(3), 401–433.

Emery, N. C., Maher, J. M., & Ebert-May, D. (2020). Early-career faculty practice learner-centered teaching up to 9 years after postdoctoral professional development. *Science Advances, 6*(25), eaba2091.

Esparza, D., Wagler, A. E., & Olimpo, J. T. (2020). Characterization of instructor and student behaviors in CURE and Non-CURE learning environments: impacts on student motivation, science identity development, and perceptions of the laboratory experience. *CBE Life Sciences Education, 19*(1), 10.

Ferguson, S. L., Moore, E. W., & Hull, D. M. (2020). Finding latent groups in observed data: A primer on latent profile analysis in Mplus for applied researchers. *International Journal of Behavioral Development, 44*(5), 458–468.

Fowler, F. J., Jr., & Cosenza, C. (2009). Design and evaluation of survey questions. *The SAGE Handbook of Applied Social Research Methods, 2*, 375–412.

Glass, G., & Hopkins, K. (1996). *Statistical methods in education and psychology*. Pearson College Division.

Gleason, J., Livers, S., & Zelkowski, J. (2017). Mathematics classroom observation protocol for practices (MCOP2): A validation study. *Investigations in Mathematics Learning, 9*(3), 111–129.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967.

Harshman, J., & Stains, M. COPUS Analyzer COPUS Profiles. http://www.copusprofiles.org/ (Accessed Feb 10, 2022).

Hayward, C., Weston, T., & Laursen, S. L. (2018). First results from a validation study of TAMI: Toolkit for Assessing Mathematics Instruction. In *21st Annual Conference on Research in Undergraduate Mathematics Education* (pp. 727–735).

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough. *Educational Researcher, 41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Hora, M. T., & Ferrare, J. J. (2014). Remeasuring postsecondary teaching: How singular categories of instruction obscure the multiple dimensions of classroom practice. *Journal of College Science Teaching, 43*(3), 36–41.

Hora, M. T., & Ferrare, J. J. (2013). Instructional systems of practice: A multidimensional analysis of math and science undergraduate course planning and classroom teaching. *Journal of the Learning Sciences, 22*(2), 212–257.

Hora, M. T., Oleson, A., & Ferrare, J. J. (2013). *Teaching dimensions observation protocol (TDOP) user's manual*. Madison: Wisconsin Center for Education Research.

Huppert, J. D., Walther, M. R., Hajcak, G., Yadin, E., Foa, E. B., Simpson, H. B., & Liebowitz, M. R. (2007). The OCI-R: Validation of the subscales in a clinical sample. *Journal of Anxiety Disorders, 21*(3), 394–406.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Praeger.

Laursen, S., Andrews, T., Stains, M., Finelli, C. J., Borrego, M., McConnell, D., Johnson, E., Foote, K., Ruedi, B., & Malcom, S. (2019). *Levers for change: An assessment of progress on changing STEM instruction*. American Association for the Advancement of Science.

Lane, E., & Harris, S. (2015). Research and teaching: A new tool for measuring student behavioral engagement in large university classes. *Journal of College Science Teaching*. https://doi.org/10.2505/4/jcst15_044_06_83

Lund, T. J., Pilarz, M., Velasco, J. B., Chakraverty, D., Rosploch, K., Undersander, M., & Stains, M. (2015). The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. *CBE-Life Sciences Education, 14*(2), 18.

Madigan, R., Lee, Y. M., & Merat, N. (2021). Validating a methodology for understanding pedestrian–vehicle interactions: A comparison of video and field observations. *Transportation Research Part F: Traffic Psychology and Behaviour, 81*, 101–114.

Manduca, C. A., Iverson, E. R., Luenberg, M., Macdonald, R. H., McConnell, D. A., Mogk, D. W., & Tewksbury, B. J. (2017). Improving undergraduate STEM education: The efficacy of discipline-based professional development. *Science Advances, 3*(2), e1600193.

Matz, R. L., Fata-Hartley, C. L., Posey, L. A., Laverty, J. T., Underwood, S. M., Carmel, J. H., & Cooper, M. M. (2018). Evaluating the extent of a large-scale transformation in gateway science courses. *Science Advances, 4*(10), e0554.

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*(2), 331.

National Council of Teachers of Mathematics. (2022). Standards and Positions. http://caepnet.org/accreditation/caep-accreditation/spa-standards-and-report-forms/nctm
National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. Berlin: National Academies Press.
Nunnally, J., & Bernstein, I., 3rd. (1994). *Psychometric theory* (3rd ed.). New York.
Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). Reformed teaching observation protocol (RTOP) reference manual. *Tempe, Arizona: Arizona Collaborative for Excellence in the Preparation of Teachers*.
Sawada, D., Eslamieh, C., & Wyckoff, S. (2003). *Reformed teacher education in science and mathematics: An evaluation of the Arizona Collaborative for Excellence in the Preparation of Teachers ACEPT*. Document Production Services.
Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253.
Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304–321.
Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: a review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior, 23*, 103445.
Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., & Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science, 359*(6383), 1468–1470.
Stains, M., Pilarz, M., & Chakraverty, D. (2015). Short and long-term impacts of the Cottrell Scholars Collaborative New Faculty Workshop. *Journal of Chemical Education, 92*(9), 1466–1476.
Smith, M., Jones, H., Gilbert, S., & Wieman, C. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE Life Sciences Education, 12*(4), 618–627.
Tomkin, J. H., Beilstein, S. O., Morphew, J. W., & Herman, G. L. (2019). Evidence that communities of practice are associated with active learning in large STEM lectures. *International Journal of STEM Education, 6*(1), 1–15.
Weston, T. J., Hayward, C. N., & Laursen, S. L. (2021). When seeing is believing: Generalizability and decision studies for observational data in evaluation and research on teaching. *American Journal of Evaluation, 42*(3), 377–398.
Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations*. Brookings Institute.
Williams, G. A., & Kibowski, F. (2016). Latent class analysis and latent profile analysis. *Handbook of methodological approaches to community-based research: Qualitative, quantitative, and mixed methods*, pp. 143–151.

Publisher's Note