Running head: PROFILES OF INSTRUCTIONAL PRACTICE

Describing Profiles of Instructional Practice: A New Approach to Analyzing Classroom Observation Data

Peter F. Halpin and Michael J. Kieffer

New York University

Paper published as: Halpin, P. F. & Kieffer, M. J. (2015). Describing profiles of instructional practice: A new approach to analyzing classroom observation data. *Educational Researcher, 44, 263-277.*

Acknowledgement: The research reported here was supported by funding from the Institute of Education Sciences (R305D140035), and the Spencer Foundation and William T Grant Foundation (Grant No. 201300093). The opinions expressed are those of the authors and do not represent views of these organizations.

Abstract

The authors outline the application of latent class analysis (LCA) to classroom observational instruments. LCA offers diagnostic information about teachers' instructional strengths and weaknesses, along with estimates of measurement error for individual teachers, while remaining relatively straightforward to implement and interpret. It is discussed how the methodology can support formative feedback to educators and facilitate research into the associations between instructional practices and student outcomes. The approach is illustrated with a secondary analysis of data from the Measures of Effective Teaching study, focusing on middle school literacy instruction.

Describing Profiles of Instructional Practice: A New Approach to Analyzing Classroom Observation Data

Recent research in teaching effectiveness has addressed the development and implementation of classroom observational instruments used to evaluate teachers (e.g., Grossman, Cohen, Ronfeldt, & Brown, 2014; Ho & Kane, 2013; Sporte, Stevens, Healey, Jiang, & Hart, 2013; White, 2014). In contrast to other methods currently used for teacher evaluation (e.g., Haertel, 2013; "Asking Students About Teaching," 2012), observational instruments are specifically intended to provide a detailed description of teachers' instructional practices. They thereby offer a strong basis for supporting formative feedback to educators (e.g., Allen, Pianta, Gregory, Mikami, & Lun, 2011). Additionally, observational instruments can facilitate research about the association between instructional practices and a wide range of important student outcomes, including but not limited to academic performance. For example, teachers' scores across a variety of instruments have been found to positively correlate with teachers' value-added (VA) measures, as well as students' self-reported school engagement and socio-emotional development (e.g., Kane & Staiger, 2012; also see Table 2 of this paper).

Despite the growing evidence that in-classroom observations can provide useful information about teaching practices, it remains less clear how that information should be summarized to support inferences about individual teachers. Researchers in teacher evaluation often summarize the observational instruments with a total score, thereby placing teachers along a single dimension of effectiveness (e.g., Ho & Kane, 2013). While this approach may facilitate comparison among teachers, it is not in line with theory and evidence that effective teaching requires the skillful coordination of multiple practices (e.g., Darling-Hammond & Bransford, 2007; Snow, Griffin, & Burns, 2007), and that different teachers may demonstrate different patterns of strengths and weaknesses (e.g., Grossman, Loeb, Cohen, & Wyckoff, 2013). Psychometric research also supports the conclusion that many observational instruments measure multiple dimensions of instructional quality (e.g., Grossman, et al. 2014; Lazarev & Newman, 2014; Savitsky & McCaffrey, 2014), suggesting that teachers' practices are not well described in terms of a single construct.

Another approach – one that has been recommended in professional development programs (e.g., Allen et al., 2011; Danielson, 2013) and appears to be commonly used in evaluation contexts – involves interpreting teachers' scores on the individual items that make up an instrument. A strong rationale for this approach is that the individual items are directly anchored to specific instructional practices, whereas total scores or subscale scores may be more difficult to use for feedback. However, this approach brings into question the reliability of the item scores. Unreliable scores can lead to inaccurate inferences about a teacher's strengths and weaknesses, which is especially problematic when those inferences are used to inform decisions in professional settings.

In this paper we address three central measurement challenges in the application of classroom observational instruments. The first challenge is to provide a measurement methodology that captures the item-level diagnostic information that the instruments are designed to provide to educators. The second challenge is to estimate the measurement error associated with the scores assigned to teachers. Current standards in assessment require that measurement error is reported whenever scores are assigned to individuals (Joint Committee on Educational and Psychological Assessment, 2014). Yet commercial vendors of observational instruments do not commonly report estimates of measurement error, either at the item level or for total scores. Without this information, it is easy for decision-makers to fall into the misconception that scores on the observational instruments are free of error, or that all teachers are measured with equal reliability, which can lead to inappropriate decisions rather than supporting the professionalism of teachers. The third challenge is to ensure that the methodology remains feasible to apply in the settings in which observational instruments are currently used.

To this end we propose the application of latent class analysis (LCA). In this paper we show how LCA can be used to obtain a small number of empirical *profiles of instructional practice*, a term that we adopt from Grossman et al. (2013) to refer to diagnostically useful patterns of practice. We explain how these profiles capture information about teachers' strengths and weaknesses, and illustrate how this information offers new possibilities for supporting formative feedback to educators and for informing research about the association between instructional practices and student outcomes. We also discuss how LCA provides an interpretable summary score and an estimate of measurement error for each teacher. A main reason that we suggest the use of LCA instead of more complicated diagnostic models (see Embretson & Yang, 2013, for a review) is because teachers' scores and their measurement errors can be easily computed, once the parameters of the measurement model have been estimated. For example, the scoring procedure can be applied to new observations using a simple spreadsheet macro.

In the following section we provide background on the three observational instruments that are the focus of the present study. We then provide a conceptual overview of LCA and its implementation in the current context. Next, we illustrate the use of LCA with a secondary analysis of data from the Measures of Effective Teaching (MET) study (Kane & Staiger, 2012). The example focuses on middle school English language arts (ELA) teachers, which is an important and under-researched domain (e.g., Carnegie Council on Advancing Adolescent

Literacy, 2010), as well as a domain in which measurement of teacher effectiveness has been particularly challenging (e.g., McCaffrey, Sass, Lockwood, & Mihaly, 2009). In the example we also address test-retest reliability and criterion-related validity with students' reading achievement, engagement, and social-emotional development.

It is important to emphasize that the empirical research we report is intended only as an illustration of LCA in the context of teacher observations. While the example illustrates the diagnostic interpretation of LCA and how it quantifies measurement error, the results do not support definitive conclusions about any particular instrument or population of teachers. We contextualize our findings in terms of ongoing measurement research in this area. Similarly, the analyses do not test a particular theoretical model of teaching and learning. When making substantive interpretations, we draw on the conceptions of teaching and learning that guided the development of the instruments. In the discussion section, we address implications and limitations of this research. Some technical details and additional figures are provided in the Online Appendices.

Observation Instruments in English Language Arts

In this paper we focus on three observational instruments that have been widely used to observe ELA middle school teachers. Two of these, the Classroom Assessment Scoring System (CLASS; Pianta, Hamre, Hayes, Mintz, & LaParo, 2008), as well as the more recently developed Framework for Teaching (FFT; Danielson, 2013), are intended for use across grades and subjects. The third, the Protocol for Language Arts Teaching Observations (PLATO; Grossman et al., 2013), is intended specifically for ELA teachers in grades four through nine. To be consistent with the data analyses that follow, we note when shortened versions of the instruments were used in the MET study. These versions of the instruments are summarized in Table 1.

CLASS was initially developed as a research tool for addressing the quality of early-childhood classroom environments (Pianta et al., 2005). Its theoretical basis is in human development and ecological systems (e.g., Bronfenbrenner & Morris, 1998), focusing on the daily interactions that take place among teachers and students (Pianta & Hamre, 2009). Multiple versions of the instrument are now used in classrooms ranging from preschool to high school, and in current implementations it is often coupled with a teacher training intervention referred to as *My Teaching Partner* (Allen et al., 2011). As used in the MET study, the instrument has four domains (see Table 1): emotional support, classroom organization, instructional support, and student engagement. The domains are measured using a total of twelve items, each of which is scored on a seven-point scale.

FFT (Danielson, 2013) is grounded in a constructivist view of teaching and learning. The developer emphasizes its use in a concerted professional development model (Danielson, 2011). The instrument has four domains: planning and preparation, professional responsibilities, classroom environment, and instruction. Only the latter two can be scored using classroom observations and are the focus of the present research. As shown in Table 1, both domains include four items, each of which is scored on a four-point scale. It is notable that FFT has recently been adopted by numerous school districts including New York City (New York City Department of Education, 2013) and Chicago (Sporte, Stevens, Healey, Jiang, & Hart, 2013).

PLATO was designed with two explicit premises in mind (Grossman et al., 2013). First, that quality teaching in ELA involves practices that are specific to the effective teaching of reading and writing. Second, that teaching quality is multidimensional, such that effective teachers “possess a range of characteristics and skills that contribute directly and indirectly to improved student outcomes” (Grossman et al., 2013, p. 447). The full instrument includes thirteen items, but the version adopted in the MET study (referred to as "PLATO prime") was shortened to six items (see Table 1). These six items are each scored on a four-point scale, and are grouped into three broader domains: disciplinary demand, instructional scaffolding, and classroom environment (Grossman et al., 2014).

In addition to their demonstrated research potential, the use of observational instruments in professional settings is promising for several reasons. First, they provide educators with a common language for talking about teaching. Second, by moving the focus from year-end student outcomes to improvement of the process of teaching, the instruments can support the professionalism of teachers. Third, by providing teachers with information about their performance on a range of instructional practices, the instruments can frame learning to teach as a long-term, developmental process, rather than a task accomplished before or at the beginning of teachers' careers.

Previous Evidence of Reliability and Validity

Of the three instruments considered in this paper, CLASS has the most extensive psychometric research base (e.g., Hamre & Pianta, 2010; Mashburn, Meyer, Allen, & Pianta, 2009). The MET study made a major contribution by providing comparative evidence for FFT, CLASS, and PLATO (as well as other instruments), drawing on data from over 3,000 teachers of ELA and mathematics in grades four through nine across six school districts in the United States. In Table 2 we summarize the main findings reported by Kane and Staiger (2012).

The reliability coefficients of FFT, CLASS, and PLATO ranged from .31 to .37 when the instruments were administered by trained raters using 15-25 minute video recordings of teachers' classroom practices. These coefficients were computed for the total scores and are interpretable in terms of the proportion of variance over multiple administrations of the instruments (i.e., test-retest reliability). By quadrupling the number of administrations, reliabilities in the range of .6 to .7 were expected (also see Ho & Kane, 2013). The correlations of the total scores with teachers' VA measures on ELA year-end state examinations were positive but relatively weak. For the SAT-9 open-ended reading examination, mean differences of .10 to .16 standard deviation units were observed between teachers in the top and bottom quartiles of the observational instruments. Similarly, mean differences on student engagement and social-emotional development were between .05 and .18 standard deviation units. In an analysis of the same dataset, Grossman et al. (2014) reported correlations between SAT-9 VA measures and each of three PLATO prime subscales, with the highest correlation for classroom environment (r = .15), followed by disciplinary demand (r = .12), and instructional scaffolding (r = .04).

These results summarize the current state of the literature and indicate that there is an important role to be played by continued measurement research. In particular, the test-retest reliability and criterion-related validity of total scores leave room for improvement, and we are not aware of any research that has addressed the standard error of measurement of scores assigned to teachers, or the application of diagnostic measurement models. These topics are the focus of what follows.

Latent Class Analysis of In-Classroom Observation Data

Conceptual Overview

LCA involves latent (unobserved) variables that are indicated by measured (observed) items. In contrast to more common measurement models such as factor analysis and item response theory, the latent variable in LCA is categorical rather than continuous. Contemporary psychometric research has seen a return to models that involve categorical latent variables, because of their diagnostic properties (see Embretson & Yang, 2013). In addition to its psychometric applications, LCA is often used as a method of model-based clustering. From this perspective, LCA is an individual-centered approach, used primarily to identify persons who cluster together based on similarities in their item scores. In contrast, factor analysis is often considered a variable-centered approach, in which the goal is to identify items that hang together with one another.

In the present application, each latent category represents a group of teachers with similar instructional practices, as measured by their scores on the items of the observational instruments. Because the observed variables are not correlated within the latent classes (see equation A2 in the appendix), the latent variable suffices to explain systematic differences in teachers' instructional practices. Therefore, the latent variable can be interpreted to distinguish what is unique about teachers' practices (i.e., signal) from the measurement error of the instrument (i.e., noise).

Within each latent category, the most probable score on each item can be estimated (see equation A3 in the appendix). We interpret these scores as representing the profile of instructional practice of the teachers in that category. LCA has the advantage of allowing for statistical inferences about the specific practices that are important for differentiating among the profiles. The diagnostic value of LCA comes from interpreting the profiles in light of theory, research, and practice in teaching and learning, which we illustrate in the example. In combination with evidence about how the profiles are related to student outcomes, this can provide a strong basis for informing feedback to educators and future research.

LCA also supports inferences about the most likely profile for each individual teacher (see equation A4 in the appendix). This is used to assign teachers a "profile membership score," which replaces the use of a total score as a summary measure. As noted, a main reason that we suggest the use of LCA is because profile membership scores can be easily computed; once the parameters of the model have been estimated, no special software is required to apply the scoring method to new observations.

Importantly, not all teachers will be equally well described by any given set of latent categories, which is why we have also emphasized the role of measurement error. LCA quantifies measurement error in terms of the profile membership distribution of each teacher, which provides information about how likely a teacher is to have been misclassified. This is a direct measure of the uncertainty associated with each teacher's score, and is also easily obtained from equation A4 once the model has been estimated.

Estimation

LCA can be estimated in readily available software programs such as Stata (StataCorp, 2013) and Mplus (Muthen & Muthen, 2014). Estimation proceeds in two phases. First, the parameters of the model are estimated for a fixed number of latent classes. Second, the number of latent classes is inferred by comparing the goodness of fit of different models. The number of latent classes that best fit the data can be assessed using multiple criteria (e.g., Lubke & Neale, 2006, 2008). In this research we emphasize classification accuracy, since our main purpose is to reliably assign teachers to profile memberships. We also report the Bayesian information criterion (BIC), which is known to perform well with discrete latent variables.

To address the possibility of model misspecification, we recommend the use of robust standard errors when inferring differences between profiles. In particular, clustering of teachers within schools should be addressed using cluster-robust methods. It is also important to note that the observational instruments are often administered repeatedly to the same teacher (e.g., over different lessons, by different raters). In research applications it is usual to aggregate scores to the teacher level (e.g., Kane & Staiger, 2012; Grossman et al. 2014), which is the approach we take here. In the discussion section we consider the limitations of this approach and the potential of methods that explicitly model the cross-classified structure of observational data.

Application to Observational Data from Middle School ELA Classrooms

This section illustrates the application of LCA to the observational instruments. We first discuss the fit of LCA to the data, then address the diagnostic interpretation of the profiles and the measurement error of teachers' profile memberships. Subsequently, we consider the consistency of profile memberships over multiple administrations of the instruments, and finally discuss the relationship between the profile memberships and a number of student outcomes. The latter two topics provide some initial indications of test-retest reliability and criterion-related validity, allowing for comparison with prior analyses of the same data reported in Table 2.

It is important to emphasize that this research should not be interpreted as validating any particular instrument, and replication of the proposed methodology with other samples would be required before drawing definite conclusions about the characteristics of teachers or of the instruments. These data analyses offer an illustration of the potential of LCA—what it can provide in terms of diagnostic information and quantifying measurement error.

Sample

We conducted a secondary analysis of the first year (AY2009-2010) of the MET longitudinal database. The sample consisted of all ELA teachers in grades six through eight for whom data was available on at least one of FFT, CLASS, and PLATO. The resulting sample consisted of 381 middle school ELA teachers from four school districts. Most teachers (94%) taught two class sections, and there were an average of 23.48 students per section. The teachers were employed in a total of 95 schools, with most schools (68%) having between two and four teachers.

Measures

The observational instruments were administered by trained raters using 15-25 minute videos of teachers' classroom practices. All instruments were administered multiple times to the same teacher, with a minimum of two and a maximum of sixteen administrations, depending on how many videos the teacher submitted. For PLATO and CLASS, most teachers had eight replications, and for FFT the mode was four. As mentioned, ratings were aggregated to the teacher level prior to analysis.

The criterion-related validity of profile memberships was addressed in terms of the association with student outcomes. To facilitate comparison with previous research, we focus on outcomes similar to those described in Table 2. These are (a) achievement on the SAT-9 open-ended reading examination, as a proxy for student learning, (b) engagement, as indexed by students' self-reported in-classroom effort, and (c) social-emotional development, as indexed by students' self-reported positive emotional experiences in the classroom. Note that we do not consider teacher's VA measures on state ELA examinations, but we do control for previous achievement on state exams as well as other student characteristics when estimating validity coefficients. The measures used for student engagement and social-emotional development both had internal consistency reliability coefficients of approximately .7. Further details on the validation methodology and outcome measures are provided in Appendix B.

Results

Estimation. LCA was conducted using all 26 items in Table 1. Using the Bayesian information criterion, a model with four classes was found to have a better fit to the data than one with fewer classes (see Table A1 in Appendix A). Increasing the number of classes to five resulted in marginal improvement in fit, but the fifth class included only eight teachers. These eight teachers shared the same profile membership in the four-class solution (profile D, described below), and were distinguished by very low scores on all items. We preferred the four-class solution because it provided sufficient sample size for comparisons among classes. Additionally, previous work has found three continuous factors when analyzing the observational instruments (e.g., Kane & Staiger, 2012; Grossman et al. 2014; Lazarev & Newman, 2014). A three-factor solution corresponds to four latent classes (see Halpin, Dolan, Grasman, & Deboeck, 2011), which provides additional rationale for preferring the four-class solution.

The classification accuracy of the four-class model was very satisfactory, as estimated by classification entropy (.942) and average classification probability (.966). This means that nearly all of the teachers in the sample had a high probability of belonging to one and only one latent class. We discuss teacher profile membership scores further below.

Interpretation of the Profiles. We labeled the classes A through D. Profile A contained the smallest number of teachers, (N_A = 66) and profile B the largest (N_B = 136), while profiles C and D contained approximately equal numbers of teachers (N_C = 93; N_D = 86). Figure 1 presents a detailed comparison of the four profiles in terms of their ratings on each item of PLATO. Similar comparisons are provided for CLASS and FFT in Figures A1 and A2 in Appendix C. The interpretation of the plots is similar to the “interaction plots’’ frequently reported with ANOVA. In each panel of Figure 1, the horizontal axis represents the individual items of PLATO, which are grouped according to the three domains identified by Grossman et al. (2014): disciplinary demand (items 1 and 2), instructional scaffolding (items 3 and 4), and classroom environment (items 5 and 6). The vertical axis represents scores on the PLATO items and the lines represent the mean performance of each profile. Note that the ordering of the groups of items on the horizontal axis is arbitrary. When interpreting the panels, the focus is on the relative elevation of the lines, rather than their upward or downward slopes.

The top left panel of Figure 1 shows all four profiles. The profiles are roughly ordinal, with Profile A scoring the highest on all competencies, and Profile D scoring the lowest. The remaining panels compare adjacent profiles, with the shaded areas representing 95% confidence intervals. These latter three plots are useful for interpreting where the profiles are differentiated from one another, and where they overlap. In general, it can be seen that not all of the profiles are differentiated on all of the PLATO items, suggesting that the profile memberships can provide meaningful diagnostic information beyond their rank ordering. We summarize our diagnostic interpretation in the following three points, relying on the theoretical framework for PLATO (e.g., Grossman et al., 2014). Effect sizes are reported using Cohen's *d* and computed using the parameter estimates from Mplus.

1) As shown in the top right panel, profile C performed better than profile D on Time Management ($d = 1.03$) and Behavior Management ($d = 1.36$), but these profiles were not differentiated on items related to disciplinary demand or instructional scaffolding. Teachers assigned to these two profiles are likely to share common weaknesses in multiple aspects of instruction. However, teachers assigned to profile D are also more likely to have weaknesses in the management of classroom environments.

2) As shown in the bottom left panel, teachers assigned to profiles B and C are likely to share common strengths in their classroom environments. However, Profile B performed better than profile C on items related to both disciplinary demand and instructional support, with effect sizes ranging from $d = 0.43$ on modeling to $d = .71$ on classroom discourse. In comparison to

teachers in profile B, those in profile C are more likely to have weaknesses in aspects of their instruction other than classroom environments.

3) As shown in the bottom right panel, profile A performed better than profile B on disciplinary demand (Intellectual Challenge: d = 0.51; Classroom Discourse: d = .59), but the profiles were not otherwise differentiated. As argued by Grossman et al. (2014), higher levels of disciplinary demand can be considered evidence of more ambitious instructional practices. This interpretation is also supported by theory and evidence suggesting that high levels of intellectual challenge (e.g., Newmann, Lopez, & Bryk, 1998) and rich classroom discourse (e.g., Nystrand & Gamoran, 1991) are important, but often rarely observed, aspects of highly effective ELA instruction. We conclude that teachers assigned to profile A exhibited more ambitious instructional practices relative to those in profile B.

Measurement error in teachers' profile memberships. LCA can be used to estimate the probability that an individual teacher belongs to each of the four profiles. The probabilities are computed using equation A4 in Appendix A, and are included as standard output with Mplus. The resulting profile membership distributions for three teachers in our sample are depicted in Figure 2.

It can be seen that Teacher 1 has a very high probability of being in profile A, and a very low probability of being in any other profile. Therefore Teacher 1 is accurately (or precisely) classified into profile A. The most likely membership for Teacher 2 is also profile A, but Teacher 2 has a non-negligible probability of being in Profile B as well. Therefore, although both Teachers 1 and 2 would be assigned the same profile membership, we are less certain about the profile membership of Teacher 2. Unlike the other two teachers, Teacher 3 was most likely to be a member of profile B, but like Teacher 2, also has a non-negligible chance of belonging to another profile.

It is notable that all three teachers had a negligible probability of belonging to more than two adjacent profiles. This was true of all teachers in our sample. Also, as reported above, most teachers had a high probability of belonging to one and only one profile – in fact, over 90% of teachers in our sample were assigned to a single profile with near certainty. Thus, the profile membership distribution of Teacher 1 was not at all exceptional for the teachers in our sample.

For teachers who were not well represented by any single profile, it is important to consider the degree of measurement error in his or her profile membership. As a starting point for addressing this issue, we recommend that profile membership distributions such as those in Figure 2 be reported when providing feedback to individual teachers and when making decisions in professional settings.

Test-retest reliability. As noted above, we used teacher-level averages to estimate LCA. This approach has the drawback that it does not provide model-based information about the consistency of a teacher’s profile membership over different administrations of the observational instruments. To address this issue, we used the parameter estimates from the teacher-level model to compute the most likely profile membership for each observation occasion (see Appendix D). Using this approach, we found that teachers were consistently classified into the same profile on 57% of administrations. However, this estimate does not take into account the base rate probability of the profiles. Correcting for the base rate resulted in an average “chance-corrected” consistency index of 44%. These measures of classification consistency leave much to be desired, but are reasonable in light of previous research on test-retest reliability (see Table 2). In Appendix D, we discuss how a teacher-level analysis may have led to an underestimation of test-retest reliability. In general, assessing classification consistency in the context of teacher observations is an area that requires further research, and we return to this point in the discussion.

Criterion-related validity. To provide evidence about criterion-related validity of the profile membership scores, we investigated their association with three student outcomes: reading achievement, as assessed by teachers' VA measures on the SAT-9 open-ended reading examination; student engagement, as indexed by students' self-reported in-classroom effort; and social-emotional development, as indexed by students' self-reported positive emotional experiences in the classroom. Figure 3 presents the standardized mean differences on these three outcomes, and the error bars represent the 95% confidence interval on the means.

The difference in SAT-9 VA measures between profiles A and D ($d = 0.22$) was comparable to a year of learning, using a previously established benchmark for the MET data (see Kane & Staiger, 2012; Grossman et al., 2014). The mean difference between profiles A and D was of similar magnitude for student engagement ($d = 0.23$), and was slightly larger for student social-emotional development ($d = 0.35$). Although the overall pattern of mean differences was the same for each outcome, there were notable distinctions. In particular, when SAT-9 was the outcome, only profile A was significantly different from the other three profiles; however, when student engagement or social-emotional development was the outcome, both profiles A and B were significantly different from profile D.

Overall, the effect sizes reported in this analysis leave room for improvement. However, they are appreciable in the context of previous research (see Table 2).

Discussion

Implications for Research on Teaching and Learning

One potential use of the diagnostic information captured by LCA is to support teachers' professional development. As illustrated in Figure 2, LCA offers empirically-derived profiles of instructional practice. Such profiles can facilitate the development of interventions that are both *targeted* at the needs of individual teachers and *coordinated* across multiple domains of practice. While it is not uncommon for interventions based on observational instruments to be targeted at specific skill deficits (Allen et al., 2011; Danielson, 2013), it is rare that such efforts explicitly consider how the targeted skills are interrelated with other aspects of instructional practice. The latter point is of particular importance when combining multiple existing programs, or for identifying teachers who have been targeted for a program that is not appropriate to their needs.

We illustrate these ideas with reference Figure 1. It is apparent that teachers in profile D could benefit from development of classroom environment techniques, whereas other teachers did not have weaknesses in this area. This demonstrates an avenue for targeting an intervention at teachers in profile D. However, it would be inadvisable for teachers in profile D to receive an intervention for disciplinary demand and/or instructional scaffolding, without also including training in classroom environment techniques. This is because teachers who are better at the former two domains (i.e., teachers in profiles A and B) also tend to be better at the latter. In other words, such an intervention would not be trying to make teachers in profile D more like their higher performing peers; it would ostensibly be trying to invent a type of teaching for which there is no empirical basis. This demonstrates how profiles of instructional practice can inform coordination of an intervention over multiple skill domains. For teachers in profile D, a basic conclusion is that any intervention program should address classroom environment techniques.

Consider next teachers in profile C. These teachers might seek to make improvements in instructional scaffolding and/or disciplinary demand, but would not benefit much from training in classroom environment techniques. Thus, teachers in profile C might be targeted for exactly the opposite kind of training identified for teachers in profile D: anything other than classroom environment techniques. Based on the overall scores on instructional scaffolding and disciplinary demand shown in Figure 1, such a program may also be appropriate for teachers in profiles A and B. However, if the goal of an intervention was to make the instructional practices of teachers in profile B more similar to those of teachers in profile A, the focus should be exclusively on disciplinary demand.

Profiles of instructional practice are certainly not the only means by which to conceptualize programs that are targeted at teachers' needs and coordinated over domains of practice. However, they do offer a rigorous quantitative basis for supporting the development and implementation of such programs. Based on a teacher's profile membership, it is possible to identify his or her specific strengths and weaknesses with respect to a given observational instrument. The certainty with which a teacher is assigned to a profile should be used to inform decisions about the appropriateness of potential programs for that teacher, and this information is available from his/her profile membership distribution (see Figure 2). Comparison of the strengths and weakness of different profiles provides an empirically-driven basis for developing and combining programs that coordinate multiple skill domains. While some configurations of skills correspond to the practices of real teachers, other combinations may be theoretically possible but lack empirical evidence. Of course, the results from our data analyses cannot speak to the efficacy of such intervention efforts – they merely illustrate the potential of LCA in the context of teachers' professional development.

This application of LCA can also inform research about the associations between teaching practices and student outcomes. For example, previous research on PLATO has used multiple regression analysis to identify that the domains of disciplinary demand and classroom environment are uniquely related to SAT-9 VA measures (Grossman et al., 2014). The results shown in Figure 3 provide further and arguably stronger evidence about the unique contributions of these two domains. In the figure it can be seen that weaknesses in classroom environment (i.e., profile D) are associated with below average SAT-9 VA measures. The remaining three profiles all demonstrated strengths in classroom environment, but only profile A was associated with SAT-9 VA measures that were above average. The skill configuration that uniquely identified teachers in profile A was relative strengths in disciplinary demand. In line with previous research, this interpretation suggests that disciplinary demand can be considered evidence of more ambitious instructional practices.

As another example, Figure 3 also suggests that the relative importance of different teaching practices depends on the student outcomes under consideration. In particular, only profile A was associated with gains in student reading achievement on SAT-9, but both profiles A and B were associated with higher levels of student engagement and social-emotional development. Accordingly, consideration of different student outcomes may lead to different conclusions about the “effectiveness” of specific teaching practices. In general, examination of the relationships between profiles of instructional practice and student outcomes can provide a rich basis for generating and testing research hypotheses.

Limitations and Future Measurement Research

As emphasized throughout this paper, the empirical results we have discussed are intended to be illustrative, not definitive. A first step to establishing more general findings is to replicate this approach with other samples of teachers. Further, we have not described how profile memberships are related to teachers’ individual differences or classroom and school contextual factors – this is an important domain of research that can be fruitfully pursued once the psychometric properties of the instruments have been more firmly established. Additionally, we have been careful to avoid causal interpretations of the associations between profiles of instruction and student outcomes. Such interpretations are not supported by the data we have analyzed. However, there is clear potential for such claims to be made by combining the proposed measurement methodology with an appropriate experimental design (e.g., random assignment of teachers to classrooms).

We have also noted that LCA conducted at the teacher level is not well suited to answer questions about within-teacher variation (e.g., variation over multiple administrations of the observational instruments). A measurement model that more directly addresses the cross-classified structure of teacher observation data, while also remaining feasible to implement in practice, would be a substantial contribution to the literature on teaching and learning. Potential methods have been developed by Savitsky and McCaffrey (2014) as well as Casabianca and Junker (2013), although their test-retest reliability, criterion-related validity, diagnostic value, and applicability remain topics for investigation. These are the key challenges to be addressed by future measurement research.

Conclusions

We have illustrated an approach to analyzing data from classroom observational instruments based on LCA. Current approaches often rely on total scores, which only allow for a rank ordering of teachers along a single dimension of effectiveness. Other approaches have relied on interpretations at the item level, but have not explicitly considered the statistical relationships among the instructional practices measured by the individual items. LCA has the advantage of providing empirically-derived profiles of instruction that describe what real teachers are doing in their classrooms. Additionally, LCA provides an estimate of the measurement error associated with each teacher's profile membership, and thereby addresses a major shortcoming of current practice. The proposed measurement methodology is readily available to researchers using software for latent variables and structural equation modeling. Once the model has been estimated, the scoring procedure and its measurement error are easily computed. This means that school districts, vendors, and researchers can readily make use of the results of LCA in practice.

The initial analyses reported here indicate that the test-retest reliability of LCA is on par with approaches based on total scores. We also found that the profile memberships were associated with students' academic achievement, engagement, and socio-emotional development, with effect sizes close to double those previously reported using the same data (see Table 2). These remain important areas for improvement.

Finally we have discussed the potential applications of this measurement methodology. We have considered how profiles of instructional practice can be used to support the development and implementation of programs that are targeted at teachers' needs and coordinated over domains of practice. We have also suggested how research into the role of instruction in student outcomes can benefit from this methodology. By illustrating the potential of LCA, and of more rigorous measurement practices in general, we hope this research can contribute to the improved use and interpretation of classroom observations in research and in professional settings.

References

Allen, J., Pianta, R., Gregory, A., Mikami, A., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333,* 1034–1037.

Asking students about teaching: Student perception surveys and their implementation. (2012). Bill and Melinda Gates Foundation.

Bronfenbrenner, U., & Morris, P. (1998). The ecology of developmental processes. In R. M. Lerner (Ed.), *Theoretical Models of Human Development.* Vol. 1 of the *Handbook of Child Psychology (5th ed.)* (pp. 993-1028). Editor-in-chief: William Damon. New York: Wiley.

Carnegie Council on Advancing Adolescent Literacy. (2010). *Advancing adolescent literacy: The cornerstone of school reform.* New York: The Carnegie Corporation of New York.

Casabianca, J. M., Junker, B. W., & Patz, R. (in press). The hierarchical rater model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Boca Raton, FL: Chapman & Hall/CRC.

Darling-Hammond, L., & Bransford, J. (Eds.). (2007). *Preparing teachers for a changing world: What teachers should learn and be able to do.* New York: John Wiley & Sons.

Danielson, C. (2011). Evaluations that help teachers learn. *The Effective Educator, 68,* 35-39.

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 instructionally focused edition* Princeton, NJ: The Danielson Group

Embretson, S., & Yang, X. (2013). A multicomponent latent trait model for diagnoses. *Psychometrika, 78,* 14–36.

Gelman A. & Hill J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education. 119,* 445-470.

Grossman, P. Cohen, J. Ronfeldt, M. & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher valued-added on multiple types of assessment. *Educational Researcher, 43,* 293-303.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores. William H. Angoff Memorial lecture series.* Princeton, NJ: Educational Testing Service.

Halpin, P., Dolan, C., Grasman, R., & Boeck, P. De. (2011). On the relation between the linear factor model and the latent profile model. *Psychometrika*, (2006), 564–583.

Ho, A. D., & Kane, T. J. (2012). *The reliability of classroom observations by school personnel.* Bill and Melinda Gates Foundation.

Joint Committee on Educational and Psychological Testing. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA Publications.

Kane T. J. & Staiger, D. O. (2012) *Gathering Feedback for Teaching: Combining high quality observations with student surveys and achievement gains*. Bill and Melinda Gates Foundation.

Lazarev, V. & Newman, D. (2014). *Can Multifactor Models of Teaching Improve Teacher Effectiveness Measures?* In. S. Corcoran (Chair), Policy Considerations in the Implementation of MMTES. Paper session conducted at The Association for Education Finance and Policy, San Antonio, TX. Available at SSRN: http://ssrn.com/abstract=2493544

Lee, W., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement, 33,* 374–390.

Lubke, G., & Neale, M. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood. *Multivariate Behavioral Research, 41,* 499–532.

Lubke, G., & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research, 43,* 592–620.

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2013). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement, 74,* 400-422.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effectiveness estimates. *Education Finance and Policy, 24,* 572606.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: John Wiley and Sons.

Muthén, B. O., & Muthén, L. (2014). *Mplus 7.3 [computer software]*. Los Angeles, CA: Muthén & Muthén.

Newmann, F. M., Lopez, G., Bryk, A. S. (1998). *The quality of intellectual work in Chicago schools: A baseline report.* Chicago, Il: Consortium for Chicago School Research.

New York City Department of Education (2013). *Introduction to NYCDOE's new teacher evaluation and development system*. Retrieved May 18, 2014, from http://www.learndoe.org/dhr/recording-new-teacher-eval/

Nystrand, M., & Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English,* 261-290.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9,* 144-159.

Pianta R.C., Hamre B.K., Hayes N., Mintz S., LaParo K.M. (2008). *Classroom Assessment Scoring System–Secondary (CLASS-S)* Charlottesville, VA: University of Virginia

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38,* 109–119.

Savitsky, T. D., & McCafrey, D. F. (2014) Bayesian hierarchical multivariate formulation with factor analysis for nested ordinal data. *Psychometrika, 79,* 275 – 302.

Snow, C., Griffin, P., & Burns, M. S. (Eds.). (2007). *Knowledge to support the teaching of reading: Preparing teachers for a changing world*. New York: John Wiley & Sons.

Sporte, S. E., Stevens, W. D., Healey, K., Jiang, J., & Hart, H. (2013). *Teacher evaluation in practice: Implementing Chicago's REACH students.* Chicago, Il: The University of Chicago Consortium on School Research.

StataCorp. (2013). *Stata [Computer Software]* (13th ed.). College Station, TX: StatCorp LP

White, T. (2014). *Adding Eyes: The Rise, Rewards, and Risks of Multi-Rater Teacher Observation Systems*. Stanford, CA: Carnegie Foundation for the Improvement of Teaching.

Table 1. *Items from observation instruments, as used in ELA classrooms in MET study*

Instrument	**Domains (in Italics) and Items**			
FFT	*Classroom Environment*			
	1. Creating an environment of respect and rapport	2. Using questioning & discussion techniques	3. Establishing a culture of learning	4. Managing classroom procedures
	Instruction			
	5. Communicating with students	6. Managing student behavior	7. Engaging students in learning	8. Using assessments in instruction
CLASS	*Emotional Support*			
	1. Positive climate	2. Negative climate	3. Teacher sensitivity	4. Regard for student perspectives
	Classroom Organization			
	5. Behavior Management	6. Productivity	7. Instructional learning formats	
	Instructional Support			
	8. Quality of feedback	9. Content understanding	10. Analysis and problem solving	11. Instructional Dialogue
	Student Engagement			
	12. Student Engagement			
PLATO prime	*Disciplinary Demand*			
	1. Intellectual challenge	2. Classroom discourse		
	Instructional Scaffolding			
	3. Modeling	4. Explicit strategy use & instruction		
	Classroom Environment			
	5. Behavior management	6. Time management		

Note: Items within each instrument are numbered to facilitate comparison with Figures 1, A1 and A2.

Table 2. *Summary of reliability and validity findings from the MET study.*

Instrument	Total score reliability (proportion of variance)	Value-added on state exams (correlation)	Mean difference: SAT-9	Mean difference: In-class Effort	Mean difference: Positive Emotional Experiences
CLASS	.31	.10	.10	.9	.18
FFT	.37	.11	.14	.12	.18
PLATO	.34	.24	.16	.08	.05

Note: All figures were originally reported by Kane & Staiger (2012). Total score reliability was reported in their Table 11; Value-added correlations were reported their Table 18; All other figures are from their Table 19. Value-added was measured using state examinations (see Kane & Staiger, 2012, p. 41). All mean differences are between teachers scoring in the top and bottom quartiles of the observational instruments, and are reported in standard deviation units at the student level. SAT-9 is an open ended ELA assessment (see Kane & Staiger, 2012, p. 6). In-classroom effort and positive emotional experiences were measured by student self-reports (see Kane & Staiger, 2012, pp. 49-50).

Table A1. *Fit statistics for LCA solutions with 2 to 5 classes.*

Number of classes	2	3	4	5
BIC	11937	10249	10010	9984
Entropy	0.964	0.953	0.942	0.950
ACP	0.990	0.988	0.968	0.971

* Note: BIC = Bayesian Information Criterion; ACP = Average Classification Probability

Table A2. *Comparison of LCA profiles with total score quantiles*

	Total Score Quantiles			
LCA	High	High-Middle	Low-Middle	Low
Profile A	57	9	0	0
Profile B	9	121	6	0
Profile C	0	6	61	26
Profile D	0	0	26	60

Note: Quantiles on total score chosen so that the quartiles had the same number of teachers as the corresponding profile. Profile A *N* = 66; Profile B *N* = 136; Profile C *N* = 93; Profile D *N* = 86; Cohen's Kappa = .70.

Figures

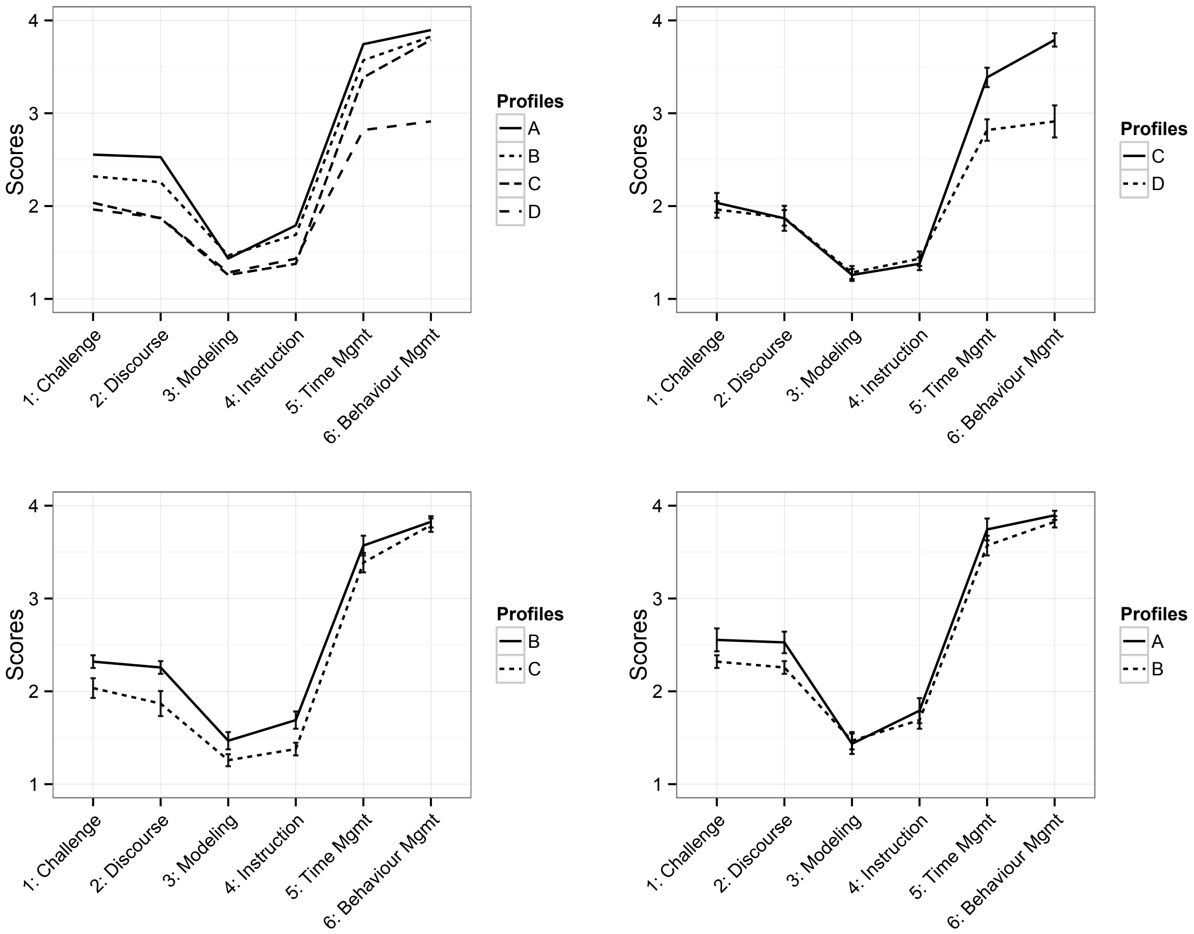

Figure 1. Mean item scores of the four profiles on PLATO Prime. Gray areas indicate 95% confidence intervals on mean estimates. The numbering of the items corresponds to Table 1, which also provides a more complete description of the items. Note that the ordering of the groups of items on the horizontal axis is arbitrary. When interpreting the panels, the focus is on the relative elevation of the lines, rather than their upward or downward slopes.

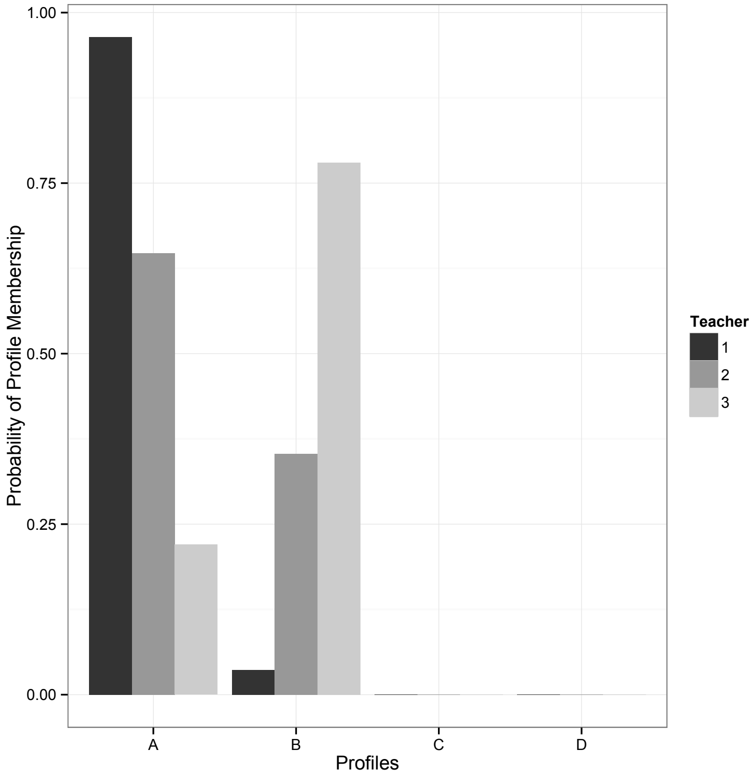

Figure 2. Profile membership distributions for three teachers

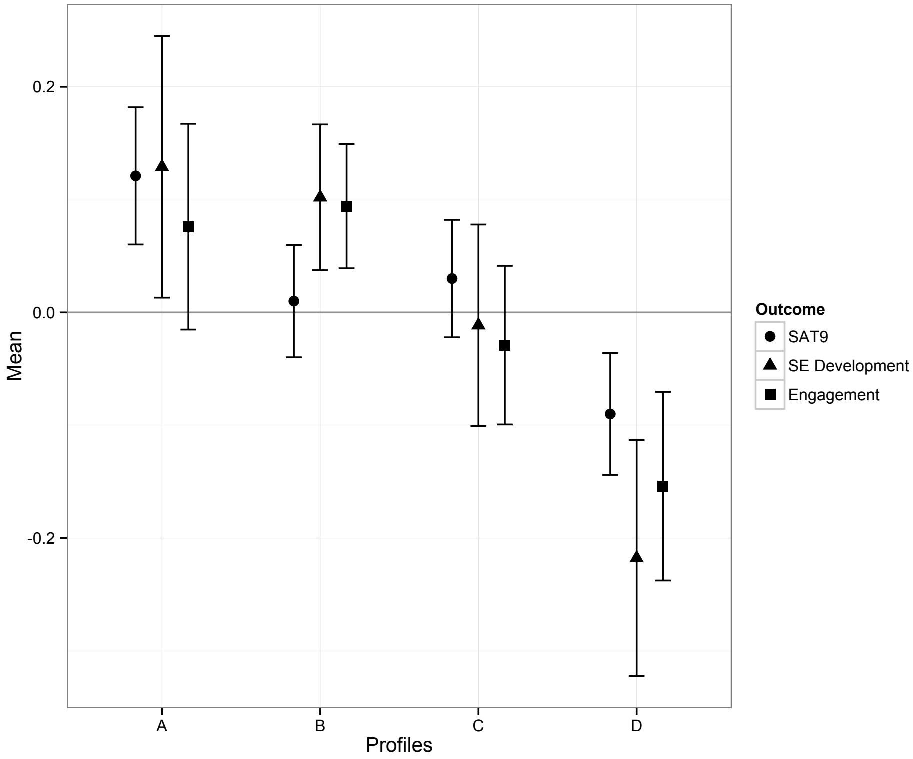

Figure 3. Profile means on three outcomes: SAT-9 open-ended reading exam; social-emotional development ("SE Development"), indexed by self-reported positive emotional experiences in the classroom, and student engagement ("Engagement") indexed by self-reported in-classroom effort. Error bars depict 95% confidence intervals. All outcomes are reported as z-scores.

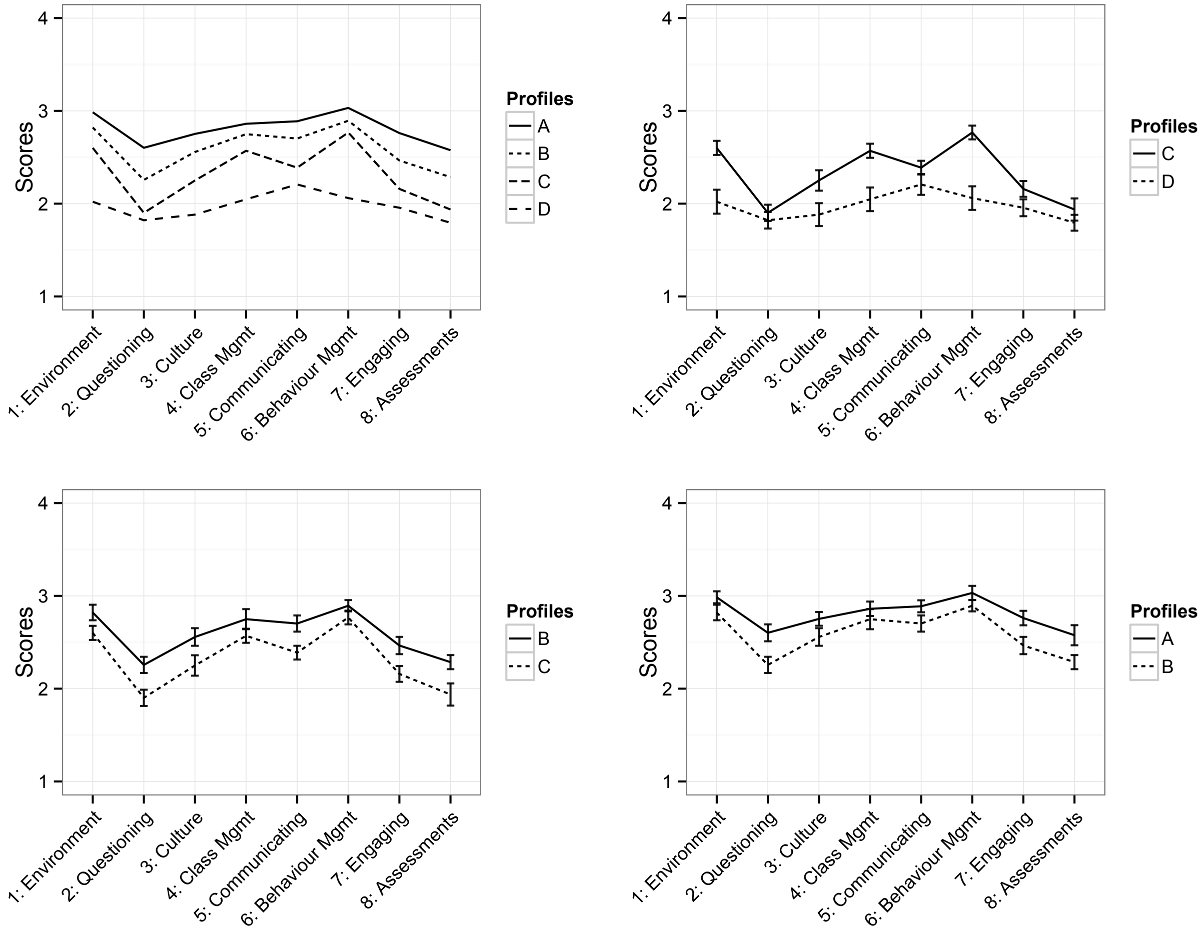

Figure A1. Mean item scores of the 4 profiles on FFT. Gray areas indicate 95% confidence intervals on mean estimates. The numbering of the items corresponds to Table 1, which also provides a more complete description of the items. Note that the ordering of items on the horizontal axis is arbitrary. When interpreting the panels, the focus is on the relative elevation of the lines, rather than their upward or downward slopes.

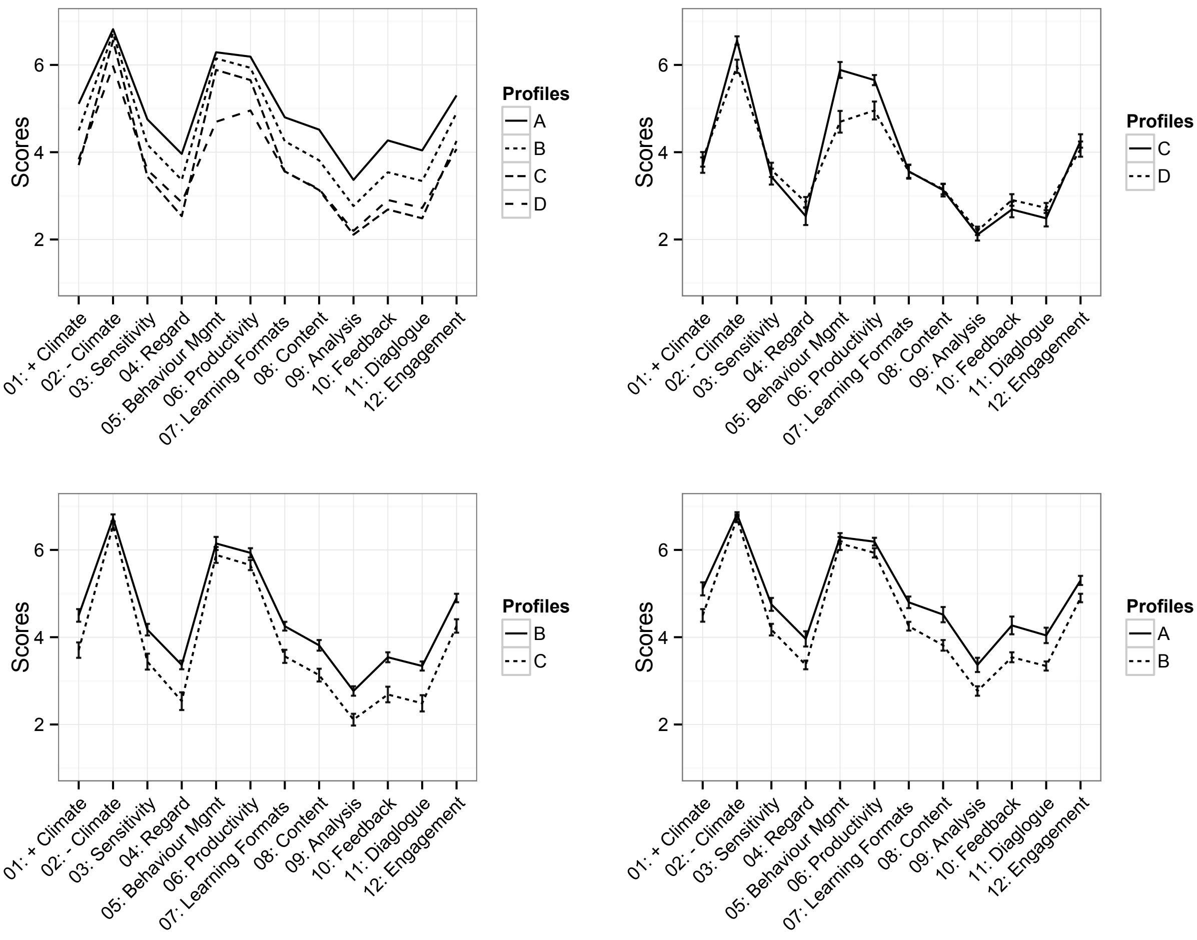

Figure A2. Mean item scores of the 4 profiles on CLASS. Gray areas indicate 95% confidence intervals on mean estimates. Item 02 (Negative Climate) is reverse coded. The numbering of the items corresponds to Table 1, which also provides a more complete description of the items. Note that the ordering of items on the horizontal axis is arbitrary. When interpreting the panels, the focus is on the relative elevation of the lines, rather than their upward or downward slopes.

Appendix A

Some Technical Details on Latent Class Analysis

Latent Class Analysis (LCA) is a latent variable model for multivariate data. The data can be represented as random *J*-vector, $\boldsymbol{X} = (X_1, X_2, \ldots, X_\mathrm{J})$. In application to the observational instruments, the X_j are the individual items (also termed "dimensions", "components", "elements"). The item level data can be conceptualized as either discrete or continuous. In the continuous case, the data are typically treated as a Gaussian mixture (see McLachlan & Peel, 2000) and the model is sometimes referred to as latent profile analysis. The following presentation is applicable to both approaches.

LCA requires two assumptions about the data. The first assumption is the existence of a categorical latent variable *Y* such that the joint distribution $p(\boldsymbol{X}, Y)$ is well defined. We denote the values of *Y* as y_c, c = 1, 2,…, *C*, where *C* is the total number of latent categories. The marginal distribution of the observed data can be written as

A1) $$p(\mathbf{X}) = \sum_{c=1}^{C} p(Y = y_c)\, p(\mathbf{X} \mid Y = y_c).$$

The second assumption is that the items are statistically independent given the latent variable

A2) $$p(\mathbf{X} \mid Y = y_c) = \prod_{j=1}^{J} p(X_j \mid Y = y_c),$$

The usual interpretation of this assumption is that *Y* explains the dependence in the observed data, since the variables X_j are statistically independent after controlling for *Y.* In terms of the observational instruments, the associations among the different components of instruction are explained by the latent classes of instructional practice.

Substitution of equation (2) into equation (1) gives the overall model:

A3)
$$p(\mathbf{X}) = \sum_{c=1}^{C} p(Y = y_c) \prod_{j=1}^{J} p(X_j \mid Y = y_c)$$

The unknown parameters of model are:

1. The number of classes, C;
2. The membership probabilities, $p(Y = y_c)$, which tell us the relative size of the classes;
3. The parameters of the conditional distributions, $p(X_j \mid Y = y_c)$, which depend on the data type (e.g., categorical, continuous).

The strategy for obtaining the unknown parameters is a two-step procedure. First we select a value for *C* and estimate the remaining parameters using, for example, the EM algorithm (e.g. McLaclan & Peel, 2000). Second, we compare various choices of *C*, for example, by using information criteria or classification error (e.g., Lubke & Neale, 2006, 2008). Table A1 reports the Bayesian information criterion, average classification error, and classification entropy for the ELA teachers in our sample.

One of the desirable properties of LCA is that it can be used to classify teachers' response patterns into one of the latent classes. In this paper we refer to this classification as a teacher's "profile membership score." The scoring formula is obtained using Bayes' rule:

4) $$p(Y = y_c \mid \mathbf{X}) = \frac{p(Y = y_c) \prod_{j=1}^{J} p(X_j \mid Y = y_c)}{p(\mathbf{X})}.$$

Equation 4 provides a distribution of profile memberships for each teacher. This is depicted in Figure 2 of the paper.

As a last point we consider the relationship between teachers' profile memberships and a categorization based on teachers' rank ordering on the total score over all items. Table A2 shows the contingency table for the two categorization schemes. It can be concluded that, while the total score quantiles and the profiles are monotonically related, they do not provide an identical classifications of teachers.

Appendix B

Criterion Validity of Teachers' Profile Memberhips

In this paper we considered the validation of teachers' profile membership scores with respect to three student outcomes:

1. Students' achievement, as measured by the SAT-9 open examination.
2. Student engagement, as measured by self-reported in-class effort.
3. Students' social emotional development, as measured by self-reported positive emotional experiences in the classroom.

The SAT-9 examination and the student survey are described in detail by Kane & Staiger (2012). In this research we measured student engagement (self-reported effort) by using responses to the following four survey items.

2a) I have pushed myself hard to completely understand my lessons in this class.

2b) In this class, I stop trying when the work gets hard.

2c) When doing schoolwork for this class, I try to learn as much as I can and I don't worry about how long it takes.

2d) My teacher pushes me to become a better thinker.

Items 2a) through 2d) were selected from a total of seven possible survey questions about classroom effort. They were chosen based on goodness of fit statistics obtained from a factor analysis for categorical data, and the final model had excellent fit to the data (RMSEA = .034, CFI = .99; TLI = .99).

The following three items were chosen from a possible four items to measure social-emotional development (positive emotional experiences).

3a) Being in this class makes me feel angry.

3b) I feel stressed out in this class.

3c) This class is a happy place for me to be.

Items 3a and 3b were reverse coded. This model is just-identified so goodness of fit statistics were not available. Inclusion of the fourth item resulted in poor model fit.

Factor scores were computed for both sets of items using an empirical Bayes estimator. The marginal reliability coefficients were estimated to be .70 for engagement and .71 for social-emotional development. The scores on these two factors, in addition to students' SAT-9 scores, were used as outcomes to assess the validity of teachers' profile membership scores.

The association between teachers' profile membership scores and the three student outcomes were assessed using a fixed-effects regression approach similar to that used in value-added modeling. In particular, for each outcome we estimated the following linear regression model using OLS:

A5)
$$Y_{ij} = \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\gamma}\mathbf{Z}_i + \delta_j + \epsilon_{ij}$$

where

Y_{ij} is the outcome for student *i* with teacher *j*;

$\mathbf{X}_i$ are the district administrative covariates for student *i*;

$\mathbf{Z}_i$ are the 2008 and 2009 Math and ELA State exams of student *i*;

δ_j is the fixed effect for teacher j;

ε_{ij} is a normally distributed error term.

Equation (5) was estimated for all 381 ELA middle school teachers for whom data were available on the observational instruments, and all of these teachers' students with data on the outcome variables and $\mathbf{Z}_i$. The model was estimated separately for each grade-by-district combination, because the administrative covariates were district specific and $\mathbf{Z}_i$ were normed within each grade-by-district combination. The outcome variables were scaled to have a mean of zero and variance within each grade-by-district combination. Teacher effects were coded such that $\hat{\delta}_j = 0$ for an average teacher in each grade-by-district combination. To address sampling error in the $\hat{\delta}_j$, we applied an empirical Bayes' "shrinkage" estimator (e.g., Gelman & Hill, 2006). The mean of the shrunken $\hat{\delta}_j$ estimates within each profile were computed. The results are presented in Figure 3.

Appendix C

Profile Plots for FFT and CLASS

Please include Figures A1 and A2 here.

Appendix D

Consistency of Profile Memberships over Multiple Administrations

In this section we address the reliability of profile scores over different administrations of the observational instruments. Since profile memberships are a categorical, their reliability can conceptualized in terms of classification consistency (Joint Committee on Educational and Psychological Testing, 2014). Here we present two classification consistency coefficients that are commonly used in the literature (e.g., Lee, Brennan, & Wan, 2009): (a) the average proportion of agreements, and (b) the average "chance-corrected" proportion of agreements. We note that existing methods for computing these coefficients are based on models with continuous latent variables, and therefore some innovation was required for the present analysis. Moreover, because our analysis was conducted at the teacher-level, this consideration of within-teacher variability is somewhat ad hoc and the procedures described here leave room for improvement, which we comment on at the end of this appendix.

As noted in the Measures section of this paper, the observational instruments were administered repeatedly to each teacher in the sample. We used equation A4 to compute the profile membership of each administration. We refer to these as the segment-level profile memberships, to be distinguished from the teacher-level profile memberships. For each teacher, the proportion of agreements, denoted as P_i, was computed as the proportion of times that a teacher's segment-level profiles agreed with the teacher-level profile. The average proportion of agreements over all teachers can then be computed as

5) $$\bar{P} = \sum_{i}^{N} P_i/N.$$

Equation (5) does not take into consideration that the segment-level profile memberships might agree with the teacher-level profile memberships "by chance." The probability of chance agreement was computed as the base rate of each profile over all segments, and is denoted P(C_i). For each teacher we can then define the chance-corrected proportion of agreement:

6) $$P_i^* = \frac{P_i - P(C_i)}{1 - P(C_i)}$$

The average of equation (6) over teachers is the average chance-corrected proportion of agreement.

As reported in the Reliability section of this paper, we found that equation (5) was equal to .57 and equation (6) was equal .44 in the present sample. These figures leave much to be desired, but they should also be contextualized by the previous findings on reliability reported in Table 2. Moreover, there are two reasons to suspect that these coefficients underestimate the classification consistency of profile membership scores. First, the LCA model parameters were estimated using teacher-level averages on each item, which are continuous; however, the segment-level scores are categorical. This may have induced misclassifications due to different levels or measurement in the two analyses. Second, the approach described here required that each segment was "hard" classified into a profile membership category. Since not all of these classification were perfectly accurate (cf. Figure 2), the consistency measures reported here are also affected measurement error in the individual profile memberships. In general, methods for computing the classification consistency of segment-level profile memberships is an important area of further research.