Commonality and Uniqueness in Teaching Practice Observation

Catherine McClellan
Clowder Consulting

John Donoghue
Clowder Consulting

Yoon Soo Park
University of Illinois-Chicago

Introduction

Focus on classroom teaching practice has never been greater than in recent years. As part of the commitment states and jurisdictions have made under federal programs such as Race to the Top and the Teacher Incentive Fund, and in seeking waivers from the constraints of *No Child Left Behind*, states and districts have committed to change policies and practices in all aspects of teacher evaluation. A large number of observation instruments have been developed to structure evaluation of the skills of teachers in the classroom and support actionable feedback. These instruments vary in their philosophy of instruction, dimensions of teaching that are valued, and specificity of application to particular content areas or grade levels, among other aspects.

One aspect of this work that presents a challenge is the selection of an appropriate observation instrument for use in a jurisdiction. Despite the differences in approach, there are substantial superficial similarities across many of the most popular tools. Component names and scale descriptions sound quite alike, even to those relatively expert in the content. Most established instruments have data and evidence to support claims of valid use for scores. In addition to the well-known instruments, there are numerous variants and modified versions cropping up throughout the country, each adapted to suit the specific needs of a particular location. There are also instrument being developed from scratch in many localities. Many of these variants and instruments suffer from the usual travails of new assessments: imperfect initial evidence supporting the planned uses and limited resources to complete the studies to provide more. There is both too much and too little information. The potential for suboptimal choices is increased by intense time pressure from grantors or governing agencies to select an instrument and get an evaluation system in place by proposed deadlines.

Despite apparent differences and similarities, it is not clear that the instruments are actually measuring different aspects of teaching practice. If they are, then choices between instruments should be given care and time. If they are not—if they measure the same aspects of teaching in more-or-less the same ways—then the selection is a less-critical factor. But few studies are available that do head-to-head comparisons of multiple teaching practice observation instruments, so comparisons generally must be made on descriptions, inferences, and interpretations.

Data Source: Measures of Effective Teaching Study

Between 2009 and 2012, the Bill & Melinda Gates Foundation funded a large-scale research project; the Measures of Effective Teaching (MET) study (see http://metproject.org/ for more information). More than 3,000 teachers in seven large school districts (Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO.; Hillsborough County, FL.; Memphis, TN; New York City, NY; and Pittsburgh Public Schools, PA.) in the US participated in MET. Pittsburgh served as the project's pilot district, so no data from this district was analyzed in the final data set from the MET study. The MET study had a number of components, including:

- a student perception survey
- a content knowledge for teaching measure

- student achievement measures, including
    - state standardized (AYP) test scores in math and ELA
    - alternative open-ended assessments
- teacher working-conditions survey
- video capture of classroom instruction sessions

It is the data from this last component that will be the main focus of this study. Most teachers in the study were captured on video four times in each of two years teaching their classes, resulting in 8 videos per teacher total. Thousands of hours of classroom video in a core sample were scored by independent raters using as many as three different observation rubrics. All videos in this sample were scored using two content-neutral instruments: CLASS (see http://www.teachstone.org/about-the-class/ for details) and the Framework for Teaching (FfT; see http://www.danielsongroup.org/article.aspx?page=FfTEvaluationInstrument for information). The 2011 version of FfT was developed based on the work done in the MET project, and that version is the one closest to that used in MET. The mathematics classes also were scored using a version of MQI (see http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&tabgroupid=icb.tabgroup120173 for information) referred to as MQI Lite and the English Language Arts (ELA) classes were scored using a version of PLATO (see http://platorubric.stanford.edu/ for details) referred to as PLATO Prime. For self-contained elementary level teachers who provide instruction in both math and ELA, there are scores from all four of these instruments. The scores assigned to the videos on these instruments as part of the MET study serve as the data source for this study.

Scoring Design and Data

In the MET study, each instrument had an associated scoring design applied. As the scoring design defines the available data, details of the design will be briefly described here. There are some notes that apply across instruments:

- Teachers had videos from Year 1 and Year 2 of the MET project, and the videos from Y1 were scored before those from Y2. The scoring was immediately adjacent, however, so as soon as the Y1 videos were exhausted, the Y2 videos began scoring. The division between the academic years of the study was known to the raters in that the site URL changed, as did the scoring software background colors—otherwise, functionality and look-and-feel remained constant throughout scoring.
- In some cases the video capture started before the beginning of the actual classroom instructional session. This was more common in schools where a staff member other than the teacher set up and started the video capture equipment. As there was no way to automatically detect when instruction began, the scoring process was standardized so that all videos scored on an instrument used the same segment timing and started at the beginning of the capture, whether the class had begun or not.
- All instruments had a specified level of double-scoring completed for quality control purposes. A double-score was triggered by the specified rate in the software system: if the rate was 10%, the 11th score assigned would be a second score of the 10th video assigned by a different rater.

Raters were selected to double-score from the pool of available raters working at the point when a double-score was triggered, with constraints on the frequency of rater pairings taken into consideration by the software.

- The overall scoring design was constrained so that a single rater could not assign the primary score to an individual teacher for more than one video across all videos available from that teacher. This constraint was relaxed in the double-scoring, so a rater who had assigned a primary score to an individual teacher was an acceptable selection as a second rater in double scoring.
- All instruments had a specified level of validity scoring completed for quality control purposes, triggered in the same way as the double scoring. Validity scoring is defined as the use of videos that have been pre-scored by an expert rater or master coder and that are inserted blindly into a rater's queue to evaluate scoring accuracy.
- All raters worked in a small team overseen by a scoring team leader. These team leaders were selected for a combination of accurate scoring skills and appropriate interpersonal skills. There were no experienced raters available at the beginning of the MET project, as all raters were newly certified. As experienced raters became available during the project, some were selected for promotion to team leader if their scoring accuracy and interpersonal skills were appropriate.
- All instruments utilized a system of back scoring by team leaders for quality control purposes. Back scoring is defined as the re-scoring of videos scored by raters within the leader's team, either blind or with knowledge of the rater score, for quality control purposes. Aberrant or disagreed scores were discussed with the rater in a joint review process.
- Scoring team leaders also provided primary scores for videos that were deferred by raters. Videos could be deferred because of technical issues (the primary score the team lead assigned could be "unscorable"), because the rater recognized the teacher in the video, or because the rater had questions about how to score the video. In the latter case, the team lead could elect to review the video with the rater in a joint review process as a training activity.
- More than 900 raters in total were trained and certified to score video on the MET study across the instruments. The majority of these raters scores CLASS and FfT, but all of the instruments examined in this study had 100 or more raters contributing to the scored data set.

CLASS

There were 5,940 videos scored in the Y1 data set on CLASS; after deletion of incomplete and problematic cases, 5,748 cases were used in Y1 analysis. There were 6,297 videos scored in the Y2 data set on CLASS; after deletion of incomplete and problematic cases, 6,252 cases were used in Y2 analysis. CLASS was scored using the Upper Elementary tool for classes in grades 4-6 and the Secondary tool for classes in grades 7-9. Most raters were certified to score on only one of these CLASS tools. For each video, only the first 30 minutes of the classroom session were scored, starting at the beginning of the video capture and ending at minute 30. All 12 dimensions and 4 domains were scored by each rater for each video segment in the main scoring. All CLASS dimensions have 7 score levels. CLASS was scored in two separately timed segments, the first from the beginning of the video until minute 15, and the second from minute 15 until minute 30. The two segments were scored independently, in that there was no effort to assign the same rater to score segment 1 and segment 2 from the same video. In fact, it was quite unlikely that the same rater would score both segments by chance, as the scheduling of raters to work shifts and CLASS video segments was complex. Since two segments were scored on CLASS, there are two sets of scores available for each video, one set for each time segment.

Framework for Teaching (FfT)

There were 7,484 videos scored in the Y1 data set on FfT; after deletion of incomplete and problematic cases, 7,439 cases were used in Y1 analysis. There were 6,294 videos scored in the Y2 data set on FfT; after deletion of incomplete and problematic cases, 6,294 cases were used in Y2 analysis. FfT does not have separate tools by grade level, so all raters were trained and certified on the same version of the instrument. For each video, minutes 0 through 15 and then minutes 25 through 35 were viewed by the rater—the scoring software system was set up so that the minutes from 15 to 25 were skipped and not seen by the rater; a screen indicating the change in time point was shown briefly to minimize rater confusion. One set of scores was assigned for the combined video segments (a total of 25 minutes of video). Eight of the ten traditional components of FfT in Domains 2 and 3 were used in MET. In Domain 2: Classroom Environment, the following components were scored:

- 2a) Creating an environment of respect and rapport
- 2b) Establishing a culture for learning
- 2c) Managing classroom procedures
- 2d) Managing student behavior

In Domain 3: Instruction, the following components were scored:

- 3a) Communicating with students
- 3b) Questioning and discussion techniques
- 3c) Engaging students in learning
- 3d) Using assessment in instruction

Components 2e) Organizing physical space and 3e) Demonstrating flexibility and responsiveness were not used in the MET project. All FfT components have four score levels. One set of FfT scores is available for each video.

MQI Lite

There were 3,414 videos scored in the Y1 data set on MQI Lite; after deletion of incomplete and problematic cases, 2,607 cases were used in Y1 analysis. There were 2,991 videos scored in the Y2 data set on MQI Lite; after deletion of incomplete and problematic cases, 2,986 cases were used in Y2 analysis. Because of the importance of the focus of the board camera (so that the accuracy of work could be judged), mathematics class videos from Y1 were deemed unscorable on MQI Lite more frequently than on the content-neutral instruments. Quality improvements made in the Y2 video capture greatly reduced this issue. MQI Lite was scored in 7.5-minute segments, the shortest segmentation in the MET study. Four segments were scored in each video: minutes 0-7.5; 7.5-15; 15-22.5; and 22.5-30. Six dimensions were scored on the version of MQI Lite used in the MET project:

- Richness of the Mathematics (RI)
- Errors and Imprecision (E&I)
- Working with Students and Mathematics (WWSM)
- Student Participation in Meaning-Making and Reasoning (SPMMR)
- Classroom Work is Connected to Mathematics (CWCM)
- Explicitness and Thoroughness (E&T)

In addition to these dimensions, two scores were assigned to each 30-minute video (*not* each segment) by the rater: an Overall Mathematical Quality of Instruction (MQI) and an Overall Guess at Mathematical Knowledge for Teaching (MKT)

All scores on all dimensions were assigned by a single rater for each video. All dimensions were scored on a 3-level score scale, except for Classroom Work is Connected to Mathematics (CWCM) which is scored Yes/No. MQI Lite does not have separate versions of the tool based on grade levels for most dimensions, so the raters were certified to score on the same version of the instrument. There is one exception to this rule, however: Explicitness and Thoroughness (E&T), which was scored only on classes with algebra content. E&T was scored interactively with another scale, Classroom Work is Connected to Mathematics (CWCM), in that if E&T was scored (the class has algebra content), then CWCM was *not* scored. The intended result was that each video segment should have a valid score on either E&T <u>*or*</u> CWCM and have a "N/A" score on the other. Each of the four time segments of video should have five valid scale scores (on the first four dimensions in the list above, plus either CWCM or E&T) and one "N/A" score. Raters were also asked to assign a holistic score across the 30 minutes of video for each dimension, and to assign the overall MQI and MKT scores.

PLATO Prime

There were 3,652 videos scored in the Y1 data set on PLATO Prime; after deletion of incomplete and problematic cases, 2,924 cases were used in Y1 analysis. There were 1,919 videos scored in the Y2 data set on PLATO Prime; after deletion of incomplete and problematic cases, 1,910 cases were used in Y2 analysis. Because of the importance of the details classroom discussions of content, ELA class videos from Y1 were deemed unscorable on PLATO Prime more frequently than on the content-neutral instruments. Improvements in audio capture in Y2 reduced this issue. PLATO Prime was scored in 15-minute segments. The first 30 minutes of the class were scored, in two segments: one from minutes 0-15 and one from minutes 15-30. PLATO Prime did not have different versions of the instrument for different grade levels, so all raters were trained and certified on the same version of the instrument. Six elements of instruction were scored on PLATO Prime:

- Intellectual Challenge
- Classroom Discourse
- Modeling

- Strategy Use and Instruction
- Time Management
- Behavior Management

All elements were scored by the raters on a 4-level score scale. On all elements, two score categories were collapsed for MET analysis purposes in the software, as the raters had difficulty distinguishing between them in practice. These data structures were: Modeling, Strategy Use and Instruction, and Time Management, for which score levels 1 and 2 were collapsed; and Intellectual Challenge, Classroom Discourse, and Behavior Management, for which score levels 3 and 4 were collapsed. The full 4-level scores were recorded and were used herein. PLATO Prime also had a number of content indicator scales (Reading, Writing, Literature, etc.) that were scored Yes/N/A depending on the content of the particular time segment, as well as an overall content representation scale scored +/-. Only the element scores were used in the analysis in this study. All scores on all elements were assigned by a single rater for each video. Since two segments were scored on PLATO Prime, there are two sets of scores available for each video, one set for each segment.

Research Questions

The broadly-stated research question is:

- What are the statistically unique aspects of each teacher practice observation instrument and what do they have in common?

In order to examine this question, we will break this down into some more specific questions:

1. What is the factor structure of each individual instrument used in the MET study?
2. What is the factor structure when the two content-neutral instruments (CLASS and FfT) are analyzed together?
3. What is the factor structure when the content-neutral and content-specific instruments are analyzed together? Specifically:
   3.1. What is the factor structure when CLASS, FfT and MQI Lite are analyzed together?
   3.2. What is the factor structure when CLASS, FfT and PLATO Prime are analyzed together?

Research question 1 will support insight into whether the factor structure of an instrument is altered when it is analyzed with data from one or more additional instruments in research questions 2 and 3.

Data Analyses

The data produced from teaching practice observation instruments are ordinal. There is some disagreement about the nature of the construct being measured, but it is generally thought that the instruments capture something most commonly referred to as “teaching effectiveness” and that this underlying latent trait is continuous and approximately normally distributed in the full population of teachers. Given ordinal data, the factor analyses were completed on the polychoric correlation matrices (Bartholomew, 1980; Jöreskog & Moustaki, 2001; Mislevy, 1986). Analyses were conducted using the R statistical software (see http://www.r-project.org/ for more information on R). In each case, varimax and promax rotations were applied to simplify the resulting data structure. The promax rotation results will be reported, as there is no reason to believe that the factors were uncorrelated and every reason to believe that they were correlated. The factor inter-correlations were reported. The factor analyses were closest to exploratory in nature, despite the apparent structure of the instruments. Variations or modifications of the observation instruments were used in many cases. None of these instruments had completed data collection via 360° video capture nor scored observation sessions in a large-scale online distributed design such as that used in MET. These novel aspects suggested that an exploratory approach might be appropriate, while keeping in mind the instruments' design. The analysis approach could be considered mixed: exploratory in Y1 data and confirmatory in Y2 data. In addition, when the instrument data are combined, it is not clear what structure one would be confirming if CFA were preferred.

The solutions selected from the factor analysis were examined for evidence of convergent and discriminant validity as well as interpretability in terms of the instrument and scales (Cronbach & Meehl, 1955). For factor analytic results, evidence of convergent validity generally was provided by showing the variables within a single factor were highly correlated through high loadings on a single common factor. As a rule of thumb, if cross-loadings of scales onto multiple factors exist in the results, the cross-loadings should be 0.2 smaller than the primary loading. Evidence of discriminant validity generally was provided by determining if the resulting factors were distinct and relatively uncorrelated. This was supported using the factor correlation matrix. It is preferable that inter-correlations between factors should not exceed 0.7, as a correlation greater than 0.7 indicates a majority of shared variance. Factors with inter-correlations greater than 0.7 may not represent distinct constructs.

Results: Single Instrument Analyses

In this first section, the results of the analyses within each instrument will be presented. Scree plots were examined for guidance about the number of apparent factors in the data. As the raters were aware of the change between Y1 and Y2 in the scoring, in most cases the data sets were analyzed separately to evaluate possible differences in scoring that may have altered the factor structure.

Single-Instrument Factor Analysis Results: CLASS

The current structure of the CLASS tools used in the MET study is shown in Table 1, indicating which dimensions are grouped together into domains. There are changes from the structure at the time of the MET study and data collection. Then, Negative Climate was part of Emotional Support and Instructional Learning Formats was part of Instructional Support, as indicated in the table note. There is also a name change to one dimension: Analysis and Problem Solving is now called Analysis and Inquiry, although the domain alignment has remained the same. Note that in analysis of CLASS data, Negative Climate is typically reverse-coded (high numeric scores on Negative Climate are "bad", whereas high scores on all other scales are "good"). We did not reverse the scale for Negative Climate in these analyses so as to reduce the possibility of introducing errors. Thus it is expected that in the results Negative Climate may have negative factor loadings.

Table 1: CLASS Dimensions and Domains[1]

Age/Grade Level	Emotional Support	Classroom Organization	Instructional Support	Student Engagement
Upper Elementary **Grades 4-6**	• Positive Climate • Teacher Sensitivity • Regard for Student Perspectives	• Behavior Management • Productivity • Negative Climate*	• Instructional Learning Formats* • Content Understanding • Analysis and Inquiry** • Quality of Feedback • Instructional Dialogue	Student Engagement
Secondary **Grades 7-12**	• Positive Climate • Teacher Sensitivity • Regard for Adolescent Perspectives	• Behavior Management • Productivity • Negative Climate*	• Instructional Learning Formats* • Content Understanding • Analysis and Inquiry** • Quality of Feedback • Instructional Dialogue	Student Engagement

*Classification Changes: Negative Climate (formerly in Emotional Support) and Instructional Learning Formats (formerly in Classroom Organization) have changed domains. Negative Climate is the third dimension in Classroom Organization; ILF is the first dimension in Instructional Support.
**Dimension Change: Analysis and Problem Solving has been renamed Analysis and Inquiry.

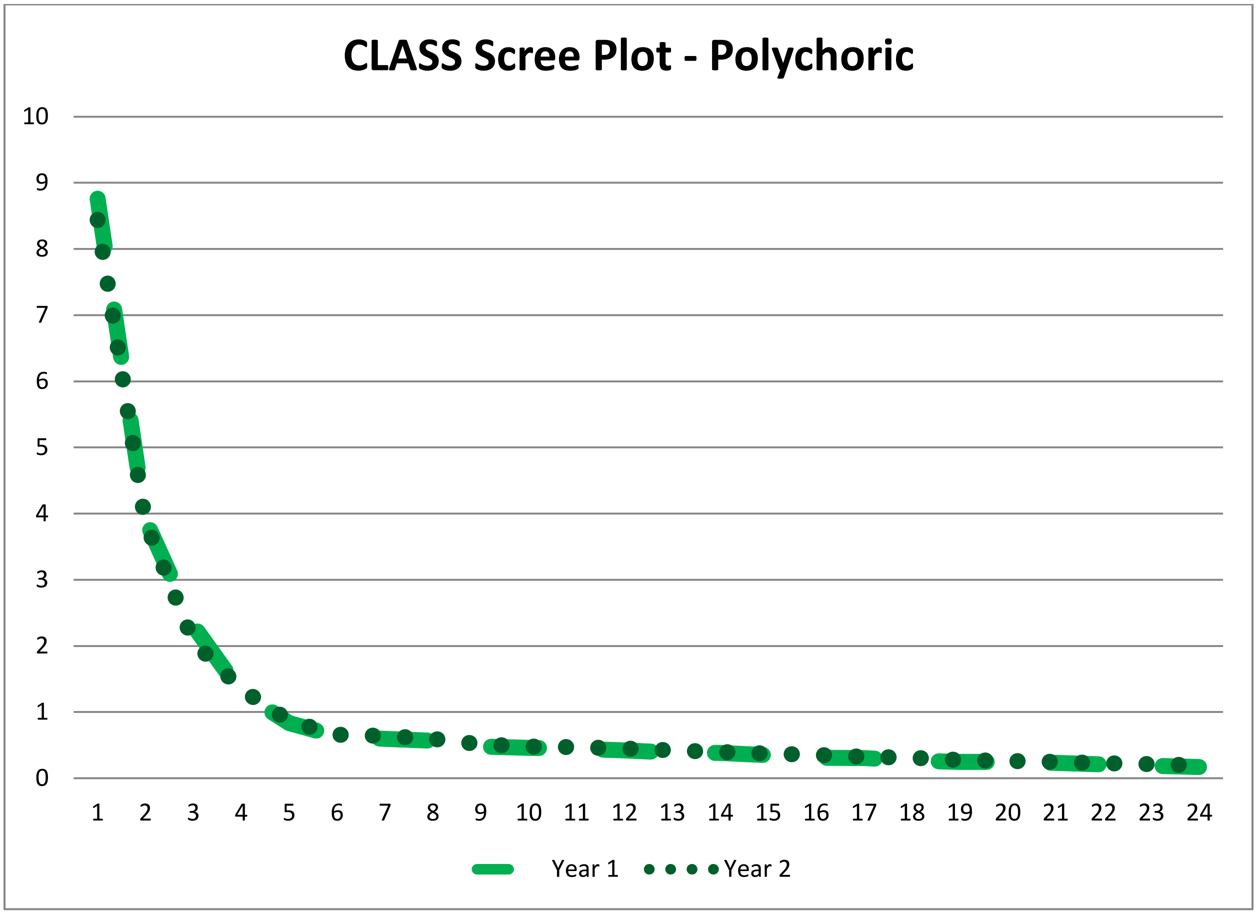
The scree plot for CLASS is presented in Figure 1. The Y1 and Y2 data were plotted together: Year 1 is shown in the dashed line, Year 2 in the dotted line. It seems that the two years' data have very similar factor structure, as the lines are nearly overlaid. The scree plot does not have a very clear "elbow", although it appears to bend between 4 and 6 factors. The eigenvalue at value 4 is just above 1.0 in both years; all eigenvalues below there are less than 1.0 in value. The scree is pretty linear below 8 factors, but there is not a clearly defined cutoff in the plot.

There was an unexpected variable that proved to be interesting that will be explored herein as part of the factor structure: the time segment of the video. All solutions up to an 8-factor were investigated. Eight was chosen as the maximum using the hypothesis that there could be four factors as defined by the domains of the instrument, either as shown in Table 1 or those at the time of the study, crossed with the two possible video time-segment variable. The full set of factor loadings for each solution is provided in Appendix A (note that the appendices are in Excel; each Appendix is a tab in the file). The chart was color-coded so that dimensions shown in rows shaded the same color indicate those belonging to the same domain at the time of the study. To assist in interpretation, for each dimension the primary loading was boxed in yellow and the factors were ordered roughly consistently across each

[1] Table adapted from http://www.teachstone.org/about-the-class/class-organization/.

analysis. It was notable that the data are quite consistent in loading size and structure when compared across the two video time segments.

Figure 1: Scree Plot for CLASS MET Data: Year 1 and Year 2

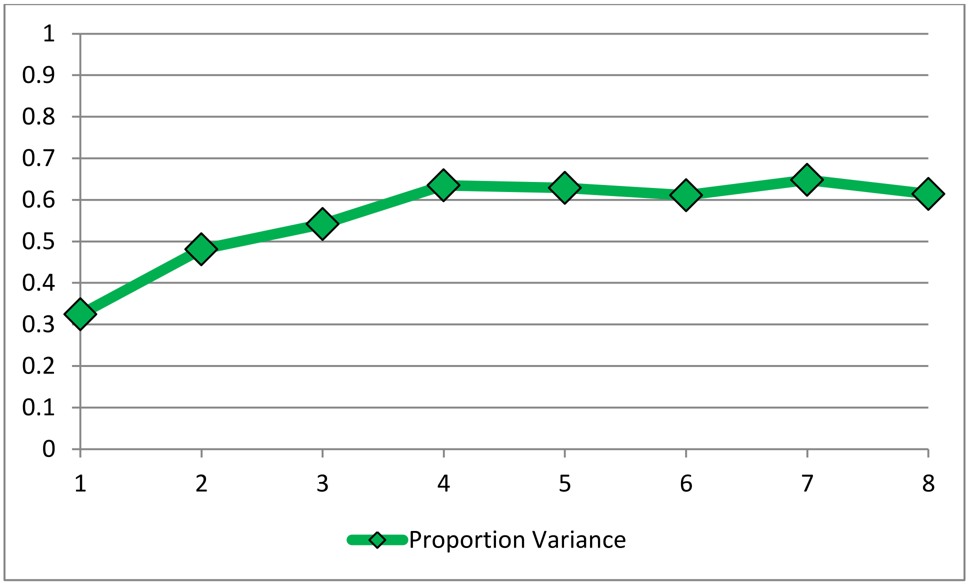


When the segment timing was included as a variable in the factor analysis, the data separated very clearly into a two-factor solution with the time-segment indicator variable. This unexpected finding was the reason for inclusion of the video time segment, rather than the year of the data (the Y2 and Y2 data were very similar), as an analysis variable. The two-factor solution clearly broke along the video time-segment lines, with all dimensions in the first segment forming factor 1 and all dimensions in the second segment forming factor 2. A pattern of the solutions with odd numbers of factors being less interpretable than the subsequent solution with an even number of factors was seen consistently, so solutions with even numbers of factors are discussed in the text. This finding also bolsters the hypothesis that the two time segments were playing an important role in the factor structure of these data.

The 4-factor solution was consistent with the move of dimension Negative Climate in domain Classroom Organization noted above—for both time segment 1 and 2, those three dimensions loaded together onto factor 3 (for time segment 1) and factor 4 (for time segment 2), with all other factors remaining loaded together in factors 1 and 2 for time segments 1 and 2 respectively. Student Engagement showed some tendency toward cross-factor loading in this solution (still within time segment), a tendency that continued as more-complex solutions were examined. The 6-factor solution was the same as the 4-factor solution, except that Positive Climate moved to a factor alone, more clearly in time segment 2 than time segment 1. The 7- and 8-factor solutions did not lend themselves to clear interpretations. The 7th factor had no dimension primarily loaded, and the 8th factor had a couple of dimensions with cross loadings.

None of the solutions were highly explanatory of the data variability. The 1-factor solution accounted for only about 32% of the variance. The 2-factor solution accounted for about 48% of the variance. Increasing to the 8-factor solution improved the variance accounted for only to about 62%. The proportion of variance accounted for in each model is shown in Figure 2. The greatest increase in variance accounted for was between 1- and 2-factor models; values leveled off at the 4-factor model.

Figure 2: CLASS Proportion Variance Accounted For in FA Models



The factor structure that was most explicable in terms of the domains defined by CLASS was probably the 4-factor model, or more accurately, a 2-factor model within each of the two time segments. In the 4-factor solution, the Classroom Organization domain (as defined more recently, not as defined at the time of data collection) was a distinct factor in each time segment. The other dimensions tended to remain clustered together in a single factor within time segment.

Table 2: CLASS Factor Inter-Correlations

Y1	MR2	MR3	MR1	MR4		**Y2**	MR3	MR2	MR1	MR4
MR2	1	0.39	0.23	0.5		MR3	1	0.37	0.51	0.27
MR3	0.39	1	0.47	0.29		MR2	0.37	1	0.3	0.51
MR1	0.23	0.47	1	0.43		MR1	0.51	0.3	1	0.4
MR4	0.5	0.29	0.43	1		MR4	0.27	0.51	0.4	1

The CLASS factor inter-correlations are shown in Table 2 for the 4-factor solution. In inter-correlation tables in this paper, values less than 0.2 in absolute value are shown shaded in green; values greater than 0.7 in absolute value are shown shaded in red; the diagonal is shaded in gray for ease in reading the table. Most of these factors retained moderate associations and there was statistical variance unaccounted for, but the factor inter-correlations did not necessarily indicate that the factors were not distinct. The correlation values greater than 0.5 in Table 2 were between the factors that have the same dimensions across the time segments (i.e. the Classroom Organization factor in time segment 1 and time segment 2).

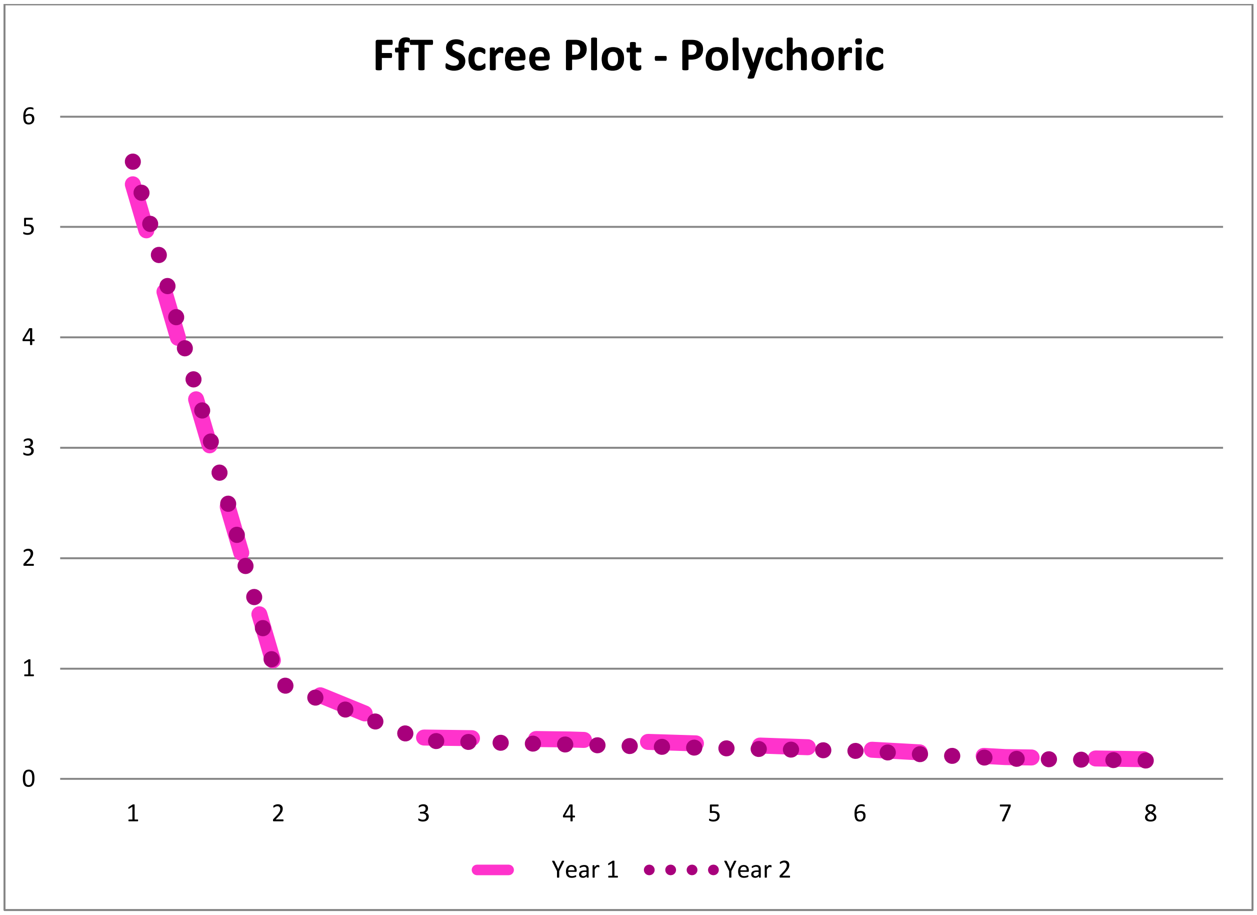
Single-Instrument Factor Analysis Results: Framework for Teaching (FfT)

The scree plot for FfT is presented in Figure 3. The Y1 and Y2 data were plotted together: Year 1 is shown in the dashed line, Year 2 in the dotted line. Again, it seems that the two years' data had very similar factor structure, as the lines were nearly overlaid. There were two apparent locations for the "elbow" in this plot, at either 2 or 3 factors. Two factors seemed more probable, both in that the 3rd eigenvalue was much smaller than 1.0 and that the instrument has two domains. The full set of factor loadings for each solution is provided in Appendix B for solutions from 1 to 4 factors. The chart was color-coded so that components shown in rows shaded the same color indicated those belonging to the same domain. To assist in interpretation, for each component the largest loading was boxed in and the factors were ordered roughly consistently across each analysis.

Neither the 3- nor the 4-factor solutions were easily interpretable in the FfT data, especially if data from both Y1 and Y2 were considered. In the 3-factor solution, Culture for Learning loaded on the 3rd factor alone, but only in Y1—there was no component with a primary loading on the 3rd factor in the Y2 data. Similarly, in the 4-factor solution, while Classroom Procedures loaded on the 3rd factor in both years' data, the 4th factor was not stable. Culture for Learning loaded on the 4th factor and Engaging Students in Learning was split across factors 2 and 4 in the Y1 data, neither patterns appeared in the Y2 data where no component had primary loading on the 4th factor. The 2-factor solution, supported by the scree plot above, was loaded consistently in the factor structure as well, and may be the best solution.

Despite the 2-factor solution appearing the best choice, the components did not all line up within the domain as specified in the instrument in the two factors. Domain 3: Instruction hung together well in a single factor, but Culture for Learning from Domain 2: Classroom Environment loaded very strongly on the same factor. The remaining components in Domain 2 loaded together on a single factor. This suggested that, at least in the MET data, Culture for Learning was statistically more similar to the Instruction components than the Classroom Environment components.

Figure 3: Scree Plot for FfT MET Data: Year 1 and Year 2



The proportion of variance accounted for in each model between 1 and 4 factors for FfT is shown in Figure 4. The two years of data were shown separately, as there were small differences in the model results. Most of the FfT factor analysis solutions explained acceptable levels of the data variability. The 2-factor solution accounted for about 70% of the variance. The 3-factor solution accounted for about 70% of the variance in Y1 but nearly 80% in Y2. Increasing to the 4-factor solution saw a drop in the variance accounted for in both Y1 and Y2, to about 68% and 75% respectively. Despite the increase in variance accounted for in the Y2 data between the 2- and 3-factor models, the basic factor structure was the same, as no component had a primary loading on the 3rd factor in the Y2 data.

Figure 4: FfT Proportion Variance Accounted For in FA Models

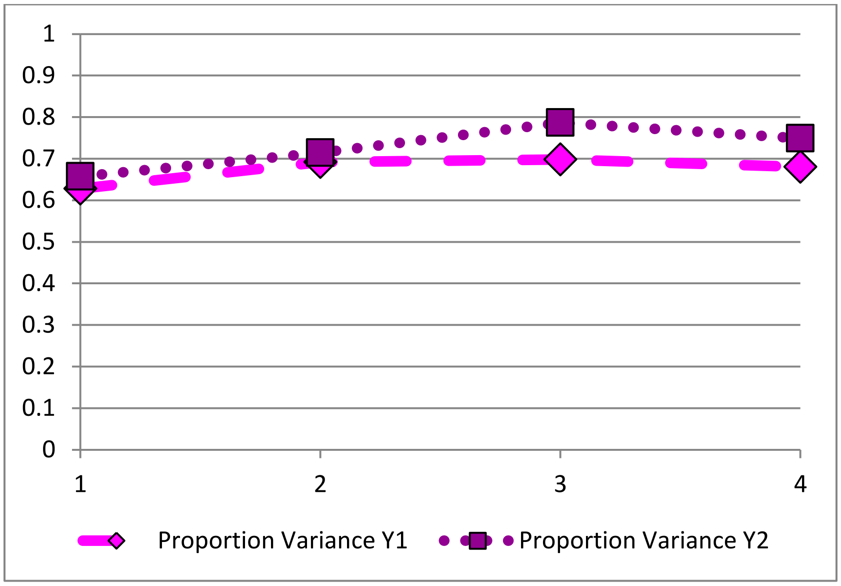


Table 3: FfT Factor Inter-Correlations

Y1	MR2	MR1		**Y2**	MR2	MR1
MR2	1	0.72		MR2	1	0.74
MR1	0.72	1		MR1	0.74	1

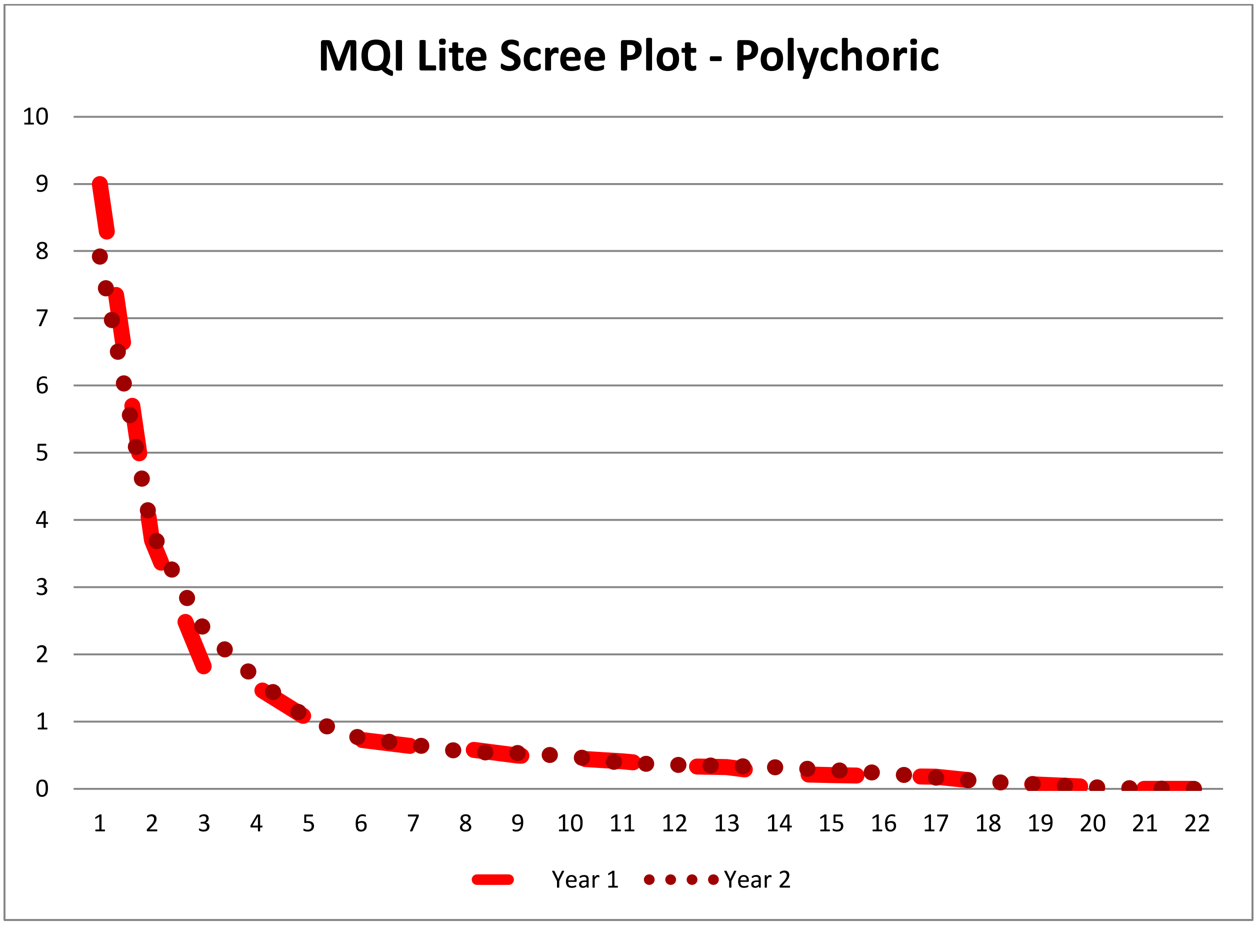
The factor inter-correlations for the 2-factor FfT solution for Y1 and Y2 are shown in Table 3. Despite satisfactory factor loadings and an interpretable structure, the two FfT factors may not be structurally distinct, given the large factor correlation values—FfT actually may have only one underlying factor.

Single-Instrument Factor Analysis Results: MQI Lite

The scree plot for MQI Lite is presented in Figure 5. The Y1 and Y2 data were plotted together: Year 1 is shown in the dashed line, Year 2 in the dotted line. Again, it seems that the two years' data had very similar factor structure, as the lines were nearly overlaid. The "elbow" was not entirely clear, but the most probable bend was at 4 factors. This was also the point about which the factor values crossed below 1.0, as the 5th factor had an eigenvalue of almost exactly 1. MQI Lite scores for all time segments within a video were assigned by the same rater, and the segment time factor did not appear to have an effect on the factor structure of the data. CWCM and E&T were not included in the factor analysis due to the complex missing-data structure these two dimensions have.

All solutions up to a 6-factor were investigated. Six was chosen as the maximum using the hypothesis that there could be one factor per dimension as defined by the instrument for a total of four (excluding CWCM and E&T), plus the overall MQI and the overall MKT. The full set of factor loadings for each solution is provided in Appendix C. The chart was color-coded so that dimensions shown in rows shaded the same color indicated those belonging to the same domain. To assist in interpretation, for each dimension the largest loading was boxed in red and the factors were ordered roughly consistently across each analysis. For the most part, the data were consistent in loading size and structure when compared across the two years of data. The most notable exception was in the 3-factor solution. The Y1 data was more muddled than the Y2 data on SPMMR; in Y2 SPMMR clearly loaded onto a factor by itself, while in the Y1 data SPMMR loaded on two factors. This may reflect increased proficiency in instrument use as the experience level in the rater pool increased through the project.

Figure 5: Scree Plot for MQI Lite MET Data: Year 1 and Year 2



The most interpretable result was the 4-factor structure. In this solution, the scores from all time segments plus the holistic score on each dimension loaded onto a single factor, and there were no cross-loadings. Of interest were the results for the overall MQI and MKT scores. Both had a moderate positive factor loading on the same factor as RI and a negative loading of about the same magnitude on the same factor as E&I, as well as small positive loadings on the same factor as WWSM. The negative loading for E&I was to be expected, as it was a reverse-coded scale (high scores are “bad”). Solutions with more than 4 factors seemed to draw out some sort of temporal factor, in that the additional factors loaded almost exclusively on the first time segment of the dimensions.

The proportion of variance accounted for in each model between 2 and 6 factors for MQI Lite is shown in Figure 6. The two years of data were shown separately, as there were some differences in the model results. Many of the MQI Lite factor analysis solutions explained acceptable levels of the data variability. The 2-factor solution accounted for only about 55% of the variance. The 4-factor solution accounted for about 72% of the variance. Increasing to the 6-factor solution augmented the variance accounted for only to about 76%. The greatest increase in variance accounted for was between 2- and 4-factor models, and the values leveled off at the 4-factor model. This added support to the selection of a 4-factor model as a reasonable choice for these data.

Figure 6: MQI Lite Proportion Variance Accounted For in FA Models

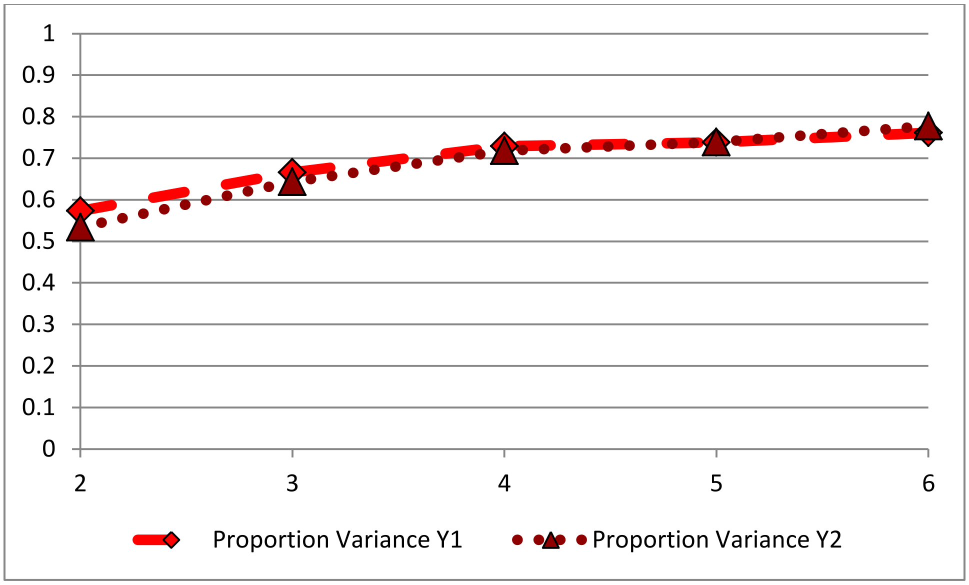


Table 4: MQI Lite Factor Inter-Correlations

Y1	MR1	MR2	MR3	MR4		**Y2**	MR2	MR3	MR1	MR4
MR1	1	-0.36	0.56	0.56		MR2	1	-0.27	-0.26	-0.08
MR2	-0.36	1	-0.26	-0.14		MR3	-0.27	1	0.46	0.41
MR3	0.56	-0.26	1	0.55		MR1	-0.26	0.46	1	0.52
MR4	0.56	-0.14	0.55	1		MR4	-0.08	0.41	0.52	1

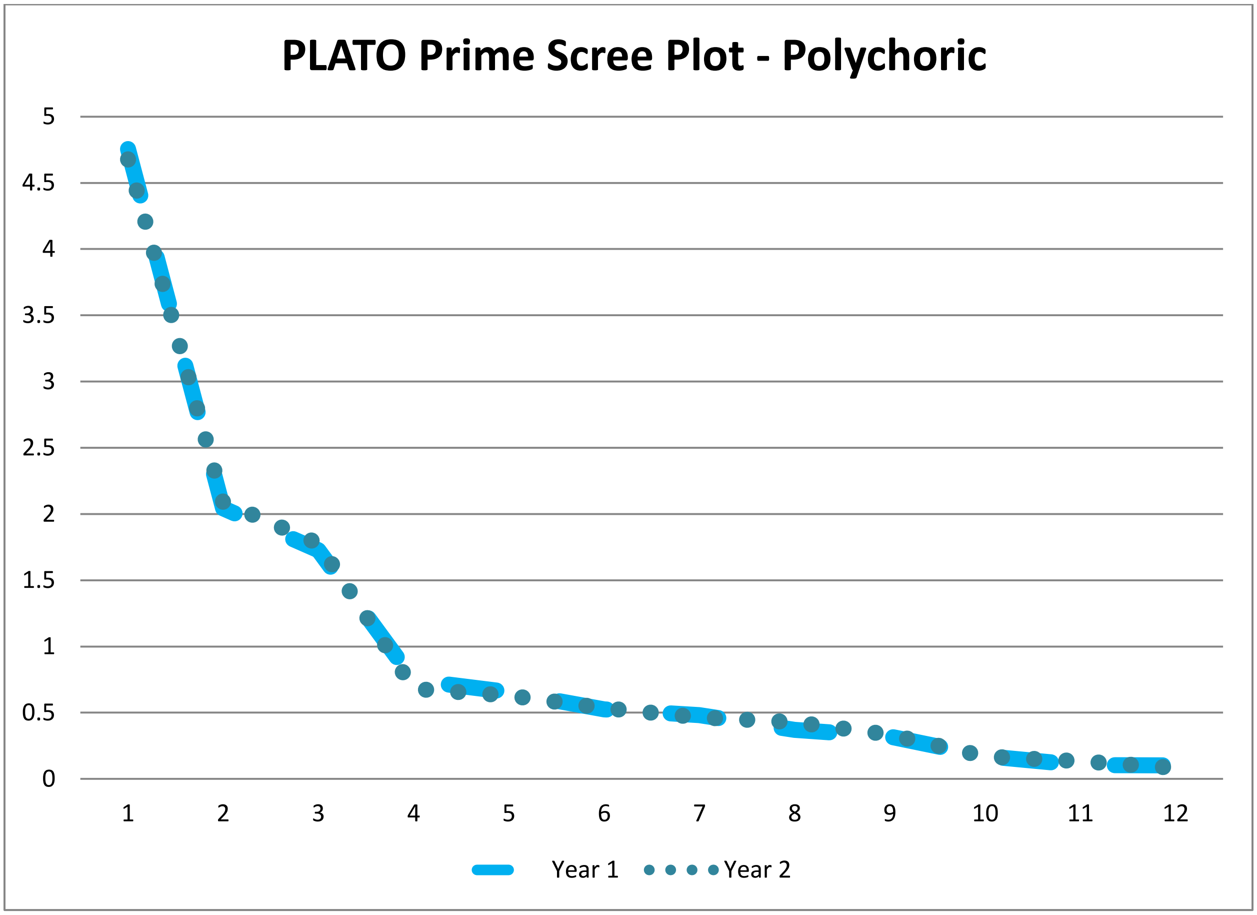
The correlation between E&I and SPMMR was the only one smaller than 0.2 in magnitude. The other values indicated moderate correlation between the factors. All inter-correlations were smaller than 0.7, so the solution may have identified distinct constructs in the data.

Single-Instrument Factor Analysis Results: PLATO Prime

The scree plot for PLATO Prime is presented in Figure 7. The Y1 and Y2 data were plotted together: Year 1 is shown in the dashed line, Year 2 in the dotted line. Again, it seemed that the two years' data had very similar factor structure, as the lines were nearly overlaid. The "elbow" was not entirely clear, as there were two points where the plot apparently bent, at 2 and at 4 factors. The factor values crossed below 1.0 between factors 3 and 4. PLATO Prime scores for all time segments within a video were assigned by the same rater, and the segment time factor did not appear to have an effect on the factor structure of the data. The full set of factor loadings for each solution is provided in Appendix D for solutions from 1 to 6 factors. The chart was color-coded so that components shown in rows shaded the same color indicated those belonging to the same domain. To assist in interpretation, for each component the largest loading was boxed in and the factors were ordered roughly consistently across each analysis.

In spite of the suggestion from the scree plot of a 2- or 4-factor solution, in terms of the instrument it was clear that the 3-factor solution was the most attractive. Although the elements scored in the MET study were not clustered in PLATO or PLATO Prime, there were pairs of elements in the 3-factor solution in data from both Y1 and Y2 that had a strong affinity for each other: Intellectual Challenge and Classroom Discourse; Modeling and Strategy Use and Instruction; and Time Management and Behavior Management. Each pair seemed to have a superficial similarity: the first in classroom instruction and interactions; the second in content approaches; and the third in classroom administration. The factor loadings for Time Management were the smallest of the set, but the clusters were each well-defined. The 4- to 6-factor solutions were somewhat inconsistent across the two years' data and more difficult to interpret as a result.

Figure 7: Scree Plot for PLATO Prime MET Data: Year 1 and Year 2



The proportion of variance accounted for in each model between 1 and 6 factors for PLATO Prime is shown in Figure 8. The two years of data were shown separately, as there are some differences in the model results. A 1-factor solution accounted for only about 34% of the variance. The 3-factor solution accounted for about 63% of the variance. Increasing to the 6-factor solution augmented the variance accounted for to about 76%; both the 5- and 6-factor solutions accounted for substantially more variance than the more-interpretable 3-factor solution. Despite the increased variance accounted for by increasing the number of factors, those structures were inconsistent across the two years' data and difficult to decode.

Figure 8: PLATO Prime Proportion Variance Accounted For in FA Models

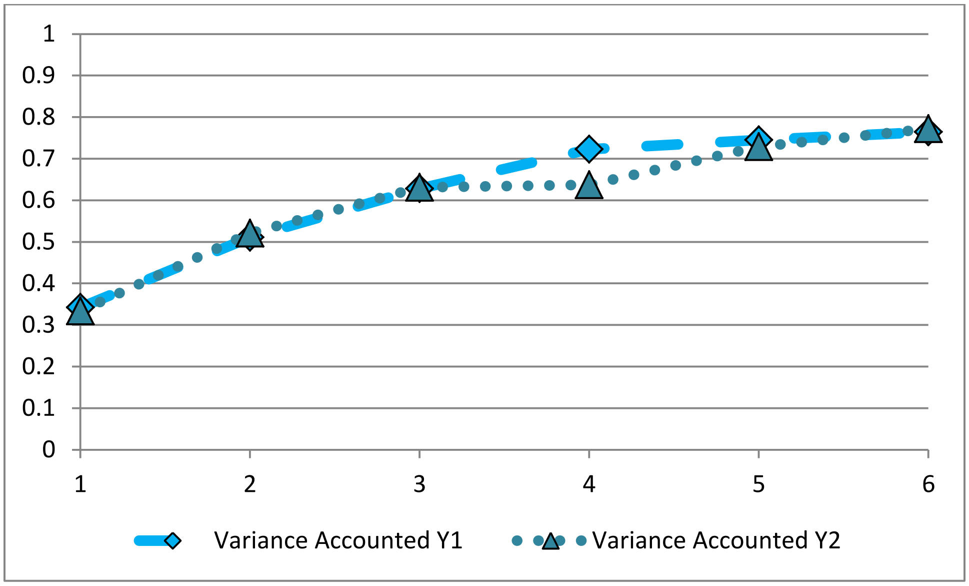


Table 5: PLATO Prime Factor Inter-Correlations

Y1	MR1	MR2	MR3		**Y2**	MR1	MR3	MR2
MR1	1	0.35	0.43		MR1	1	0.47	0.33
MR2	0.35	1	0.45		MR3	0.47	1	0.36
MR3	0.43	0.45	1		MR2	0.33	0.36	1

All values indicated moderate correlation between the factors. All inter-correlations were smaller than 0.7, so the solution may have identified distinct constructs in the data.

Results: Combined Instrument Analyses

In this next section, the results of the analyses combining data from various instruments are presented. As before, scree plots were examined for guidance about the number of apparent factors in the data. As the raters were aware of the change between Y1 and Y2 in the scoring, the data sets were analyzed separately to evaluate possible differences in scoring that may have altered the factor structure.

There are some aspects of the data sources that are important to keep in mind when considering the results. For each set of results, the scores were assigned to the exact same videos. If the coding had occurred in a live classroom, it would be the equivalent of the following:

- For CLASS and FfT combined: At the beginning of the classroom instructional session, a trained CLASS observer and a trained FfT observer were both present and scoring independently. At the end of the first 15 minutes, the original CLASS observer was (magically) replaced with a different trained CLASS observer for the next 15 minutes (the two time segments were scored by different raters); this rater would disappear at minute 30. Each CLASS rater recorded a complete set of scores. Between minutes 15 and 25 of the class, the FfT rater (also magically) disappeared, reappearing at minute 25 to stay through minute 35, recording a single set of scores and then disappearing.

- For CLASS, FfT, and MQI Lite, in the math classes, all the above conditions apply, plus there was a trained MQI observer in the room as well. The MQI Lite rater was present during the same 30 minutes as the two CLASS raters. During this period the MQI Lite rater recorded 4 sets of scores plus the overall MQI and MKT scores.
- For CLASS, FfT, and PLATO Prime, in the ELA classes, the above conditions for the CLASS/FfT combination apply, plus there was a trained PLATO Prime observer in the room as well. The PLATO Prime rater was present during the same 30 minutes as the CLASS raters. During this period the PLATO Prime rater recorded 2 sets of scores plus a number of content codes.

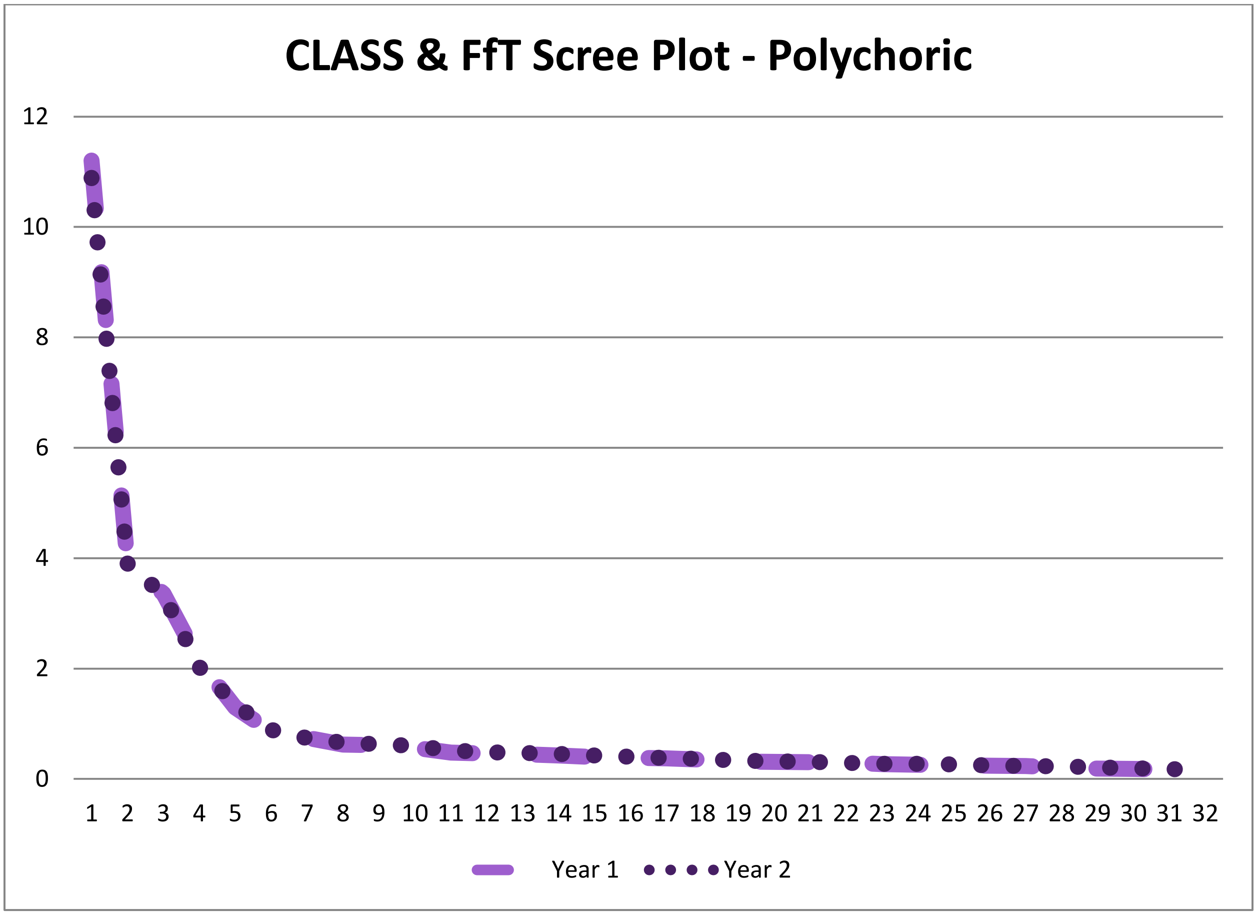
This series of events obviously would be very disruptive in a live classroom, what with multiple raters popping in and out, but worked quite well using the videos and scoring software platform. The most important fact to keep in mind though, is that all of the raters within a combined data set were observing the same teaching practices—just through the lens of the different instruments. The question examined in this section is whether the different lenses enabled them to see different things or not.

Combined-Instrument Factor Analysis Results: CLASS and FfT

There were 5,709 cases used in Y1 analysis and 6,251 cases used in Y2 analysis of the combined CLASS/FfT data set. The scree plot for CLASS and FfT combined is presented in Figure 9. Recall that we accepted a 4-factor solution for CLASS as most probable and interpretable (actually the same 2-factor solution within each of the two time segments) and a 2-factor solution for FfT. If the factor structures remain relatively intact within each instrument, we might expect to see a 6-factor solution. If these two instruments have some “common” factors, we might expect to see a 2- or 4-factor solution dominate, depending on whether CLASS’ time-segment variable comes through in this larger analysis.

Given that both CLASS and FfT had minimal to no differences between Y1 and Y2 in the scree plots, it was unsurprising that the combined data set also showed apparently the same structure in the combined scree plot. From the scree plot, a 2- factor solution suggested itself, as did a 6-factor solution; the 6-factor solution was bolstered somewhat by the fact that the 6th factor was the point at which the eigenvalues crossed below 1.0.

Figure 9: Scree Plot for CLASS and FfT MET Data Combined: Year 1 and Year 2



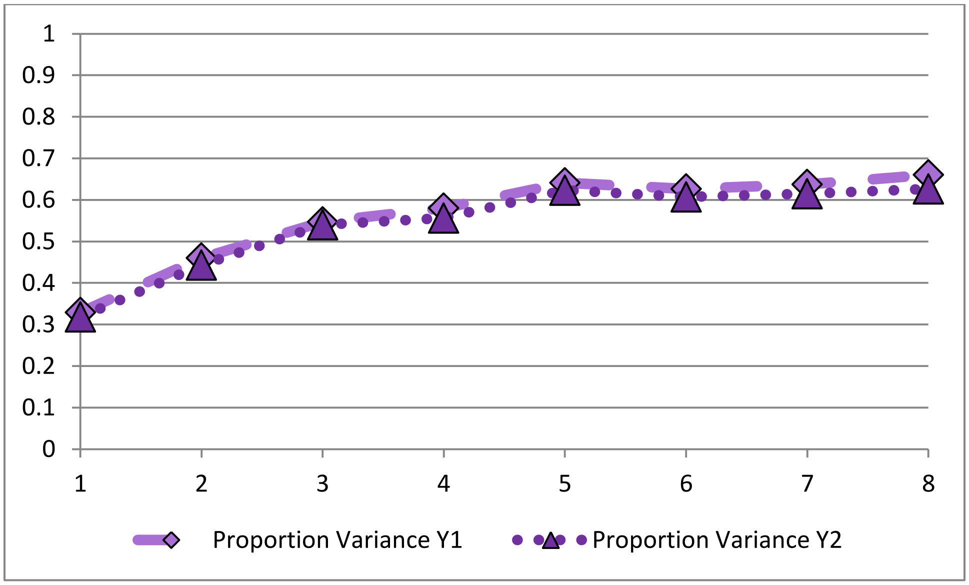
The full set of factor loadings for each solution is provided in Appendix E for solutions from 1 to 8 factors. The chart was color-coded so that components shown in rows shaded the same color indicated those belonging to the same domain; the CLASS domains were shaded as they were at the time of the MET study. The CLASS dimensions begin with a "C_" in the list, and the number at the end of the dimension indicates the time segment. The Framework for Teaching components begin with an "F_" in the list. Y1 and Y2 data were shown separately. To assist in interpretation, for each component the largest loading was boxed in and the factors were ordered roughly consistently across each analysis.

The combined data had a curious structure. The 2-factor results had all of CLASS time segment 1 and all of FfT loaded together on the first factor, and all of CLASS time segment 2 on the second factor. It seemed clear that the strong time-segment effect seen in the single-instrument analysis of CLASS had persisted into the combined analysis. The 3-factor solution, more clearly in Y1 than in Y2, showed the CLASS (recently defined) Classroom Organization domain separated from the other dimensions of CLASS onto the 3rd factor with all of the FfT components; factor 1 was the remaining time-segment 1 CLASS dimensions, and factor 2 was the time-segment 2 CLASS dimensions. The 4-factor solution showed the same basic structure except that most of the FfT components moved into a factor separate from the Classroom Organization dimensions. There was cross-loading of the Domain 2 FfT components, except

Culture for Learning, with the CLASS Classroom Organization dimensions in this solution. In the 5-factor solution, CLASS had resolved into the structure seen in the single-instrument analysis: a 2-factor solution within each time segment, with the Classroom Organization dimensions loaded on one factor and all other CLASS dimensions on the other. FfT in the 5-factor solution had returned to loading on a single factor. In the 6-factor solution, the CLASS structure remained stable from the 5-factor solution and the single-instrument analysis. FfT resolved into approximately its single-instrument analysis solution as well, with all Domain 2 components except Culture for Learning loaded onto a single factor, and the Domain 3 components plus Culture for Learning loaded together. There were more cross-loadings in the combined data from the Domain 2 components onto the latter factor than in the single-instrument results. In the 7- and 8-factor results, the CLASS dimension Positive Climate cross-loaded onto a factor alone, and the FfT Domain 2 cross-loadings disappeared except for Culture for Learning, which was primarily loaded on the FfT Dimension 3 factor.

The proportion of variance accounted for in each model between 1 and 8 factors for the combined CLASS and FfT data is shown in Figure 10. The two years of data were shown separately, as there are small differences in the model results. A 1-factor solution accounted for only about 32% of the variance. The 3-factor solution accounted for about 54% of the variance. Increasing to the 6-factor solution augmented the variance accounted for to about 61% with limited gains in the models with more factors.

Figure 10: CLASS & FfT Combined Proportion Variance Accounted For in FA Models



All inter-correlations in Table 6 are smaller than 0.7, although there are numerous moderate correlation values, so the solution may have identified distinct constructs in the data.

Table 6: CLASS & FfT Combined Factor Inter-Correlations

Y1	MR3	MR2	MR1	MR4	MR5	MR6		**Y2**	MR3	MR2	MR1	MR4	MR5	MR6
MR3	1	0.38	0.46	0.24	0.52	0.2		MR3	1	0.38	0.46	0.24	0.52	0.2
MR2	0.38	1	0.42	0.48	0.3	0.17		MR2	0.38	1	0.42	0.48	0.3	0.17
MR1	0.46	0.42	1	0.36	0.46	0.51		MR1	0.46	0.42	1	0.36	0.46	0.51
MR4	0.24	0.48	0.36	1	0.45	0.42		MR4	0.24	0.48	0.36	1	0.45	0.42
MR5	0.52	0.3	0.46	0.45	1	0.47		MR5	0.52	0.3	0.46	0.45	1	0.47
MR6	0.2	0.17	0.51	0.42	0.47	1		MR6	0.2	0.17	0.51	0.42	0.47	1

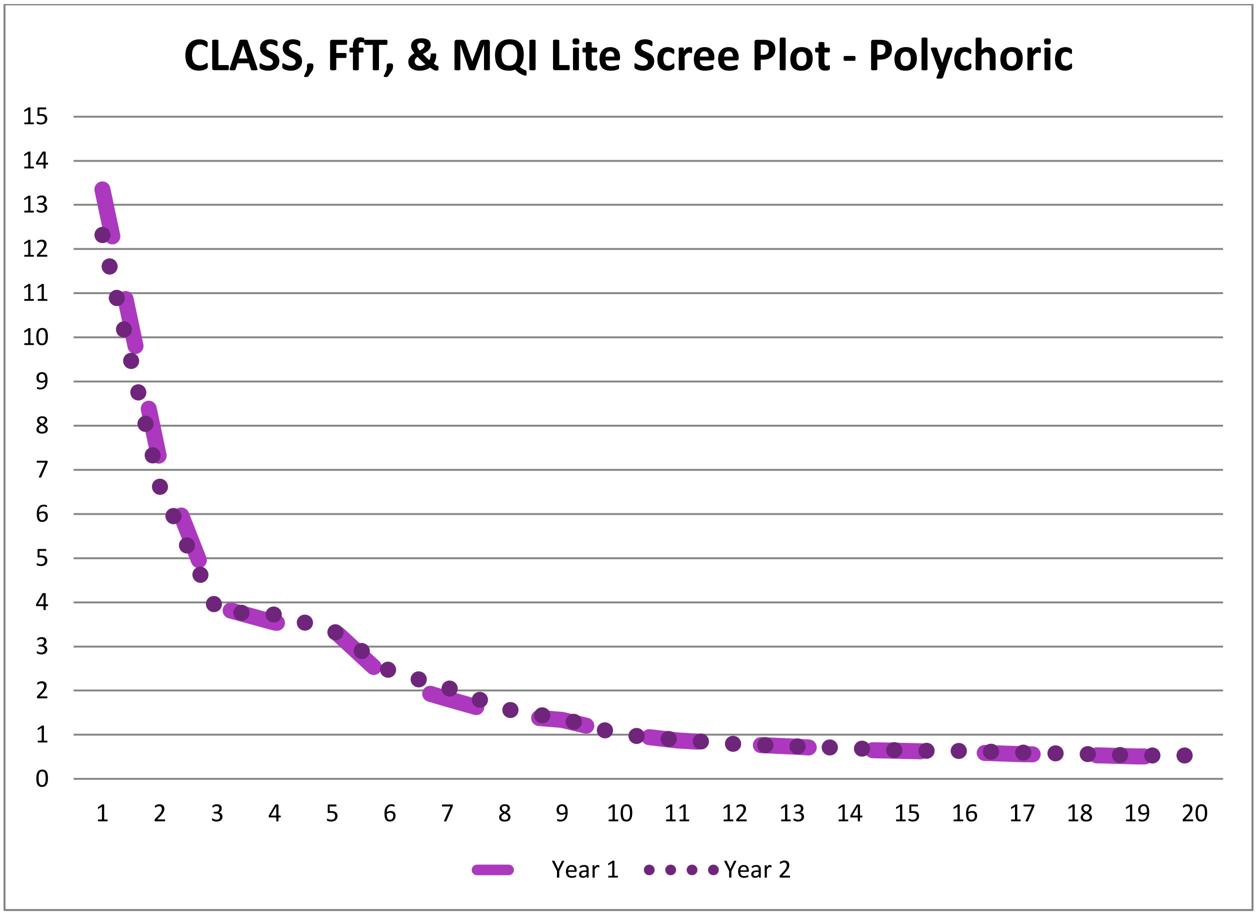
Despite some cross-loadings within FfT, the 6-factor solution was the most interpretable. The union of the 4-factor solution from the single-instrument analysis of CLASS and the 2-factor solution from the single-instrument analysis of FfT was the 6-factor solution for the combined data set. There were no cross-loadings between the instruments. The results of this analysis implied that the solution where MQI Lite and PLATO Prime data were included also may have 6 or more factors, if this basic CLASS/FfT solution remained stable.

Combined-Instrument Factor Analysis Results: CLASS, FfT, and MQI Lite

There were 2,576 cases used in Y1 analysis and 2,976 cases used in Y2 analysis of the combined CLASS/FfT/MQI Lite data set. The scree plot for CLASS, FfT, and MQI Lite combined is presented in Figure 11. Recall that we accepted a 4-factor solution for MQI Lite as most probable and interpretable, and that the combination of CLASS and FfT resulted in each instrument retaining its factor structure from the individual analysis—so a 6-factor solution was chosen. If the factor structures remained relatively intact within each instrument, we might expect to see a 10-factor solution for this combined data set. Since MQI Lite was focused exclusively on mathematics and instruction of content, it would seem to have little in common with the content-neutral instruments. There are 54 factors in the analysis, but for the sake of simplicity in examining the scree plot, only the first 20 were shown in Figure 11. At that point, the eigenvalues were well below 1.0 in value (eigenvalues starting at 11 were less than 1.0 in value) and the plot appeared to have leveled off.

Based on the scree plot, there was a noticeable bend at 3 factors and at 5 factors. From there the plot was difficult to interpret, as it sloped smoothly down, a pattern that continued through to the 54th factor. The full set of factor loadings for each solution is provided in Appendix F for solutions from 1 to 12 factors. The chart was color-coded so that components shown in rows shaded the same color indicated those belonging to the same domain; the CLASS domains are shaded as they were at the time of the MET study. The CLASS dimensions begin with a “C_” in the list, and the number at the end of the dimension indicates the time segment. The Framework for Teaching components begin with an “F_” in the list. The MQI Lite dimensions begin with “M_” in the list. Y1 and Y2 data were shown separately. To assist in interpretation, for each component the largest loading was boxed in.

Figure 11: Scree Plot for CLASS, FfT, and MQI Lite MET Data Combined: Year 1 and Year 2



The 2-factor solution was notable in that MQI Lite separated into one factor, with CLASS and FfT combined into the other. In the 3-factor solution, MQI Lite remained loaded on one factor, CLASS time segment 1 and FfT on a second, and CLASS time segment 2 on the third. The 4-factor solution had MQI Lite's E&I dimension separated onto a factor of its own, and the overall MQI and MKT showed negative cross-loadings with it. In the 5-factor solution, where the next bend in the scree plot appeared, the CLASS Classroom Organization domain appeared loaded onto a single factor. This pattern of the factor structure seen within each individual instrument asserting itself continued across the 6- through 9-factor solutions, with one exception: FfT. For every solution examined, all FfT components remained loaded on a single factor, and this remained the case through the 12-factor solution. FfT never broke into the 2-factor solution seen when the FfT data were analyzed alone. The 9-factor solution had the 4-factor (or 2-factor within two time segments) solution from the individual CLASS analysis, the 4-factor solution from the individual MQI Lite analysis, and FfT loaded onto a single factor. This was the most interpretable solution of the set.

The proportion of variance accounted for in each model between 1 and 12 factors for the combined CLASS, FfT, and MQI Lite data is shown in Figure 12. The two years of data were shown separately, although there were only small differences in the results. A 1-factor solution accounted for only about

23% of the variance. The 3-factor solution accounted for about 43% of the variance, and the 5-factor solution accounted for about 56%. The interpretable 9-factor solution accounted for about 70% of the variance with limited gains in the models with more factors, although the curve was still climbing slowly past 9 factors.

Figure 12: CLASS, FfT & MQI Lite Combined Proportion Variance Accounted For in FA Models

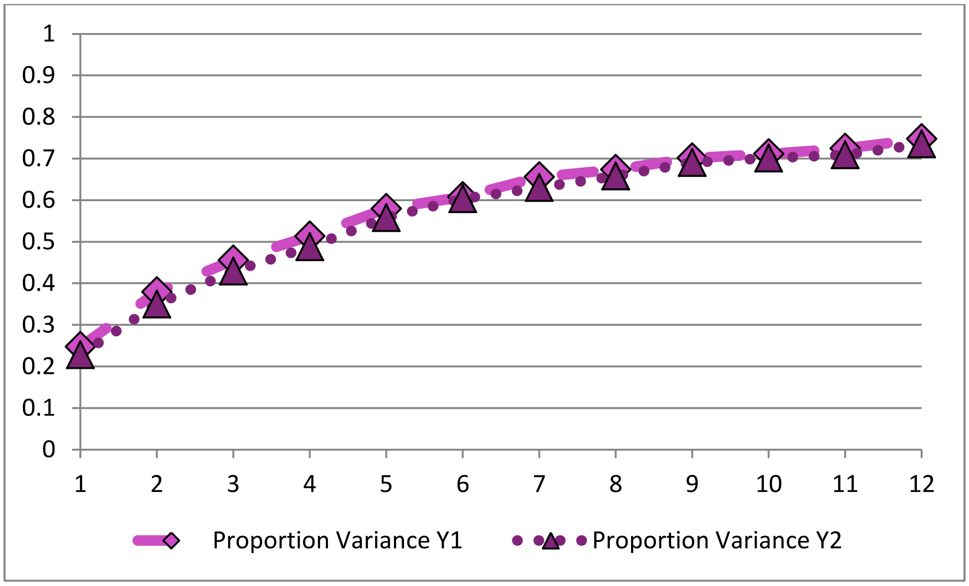


Table 7: CLASS, FfT & MQI Lite Combined Factor Inter-Correlations

Y1	MR5	MR3	MR1	MR2	MR4	MR7	MR8	MR9	MR6
MR5	1	0.36	0.41	0.17	0	0.13	0.29	0.44	0.18
MR3	0.36	1	0.34	0.14	-0.02	0.12	0.29	0.26	0.4
MR1	0.41	0.34	1	0.19	-0.02	0.17	0.28	0.48	0.37
MR2	0.17	0.14	0.19	1	-0.35	0.53	0.53	0.18	0.11
MR4	0	-0.02	-0.02	-0.35	1	-0.26	-0.14	0.02	0.02
MR7	0.13	0.12	0.17	0.53	-0.26	1	0.5	0.14	0.08
MR8	0.29	0.29	0.28	0.53	-0.14	0.5	1	0.16	0.11
MR9	0.44	0.26	0.48	0.18	0.02	0.14	0.16	1	0.41
MR6	0.18	0.4	0.37	0.11	0.02	0.08	0.11	0.41	1

Y2	MR1	MR5	MR4	MR3	MR6	MR2	MR8	MR9	MR7
MR1	1	0.4	0.36	-0.02	0.2	0.14	0.26	0.24	0.39
MR5	0.4	1	0.34	-0.01	0.2	0.1	0.17	0.39	0.4
MR4	0.36	0.34	1	-0.02	0.18	0.11	0.23	0.41	0.24
MR3	-0.02	-0.01	-0.02	1	-0.26	-0.26	-0.08	-0.02	0.02
MR6	0.2	0.2	0.18	-0.26	1	0.45	0.37	0.16	0.17
MR2	0.14	0.1	0.11	-0.26	0.45	1	0.49	0.05	0.07

MR8	0.26	0.17	0.23	-0.08	0.37	0.49	1	0.04	0.06
MR9	0.24	0.39	0.41	-0.02	0.16	0.05	0.04	1	0.39
MR7	0.39	0.4	0.24	0.02	0.17	0.07	0.06	0.39	1

The data in Table 7 contained a substantial number of small inter-correlations, with no values greater than 0.5 in magnitude.

Combined-Instrument Factor Analysis Results: CLASS, FfT, and PLATO Prime

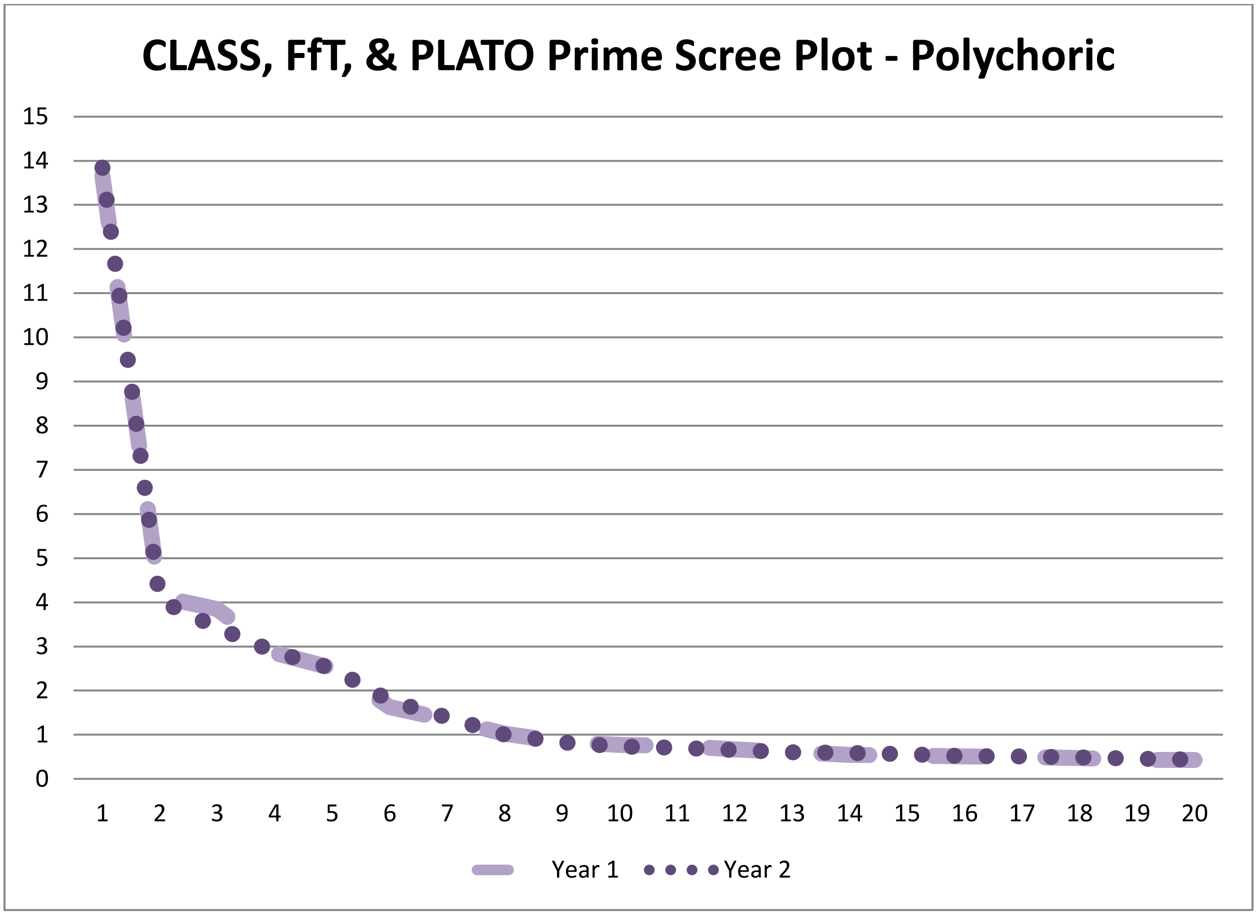
There were 2,910 cases used in Y1 analysis and 1,906 cases used in Y2 analysis of the combined CLASS/FfT/PLATO Prime data set. The scree plot for CLASS, FfT, and PLATO Prime combined is presented in Figure 13. Recall that we accepted a 3-factor solution for PLATO Prime as most probable and interpretable, and that the combination of CLASS and FfT resulted in each instrument retaining its factor structure from the individual analysis—so a 6-factor solution was chosen. If the factor structures remain relatively intact within each instrument, we might expect to see a 9-factor solution for this combined data set. If the FfT data behave as they did when combined with CLASS and MQI Lite and load on only a single factor, we might expect to see an 8-factor solution.

PLATO Prime has a primary focus on ELA and instruction of content, but, unlike MQI Lite, PLATO Prime includes two elements that measure general classroom characteristics, giving it more in common with the content-neutral instruments. There are 44 factors in the analysis, but for the sake of simplicity in examining the scree plot, only the first 20 were shown in Figure 13. At that point, the eigenvalues were well below 1.0 in value (eigenvalues starting at 9 were less than 1.0 in value) and the plot appeared to have leveled off.

The scree plot appeared to have a sharp bend at factor 2 and a fairly smooth curve sloping down from there. The tail appeared level by about factors 8 to 10. The full set of factor loadings for each solution is provided in Appendix G for solutions from 1 to 12 factors. The chart was color-coded so that components shown in rows shaded the same color indicate those belonging to the same domain; the CLASS domains were shaded as they were at the time of the MET study. The CLASS dimensions begin with a “C_” in the list, and the number at the end of the dimension indicates the time segment. The Framework for Teaching components begin with an “F_” in the list. The PLATO Prime dimensions begin with “P_” in the list. Y1 and Y2 data were shown separately. To assist in interpretation, for each component the largest loading was boxed in.

Given the large number of FA models examined, only a few will be discussed here. The 3-factor solution, where the first bend in the scree plot appears, was not easily interpretable. All of the FfT components, the Classroom Organization dimensions of CLASS, and PLATO Prime’s Time and Behavior Management elements all loaded onto a single factor that could be interpreted as representing some version of classroom organization and structure. The other two factors were each a single time-segment factor for the remaining CLASS dimensions plus some small loadings on the rest of the PLATO Prime elements, some of which are cross-loaded.

Figure 13: Scree Plot for CLASS, FfT, and PLATO Prime MET Data Combined: Year 1 and Year 2

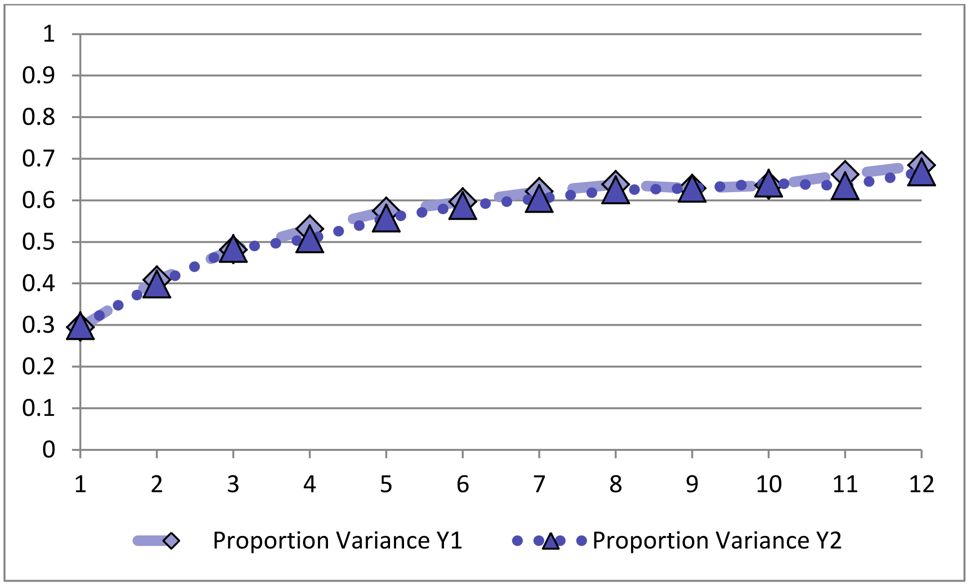


The 6-factor solution was interesting. In it, CLASS had sorted itself into a version of the solution seen in every analysis of the CLASS data, except that here the Classroom Organization domain for both time segments was loaded onto the same factor; the rest of the CLASS dimensions within each time segment load onto one factor each. Also loaded onto the factor with CLASS' the Classroom Organization domain were PLATO Prime's Time and Behavior Management elements and FfT's Managing Student Behavior component—leading to a possible interpretation of this factor as classroom procedures and management. The last 3 factors were loaded with the remaining 7 FfT dimensions; PLATO Prime's Strategy Use and Modeling; and PLATO Prime's Intellectual Challenge and Classroom Discourse. So in the 6-factor solution PLATO's single-instrument solution had reappeared, as had a version of CLASS'. FfT had largely remained loaded onto a single factor as seen in the CLASS/FfT/MQI Lite analysis, and a classroom procedures and organization factor had emerged across the three instruments. But this solution was not stable when more factors were introduced.

Increasing the number of factors resulted in the factor solutions from the single-instrument analyses reasserting themselves. CLASS returned to its 2-factor within two time segments solution in the 7-factor solution, with the time segment 2 Classroom Organization dimensions separating from the cross-instrument classroom procedures and management factor seen in the 6-factor solution. In the 8-factor solution, the separation of the instruments was nearly complete, with only FfT's Managing Student Behavior component loaded jointly with PLATO Prime's Time and Behavior Management elements. The 9- and 10-factor solutions were not more interpretable, as they produced one or a few scales loading onto the additional factors. In the 11- and 12-factor solutions, the 2-factor structure seen in the FfT single-instrument analysis finally emerged. CLASS' 2-factor by 2-time-segment structure was relatively immutable even in these large solutions, but PLATO's interpretable 3-factor solution began to "smear" and lose coherence as cross-loadings and single-element factors appeared. The 8-factor solution was probably the best choice, based on interpretability as well as acceptable support from the scree plot.

The proportion of variance accounted for in each model between 1 and 12 factors for the combined CLASS, FfT, and PLATO Prime data is shown in Figure 14. The two years of data were shown separately, although there are only small differences in the results. A 1-factor solution accounted for only about 29% of the variance. The 3-factor solution accounted for about 48% of the variance, and the 6-factor solution accounted for about 59%. The interpretable 8-factor solution accounted for about 63% of the variance. The curve leveled out briefly from there through the 11-factor solution, and began climbing slowly again at the point where FfT breaks into a 2-factor structure.

Figure 14: CLASS, FfT & PLATO Prime Combined Proportion Variance Accounted For in FA Models



Although there are numerous moderate inter-correlations between the factors shown in Table 8, none are unacceptably large.

Table 8: CLASS, FfT & PLATO Prime Combined Factor Inter-Correlations

Y1	MR2	MR3	MR4	MR1	MR5	MR6	MR7	MR8
MR2	1	0.36	0.46	0.36	0.26	0.2	0.41	0.48
MR3	0.36	1	0.4	0.31	0.21	0.44	0.36	0.26
MR4	0.46	0.4	1	0.56	0.29	0.41	0.39	0.5
MR1	0.36	0.31	0.56	1	0.42	0.57	0.35	0.63
MR5	0.26	0.21	0.29	0.42	1	0.21	0.41	0.28
MR6	0.2	0.44	0.41	0.57	0.21	1	0.17	0.42
MR7	0.41	0.36	0.39	0.35	0.41	0.17	1	0.18
MR8	0.48	0.26	0.5	0.63	0.28	0.42	0.18	1

Y2	MR2	MR3	MR4	MR1	MR6	MR5	MR7	MR8
MR2	1	0.36	0.44	0.39	0.39	0.2	0.52	0.24
MR3	0.36	1	0.41	0.39	0.36	0.18	0.35	0.49
MR4	0.44	0.41	1	0.55	0.34	0.16	0.48	0.35
MR1	0.39	0.39	0.55	1	0.35	0.34	0.59	0.55
MR6	0.39	0.36	0.34	0.35	1	0.31	0.2	0.09
MR5	0.2	0.18	0.16	0.34	0.31	1	0.16	0.17
MR7	0.52	0.35	0.48	0.59	0.2	0.16	1	0.39
MR8	0.24	0.49	0.35	0.55	0.09	0.17	0.39	1

Limitations

It is important to remember that at least one of the circumstances that made this study possible is also a limitation: it is a single data set. It is a very large and high-quality data set, including a large number of teachers drawn from a range of school districts across the country, and a large number of class sessions are included in it. There were hundreds of trained observers from across the country as well. But it is still one study conducted under one set of design constraints, and whatever aspects make the data unique also may limit the generalizability of these findings. The precise versions of MQI (MQI Lite) and PLATO (PLATO Prime) that were used in the MET study were not used in any other studies of which the authors are aware. CLASS has reorganized the instrument domains, as noted in the description. FfT has been revised since MET in two newer versions, one in 2011 that is very close to the MET version and one in 2013 incorporating small changes to align with the Common Core State Standards. As a result, the findings herein regarding the factor structure of the data from these instruments may not be applicable to data collected using other versions.

Interpretation and Conclusions

Most jurisdictions simply cannot invest the resources in repeated scoring of teacher practice with multiple instruments. The MET data offer a unique opportunity to examine the degree to which multiple instruments capture the same aspects of teaching and whether each contributes something unique.

Despite the fact that the data set is relatively distinctive, it offered an opportunity to investigate a question relevant to policy decisions being made regularly in the field: do different teaching practice observation instruments measure different things? The answer, based on these data, is: yes, they do.

If the instruments captured aspects of teaching practice that were—statistically at least—the same, then the factor analysis of the polychoric correlation matrices should have shown factor loadings indicating the common structure. Almost without exception, this was not the case. Whatever factor structure appeared in the MET data in the single-instrument analysis persisted when data sets from multiple instruments were combined. The exception was FfT. Analyzed alone and analyzed with CLASS, FfT has a 2-factor structure that mirrors its domain structure, with the exception of one component aligning with the other domain. But adding a third instrument, either MQI Lite or PLATO Prime, caused FfT to load on a single factor pretty consistently. It is worth recalling that the factor inter-correlations from the single-instrument analysis of FfT were quite large, an indicator that the factors may not have identified distinct constructs. Based on the analysis results of the MET data set, FfT is the instrument most sensitive to the presence of another instrument's data in analysis, but only if that instrument is a content-specific one. And, oddly, the presence of the content-specific instrument data seems to have the effect of compressing the 2-factor structure of FfT into 1 factor, not emphasizing it or expanding the structure into cross-loadings with the added data. This may be an indication that FfT truly has only one underlying construct, but the evidence from this study is ambiguous.

If there was an expectation that any instruments would share commonality, it seemed most probable that this would occur when combining data from CLASS and Framework for Teaching. Both instruments are content-neutral and can be used in classrooms across a range of grade levels[2]. The two rubrics were developed based on different philosophies of instruction. However, the superficial similarities of the dimensions/components would seem to suggest that some coalescing of data into common factors was likely to occur in these two instruments, if in no other combination. It didn't happen. Given sufficient (factor) space, the data retreated neatly within each instrument, with minimal or no cross-loadings, and re-formed the within-instrument factor structure alone.

So CLASS and FfT do indeed allow observers to see different things. The dimensions of CLASS and the components of FfT did not load together in this analysis—indeed, they come quite close to repelling each other! Despite apparent similarities in scales named "Behavior Management" on CLASS and "Management of Student Behavior" on FfT, the applied or operational definition of what the observers are looking for seems to be unalike.

Other than the lack of combining of data into common factors, the most intriguing finding in the analyses was the split of the CLASS data into factors within time segment. The strong effect of the time segment in these data is thought-provoking. CLASS typically is not scored in this design, with different raters for sequential time segments of the same class session. Some possible explanations of the observed outcome include:

[2] CLASS does have different tools for different grade levels, but for the two levels used in the MET study, the differences are limited are occur primarily in the training exemplars; the instrument description is very similar.

- The effect is due to the change of rater;
- The effect is due to the change in time in the class session (i.e., teachers alter instruction between the beginning and middle of the class sufficiently that raters assign qualitatively different scores to the time segments);
- Both of the above are true; or
- Some other alternative.

There seems no reason to expect that the academic instruction differs strongly between the first and second 15 minutes of the same class, but there time-segment effect in the CLASS MET data is very clear in the results. The finding would seem potentially to cast doubt onto the exchangeability of the CLASS MET raters—if the instruction does not differ, the other logical explanation is that the raters are scoring differently between time segments because they are different people. This explanation is challenged by the fact that the same rater pool was used interchangeably to score CLASS segment 1 and segment 2. They were scheduled in rotation to score one or the other depending on the days and shifts the raters were scheduled to work, but basically all CLASS raters scored both segments 1 and 2 in MET. There may be some effect of being "dropped in" to a session already in progress without the context from the initial 15 minutes that qualitatively and quantitatively alters the scoring. There are data in the MET study that may allow for more examination of these hypotheses, but the work is out of the scope of this study.

Summary

In terms of commonality across the data from the instruments analyzed in this study, there was effectively none. The instruments each cause the raters to see different aspects of teaching practice and to assign scores accordingly. Jurisdictions making decisions about instrument selection should not be misled by superficial similarities into believing that it does not matter which teaching practice observation instrument they select. It is clear from these analyses that it does matter—quite a lot—and jurisdictions must spend the time and gain the expertise necessary to make a thoughtful and well-informed decision about adoption of the instrument that most closely aligns to the local values and priorities.

References

D. J. Bartholomew. (1980). Factor Analysis for Categorical Data. *Journal of the Royal Statistical Society. Series B (Methodological).* Vol. 42, No. 3 (1980), pp. 293-321

Lee J. Cronbach & Paul E. Meehl. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, 52, 281-302.

Karl G. Jöreskog & Irini Moustaki (2001): Factor Analysis of Ordinal Variables: A Comparison of Three Approaches, *Multivariate Behavioral Research*, 36:3, 347-387

Robert J. Mislevy. (1986). Recent Developments in the Factor Analysis of Categorical Variables. *Journal of Educational Statistics,* Vol. 11, No. 1 (Spring, 1986), pp. 3-31.