Using Latent Profile Analysis to Assess Teaching Change

Timothy J.Weston
University of Colorado

Sandra L. Laursen
University of Colorado

Charles N. Hayward
University of Colorado

Teaching observations can be used in multiple ways to describe and assess instruction. We addressed the challenge of measuring instructional change with observational protocols, data that often do not lend themselves easily to statistical comparisons. We first grouped 790 mathematics classes using Latent Profile Analysis and found four reliable categories of classes. Based on the grouping we proposed a proportional measure called Proportion Non-Didactic Lecture (PND). The measure is the proportion 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. The measure correlated highly with a dependent measure derived from the Reformed Teaching Observation Protocol (RTOP), a holistic observational measure. The PND also provided effective descriptions and visualizations of instructional approaches and how these changed from pre to post.

Keywords: Structured Observations, Undergraduate STEM Teaching,

Introduction

Numerous studies show that active, engaging, and collaborative classrooms help students learn and persist in college, but adoption of new teaching practices has been slow (American Association for the Advancement of Science, 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 one tool 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). But learning whether or not these programs change teaching practices is challenging because typical means of measurement, 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).

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).

Shortcomings of Segmented Observational Protocols as Dependent Measures

Segmented observational protocols such as the COPUS and TDOP are employed in comparative research designs but pose measurement challenges. Typically, these instruments code each 2-minute segment of class time for instructor and student behaviors such as lecture or

group work. Difficulties arise in using segmented observational protocols in research studies for several reasons. First, 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). In effect, this can oversimplify what is occurring the classroom. Data drawn from a segmented protocol 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., 2019). 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 the high number of codes generated by segmented protocols compared to a holistic protocol's single aggregate 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 doesn't exist) increases substantially (Abdi, 2007), which can lead to false claims about the efficacy of an intervention.

Shortcomings of Holistic Observational Protocols as Dependent Measures

Many studies that employ observational data to assess change use the Research Teaching Observation Protocol (RTOP), a holistic observational measure (Sawada et al., 2002). Holistic instruments ask observers to rate elements of a class such as "The lesson promoted strongly coherent conceptual understanding." These types of instruments often ask for more expert judgments of teaching quality versus observations of behaviors (Hora& Ferrare, 2013). While the measures derived from the RTOP have high internal reliability and some criterion validity, the measure seemed 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). Those using the measure also seem limited in their ability to extrapolate from scores to more concrete descriptions of teaching. This is partly caused by the somewhat vague wording of some score range categories that are presented in early RTOP validity documents (Sawada et al., 2003) and studies using the RTOP for outcome comparisons (Ebert-May et al., 2011). An example would be the score range category "46-60 Significant student engagement with some minds-on as well as hands-on involvement," which provides little guidance on what instructors and students are doing in the classroom. 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 teaching descriptions varied even more widely from study to study.

Rationale for Study & Research Questions

In the current study, we consider two 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. 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. Research questions include:

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) How do the RTOP aggregate dependent measure and the segmented TAMI-OP dependent measure function with statistical tests?
4) How can the segmented TAMI-OP dependent measure be extrapolated to provide descriptions of teaching and teaching change?

Methods

Instruments

We developed segmented observational protocol called the Toolkit for Assessing Mathematics Instruction-Observation Protocol (TAMI-OP) (Hayward et al., 2018). At two-minute 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*, including codes for Lecture, Student Questions, Group Work and Student Presentation among other activities. We also completed the RTOP for a subset of 484 of the same classes observed with the TAMI-OP. Both the TAMI-OP and RTOP had adequate interrater reliability, generalizability and internal reliability.

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. 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. All data were collected with human subjects approval.

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.

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). 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 a Latent Profile Analysis of the 790 classes in our database. The component variables for analysis all used class-level proportions of activity codes. While these variables are continuous, most did not form normal univariate distributions. We used a Maximum Likelihood (ME) estimation, 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.

Results

We found four reliable profiles that characterized 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. 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*. Figure 1 presents the individual averages for each observation code for each profile.

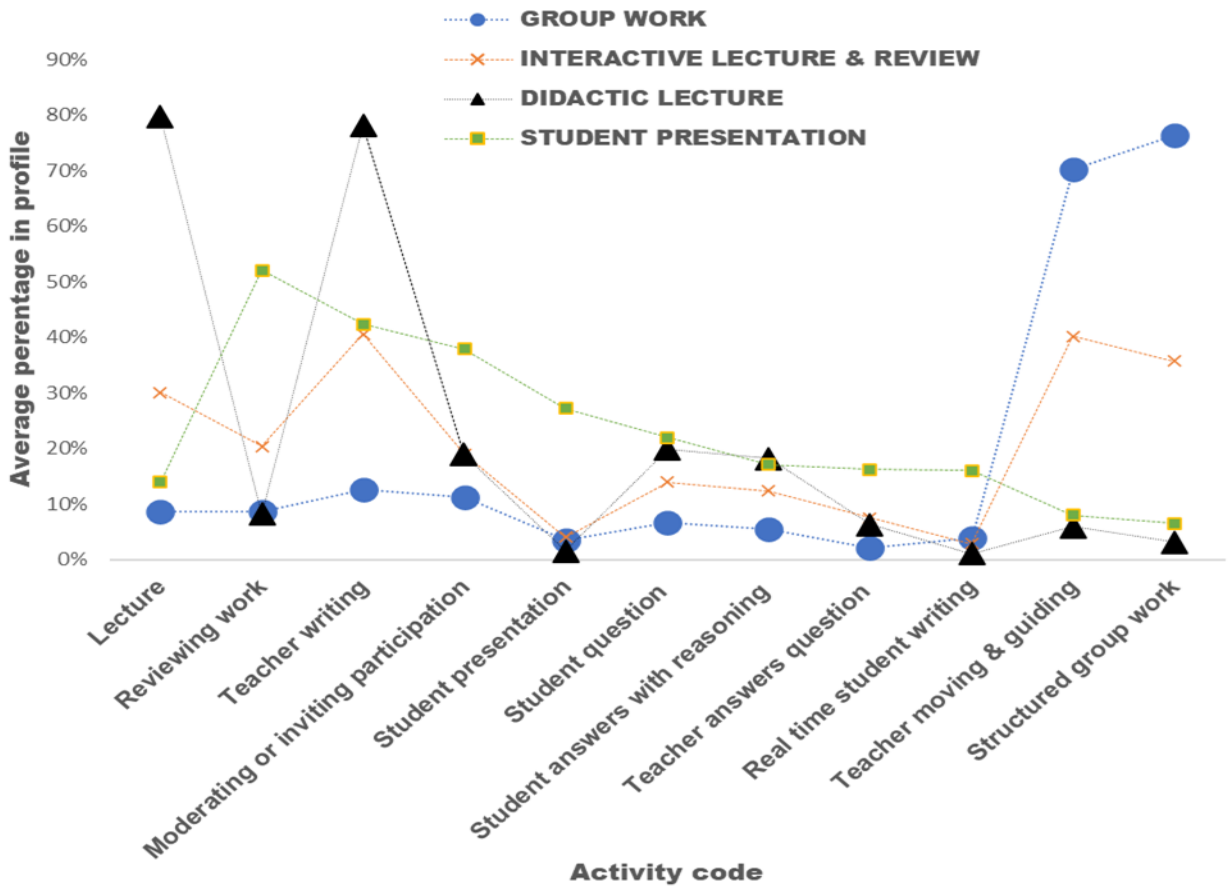
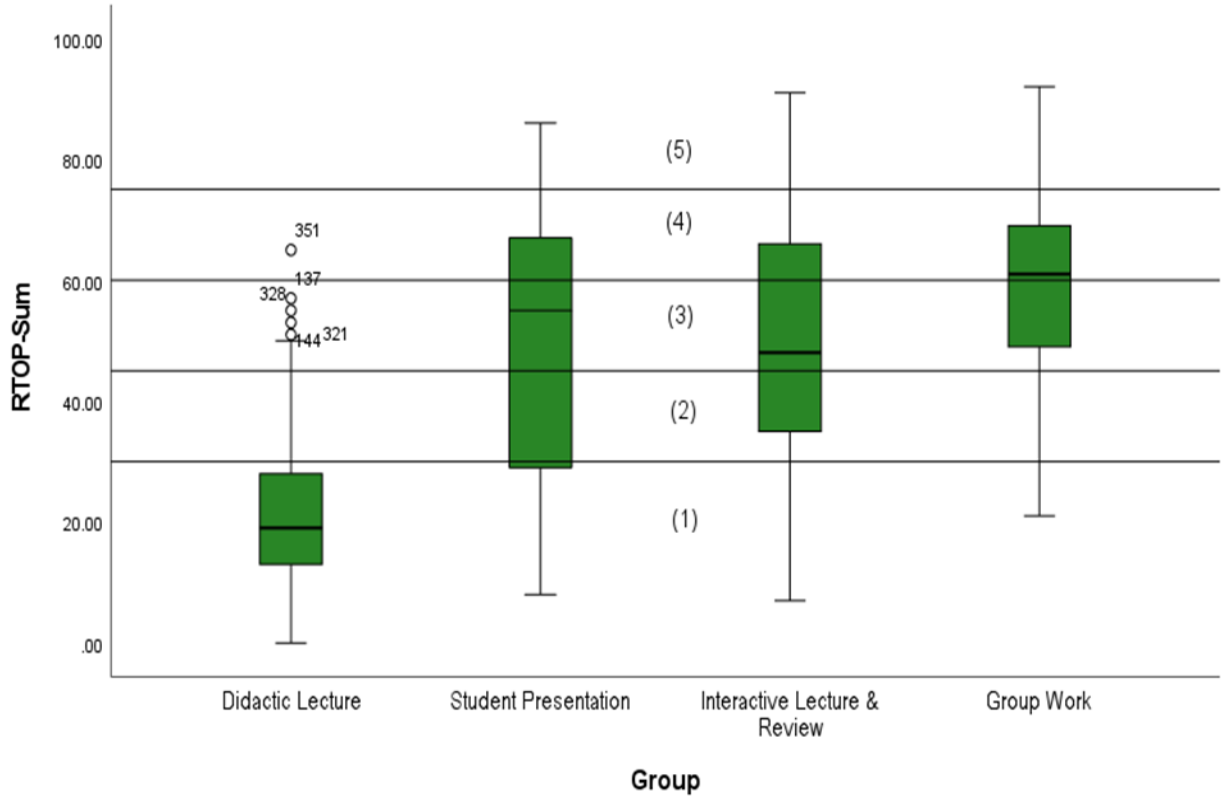


Figure 1. Individual averages for each observation code for each profile.

We first attempted to derive outcome measures based on the TAMI-OP with factor analysis but found resulting dependent measures were not reliable enough to use in analyses. A viable outcome measure derived from the LPA was the simple proportion of non-didactic lecture classes used by each teacher: *Proportion Non-Didactic Lecture* (PND). This is a teacher level measure that is the number of Non-Didactic classes divided by total class observed for the each instructor. 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.

We also examined some of the psychometric qualities of the RTOP-Sum, the dependent measure derived from a total of 25 RTOP numerical ratings. The resulting measure showed high

internal reliability ($\alpha = 0.97$), and the RTOP-Sum and the PND had a very high correlation at r=0.81. Attempted Exploratory and Confirmatory Factor Analyses did not find that proposed RTOP subscales presented as item blocks in the instrument formed separate factors.
The relationship between the four latent profiles found with LPA and RTOP-Sum scores can be seen in figure 2.



Note: 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.

Figure 2. individual averages for each observation code for each profile.

We applied these measures to a sample data set of 15 teachers and 278 classes that included both pre and post observations for the same group of teachers, each of whom provided data from the same or similar courses taught before and after a professional development intervention. To learn how the PND functioned with basic statistical tests, 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 gains. The results presented in Table 1 show statistically significant results for change in both the RTOP and PND measures. While both measures detect significant differences in a pre/post comparison study, the RTOP-Sum has a bigger effect size and lower p-value than the PND, indicating that the RTOP-Sum has greater statistical power in this study.

Table 1: 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 (one-sided)	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 (two-sided)	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

Note: significance levels are indicated by p< 0.05*, p< 0.01 **, p< 0.001***

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 also graphed to learn which activities changed from pre to post (not shown here due to space considerations). 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 among activities also increased.

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*. 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.

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 most statistical tests are robust to non-normality (Glass and 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 comparison with the RTOP-Sum, but in our case provided similar statistical inferences as the RTOP about pre-post change.

There are several other critical caveats to the use of a measure based on LPA or any other clustering technique. The final categorization of classes is dependent on both the sample used and the variables included in the model. The ultimate category where classes end up can vary given the characteristics of the initial pool of classes and the specification of the model (Williams and 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; those using the COPUS can take advantage of the COPUS Analyzer (Harshman and Stains, 2020) an online method for profiling observational data. We also can categorize new classes based on the original clustering algorithm. 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. Also, 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. Related to this are possible interactions between the number of classes sampled for each instructor and the probability that rarer classes will show up in the classes sampled. If greater or fewer classes for each teacher are sampled from pre to post this can create bias in estimates of teaching change. Unequal sampling occurred in our small study because of logistical concerns, ideally pre and post samples should be balanced. 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). 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%).

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.*

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.

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), ar3.

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.

Ferguson, S. L., G. 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 Behaviorial Development*, *44*(5), 458-468.

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

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).

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.

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.* Washington, DC: American Association for the Advancement of Science.

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), ar18.

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 etent of a large-scale transformation in gateway science courses. *Science Advances*, *4*(10), eaau0554.

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., 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.

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*, 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.

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.

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*, 143-151.