Abstract Title Page

Title: Comparing Beginning Teachers' Instructional Quality Growth on Subject-Specific and Global Measures

Authors and Affiliations: Laura Neergaard, Vanderbilt University; Tom Smith, Vanderbilt University

Abstract Body

Background / Context:

Observation measures of instructional quality tend to fall into two broad categories – those for use across subject areas and those intended for use in specific subject areas. The move toward content-specific measures is a result of research suggesting that effective teaching looks different across subject areas (Evertson, Anderson, Anderson, & Brophy, 1980; Graeber, Newton, & Chambliss, 2012; Grossman & Stodolsky, 1995; McDonald & Elias, 1976; Stodolsky & Grossman, 1995) and that both content knowledge and pedagogical content knowledge are critical for effective teaching (Ball, Thames, & Phelps, 2008; Darling-Hammond, 2000; Harris & Sass, 2007; Hill, Rowan, & Ball, 2005; Shulman, 1986). Choosing either across-subject and subject-specific observational evaluation measures has interesting implications for beginning teachers who often struggle with global instructional practices like establishing a positive classroom climate and classroom management, while concurrently implementing rigorous content-area instruction (Grossman, 1992; Meister & Melnick, 2003). Despite these challenges, evidence has found that beginning teachers' effectiveness at increasing student achievement improves during the first few years on the job (Clotfelter, Ladd, & Vigdor, 2007; Harris & Sass, 2007; Kane, Rockoff, & Staiger, 2006). Yet, no research has examined whether beginning teachers are more likely to show improvement on a content-focused measure or a more general measure or whether improvements on one type of measure are associated with improvements on the other. These questions are important given the emphasis placed on evaluation of teachers' instructional practices as means for improving overall teacher quality.

Purpose / Objective / Research Question / Focus of Study:

Until the recent Measures of Effective Teaching (MET) project, which is comparing multiple measures of teaching effectiveness, studies of teaching quality using observational measures have largely used only one measure of instructional quality (Gates Foundation, 2010). Thus, the relationship between subject-specific and general measures of instructional effectiveness has rarely been explored. Furthermore, no study has compared these two types of measures for beginning teachers over time. This study will answer the following research questions: (a) To what extent do beginning teachers improve their instructional quality during their first three years of teaching, as measured by a standardized observation measure of global classroom quality, the Classroom Assessment Scoring System (CLASS), and a subject-specific rubric in mathematics, the Instructional Quality Assessment (IQA)?, and (b) Do teacher instructional quality ratings from the IQA and the CLASS, similarly identify high quality and low quality middle school mathematics teachers during their first three years of teaching? As part of the second question, we examine whether teachers who rate highly on measures of general instructional quality (e.g., classroom climate and organization) either initially or through improvement over time more likely to engage students in rigorous math activity and discussion.

Population / Participants / Subjects / Setting:

Data used includes classroom observations of teachers participating in a longitudinal study of beginning middle school math teachers' induction and mentoring experiences (the Assessing Induction and Mentoring project, AIM), funded by the National Science Foundation. Teachers were invited to participate in AIM if they met two inclusion criteria: (1) served as the teacher of record for at least one seventh or eighth grade math class; and (2) had no prior

experience as a teacher of record. All AIM participants who were observed at least once are included in this study.

Participants include 62 teachers in 11 districts across 4 states from three cohorts who began teaching in either 2007-2008, 2008-2009, and 2009-2010. For analysis purposes, the data have been pooled to examine the first, second, and third year teaching experiences across cohorts. The districts range in size and student composition (see Table 1). The largest district enrolled about 98,000 students and smallest enrolled about 7,000. Percent of the district's students receiving free or reduced price lunch (FRPL) ranged from 9 percent to 66 percent. The table also shows the number of teachers in the study from each district. About one-third of the teachers in the study are from the largest participating district.

About one-third of the teachers were male, most were white, and all had bachelor's degrees. Teacher degree focus was categorized as education, math education, math or other based on the teachers' majors in their undergraduate and graduate degrees. Compared to first year middle school mathematics teachers in the 2007-2008 Schools and Staffing Survey sample (N=44), AIM teachers are similar in age, gender, race, and the percentage with alternative certification, but AIM participants are more likely to have math education degrees and less likely to have mathematics degrees and student teaching experience (see Table 2).

Research Design:

The longitudinal nature of the data allow us to use growth curve analysis to investigate the extent to which teachers improve in the IQA and CLASS-S ratings over time. Growth curve modeling has rarely been used to examine ratings of teacher's instructional quality, though it makes sense to expect to see improvement over time, especially in the teachers' beginning years. A few studies have used growth curve modeling to examine change over time in university professors' instructional ratings (Lang & Kersting, 2007; Marsh, 2007). Advantages of individual growth curve modeling are that assessment times do not have to be identical, allowing respondents with missing data to remain in the analysis and that it captures the time-ordered nature of the observations.

Data Collection and Analysis:

Researchers videotaped participating teachers' instruction on two consecutive days during the same class period at four time points: winter of the first year of teaching and the spring of the first, second, and third years. All recordings were coded using both the CLASS-S and the IQA. Ratings across days were averaged, providing one set of ratings per teacher for each of the four observation periods.

The Instructional Quality Assessment (Junker et al., 2006; Matsumura et al., 2006) bases ratings on the use of cognitively demanding problem-solving tasks and discussions emphasizing reasoning and connections among mathematical ideas. The IQA assesses the quality of observed classroom instruction on three dimensions: (1) task potential, (2) task implementation, and (3) class discussion; ratings range from zero to four. A score of 0 specifies absence of mathematical activity or discussion; 1 points out instruction emphasizing facts and memorization; 2 indicates instruction emphasizing unambiguous application of procedures and single representations of concepts; and 3 or 4 designates instruction characterized by open-ended tasks, multiple representations of mathematical concepts, and connections among mathematical ideas.

While the IQA is specifically geared toward mathematics instruction, the CLASS-S captures a broader range of instructional practices (Pianta, Hamre, & Mintz, 2011). CLASS

dimensions are based on developmental theory and research suggesting that interactions between students and adults are the primary mechanism for student development and learning. The secondary version of the CLASS measures middle and secondary teachers' instructional quality across content areas in three broad domains: (a) Emotional Support (ES), (b) Classroom Organization (CO), and (c) Instructional Support (IS). Raters rate multiple short segments of instruction during a class period instead of rating a single class period of instruction in its entirety. Each domain is organized into multiple dimensions, and each dimension consists of several indicators. The *Emotional Support* domain includes positive climate, negative climate, teacher sensitivity, and regard for adolescent perspective. The *Classroom Organization* domain includes behavior management, productivity, and instructional learning formats. The *Instructional Support* domain consists of content understanding, analysis and problem solving, quality of feedback, and instructional dialogue. The CLASS also assesses student engagement, the degree to which students in the class are focused and participating in learning activities. Raters rate each dimension as low (1, 2), mid (3, 4, 5), and high (6, 7).

While the IQA and the CLASS-S are both concerned with the quality of a teacher's instructional practices, they privilege different criteria. The IQA focuses on elements that the developers felt were critical to developing students' conceptual understanding of mathematics. These elements are the rigor of the task provided, student work time, and discussion. In contrast, two-thirds of CLASS-S dimensions are dedicated to elements of classroom climate and organization, while one-third describes general instructional strategies. Table 3 shows which IQA rubrics and CLASS-S Dimensions measure similar concepts. All of the CLASS-S dimensions are from the Instructional Support domain. There is no conceptual overlap between the IQA and the CLASS-S Emotional Support and Classroom Organization domains.

To examine whether teachers are improving in their instruction, growth curve analysis will be used. We will model the teachers' IQA and CLASS-S score trajectories across the four time periods using a multilevel approach to growth curve modeling (Raudenbush & Bryk, 2002; Rogosa, Floden & Willett, 1984). This will be done this for an overall composite score of the IQA and for the emotional support, organizational support, and instructional support domains of the CLASS-S. At Level 1, the repeated IQA observations can be used to model a latent growth trajectory, the shape of which depends on a set of individual growth parameters. These parameters are the outcome variables in the Level-2 model, where they depend on teachers' individual characteristics. Background variables, including mathematics and education background and student teaching experience, will be included to control for pre-teaching characteristics that may be associated with instructional quality.

To evaluate the correspondence of IQA and CLASS-S instructional quality ratings, I will look at the three CLASS domain scores for teachers who score above and below a 3 average (across days and raters) on the IQA at each time point. This will reveal whether teachers who rate as "high" on the IQA have higher means on the CLASS than teachers who rate below "high" on the IQA. Eventually, OLS regression analysis will be used to determine whether having higher scores on ES and CO is predictive of having higher task implementation and discussion scores on the IQA. This will reveal whether teachers who rate highly on measures of global instructional quality (e.g., classroom climate and organization) either initially or through improvement over time are more likely to engage students in rigorous math activity and discussion. At this point, the analysis has only been completed for those teachers who remained in the study and in teaching for their first three years, as all coding has not yet been completed.

Findings / Results:

Mean IQA and CLASS-S scores for the four observations periods (first year winter, first year spring, second year spring, third year spring) are presented in Figures 1 and 2. Teachers scored the highest on Classroom Organization across all time points and the lowest in Instructional Support. Teachers' average scores do get higher over time. The means of all three domains across time all were in the "mid" category of the CLASS-S. On the IQA, teachers scored highest on average on Task Potential, though the averages indicate mostly procedural task assignment. On average, teachers score lowest on Discussion, indicating discussion characterized by one-word answers. Teachers' average scores on the IQA did not change much over time.

To analyze whether teachers are improving in their instruction, I used hierarchical linear growth modeling with repeated measures of the IQA/CLASS-S nested within teachers. Time is measured in months. Growth parameters for each of the CLASS-S domains and the IQA Overall scores are reported in Table 4. The coefficient on month is not significant for IQA, but is significant for all three CLASS-S domains. Teachers are predicted to increase on Emotional Support by about a fourth of a point over a year and by about a third of a point on Classroom Organization. Teachers are predicted to improve less on Instructional Support. Beginning math teachers appear to improve more in aspects of classroom management and organization over their first three years than they do in their instructional rigor.

Teachers who score a 3 or 4 on the IQA Task Implementation rubric score have higher average CLASS-S scores than teachers scoring a 2 or lower on the IQA. This relationship appears strongest with the Instructional Support domain and does not appear to change over time. The IQA Overall measure is most highly correlated with Instructional Support domain (see Table 5).

Conclusions:

While the teachers' IQA ratings did not change over time, teachers improved on the emotional support and classroom organization dimensions of the CLASS. Findings also show that teachers who exhibit higher levels of emotional support and classroom organization during also rate at higher levels of the math-specific IQA, indicating that teachers with stronger relationship with students and classroom management may be more be likely to employ reform math ideals.

A better understanding of the relationship between subject-subject and content neutral measures of instructional quality and how beginning teachers' instructional quality may improve over time can inform better evaluation system design. In the last decade, researchers have called for differentiated evaluation procedures based on teachers' development levels and content areas (Holland, 2006; Ponticell & Zepeda, 2004). One concern with current evaluation systems is that they use the same rating criteria regardless of experience level, despite findings that beginning teachers have different struggles than their more experienced colleagues. Teachers themselves have voiced that a particular evaluation system was appropriate for evaluating beginning teachers, but unlikely to meet the needs of experienced teachers (Peterson & Comeaux, 1990). Results from this study also question the appropriateness of a single, system-wide evaluation measure. They also bring up the question of what type of supports we should be providing to beginning teachers to help them implement rigorous content-area instruction.

Appendices

Appendix A. References

Bill & Melinda Gates Foundation (2010). Learning about teaching: Initial findings from the Measures of Effective Teaching Project, from http://www.metproject.org.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Evertson, C., Emmer, E. T., & Brophy, J. E. (1980). Predictors of effective teaching in junior high mathematics classrooms. *Journal for Research in Mathematics Education, 11*, 167-178.

Graeber, A. O., Newton, K. J., & Chambliss, M. J. (2012). Crossing the borders again: Challenges in comparing quality instruction in mathematics and reading. *Teachers College Record, 114*(4), 30 pages.

Grossman, P. L. (1992). Why models matter: An alternative view on professional growth in t teaching. *Review of Educational Research,* 62(2), 171-179.

Harris, D. N., & Sass, T. R. (2007). Teacher training, teacher quality, and student achievement. *National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Working Paper*, *3*.

Holland, P.E. (2006). The case for expanding standards for teacher evaluation to include an instructional supervision perspective. *Journal of Personnel Evaluation in Education, 18*, 67-77.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City*. National Bureau of Economic Research.

Lang, J. W., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science*, *35*(3), 187-205.

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology*, *99*(4), 775.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., et al. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment.* (CSE Technical Report #681). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

McDonald, F. J., & Elias, P. (1976). The Effects of Teaching Performance on Pupil Learning: Final Report, Volume I, Beginning Teaching Evaluation Study, Phase II, 1974-1976. *Princeton, NJ: Educational Testing Service*.

Meister, D. G., & Melnick, S. A. (2003). National new teacher study: Beginning teachers' concerns. *Action in Teacher Education*, *24*(4), 87–94.

Peterson, P. L., & Comeaux, M. A (1990). Evaluating the systems: Teachers' perspectives on teacher evaluation. *Educational Evaluation and Policy Analysis, 12*(1), 3-24.

Pianta, R. C., Hamre, B. K., & Mintz, S.L. (2011). *Classroom Assessment Scoring System – Secondary Manual*.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data*

analysis methods (2nd ed.). Thousand Oaks, CA: Sage.
Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology*, *76*(6), 1000.
Stodolsky, S. & Grossman, P. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal,* 32, 227-249.

Appendix B. Tables and Figures

Table 1. Characteristics of Participating School Districts

State	District	# of Teachers in Study	# of District Secondary Teachers	Urbanicity	Schools	Students	% Black or Hispanic	% FRPL
1	A	12	520	Urban	70	37,000	33	47
	B	21	1,480	Urban	180	98,000	41	59
	C	3	120	Rural	10	8,000	7	51
2	D	2	300	Suburban	20	11,000	14	14
3	E	1	190	Suburban	10	7,000	69	63
	F	1	430	Suburban	20	12,000	12	9
4	G	7	1,420	Urban	140	75,000	63	66
	H	4	780	Rural	40	38,000	25	37
	I	7	820	Suburban	50	48,000	42	31
	J	3	560	Suburban	50	27,000	14	35
	K	1	580	Rural	40	31,000	8	10

Source: Common Core of Data (2009-2010). Number of schools rounded to the nearest ten and number of students rounded to the nearest thousand to protect the identity of school districts.

Table 2. Teacher Background Characteristics from AIM and 2007-2008 SASS

	AIM (n=62)	SASS 1st Year (n=44)	SASS 1-5 Years (n=275)
Age	27.9	27.4	
Male	31%	25%	30%
White	89%	86%	87%
Education Degree	42%	36%	44%
Math Education Degree	24%	9%	10%
Math Degree	8%	16%	20%
Other Degree	26%	39%	26%
Alternative Certification	27%	32%	28%
Student Taught	62%	73%	81%

Table 3. Hypothesized Alignment Between IQA Rubrics and CLASS-S Dimensions

IQA Rubrics	CLASS-S Dimensions (indicators)
Task Potential	Analysis & Problem Solving (Inquiry & analysis, Opportunities for novel application, Metacognition)
Task Implementation	Analysis & Problem Solving (Inquiry & analysis, Opportunities for novel application, Metacognition)
Discussion	Content Understanding (Depth of understanding) Quality of Feedback (Feedback loops) Instructional Dialogue (Cumulative content-driven exchanges)
Participation	Quality of Feedback (Encouragement & affirmation) Instructional Dialogue (Distributed talk)
Teacher Linking	Quality of Feedback (Building on student responses) Instructional Dialogue (Cumulative content-driven exchanges)
Student Linking	Quality of Feedback (Building on student responses) Instructional Dialogue (Cumulative content-driven exchanges)
Teacher Press	Quality of Feedback (Scaffolding) Instructional Dialogue (Facilitation strategies)
Student Providing	Instructional Dialogue (Facilitation strategies)

Table 4. Growth Coefficients

	IQA Overall	CLASS-S Emotional Support	CLASS-S Classroom Organization	CLASS-S Instructional Support
Month	0.00 (0.00)	0.02*** (0.00)	0.03*** (0.01)	0.01** (0.01)
Constant	1.67 (0.08)	4.44 (0.12)	4.45 (0.17)	3.63 (0.12)

Table 5. CLASS-S Domain Correlations with Overall IQA Score

	FYR1	SYR1	SYR2	SYR3
Emotional Support	0.45	0.44	0.11	0.11
Classroom Organization	0.32	0.18	0.20	0.00
Instructional Support	0.59	0.46	0.22	0.50

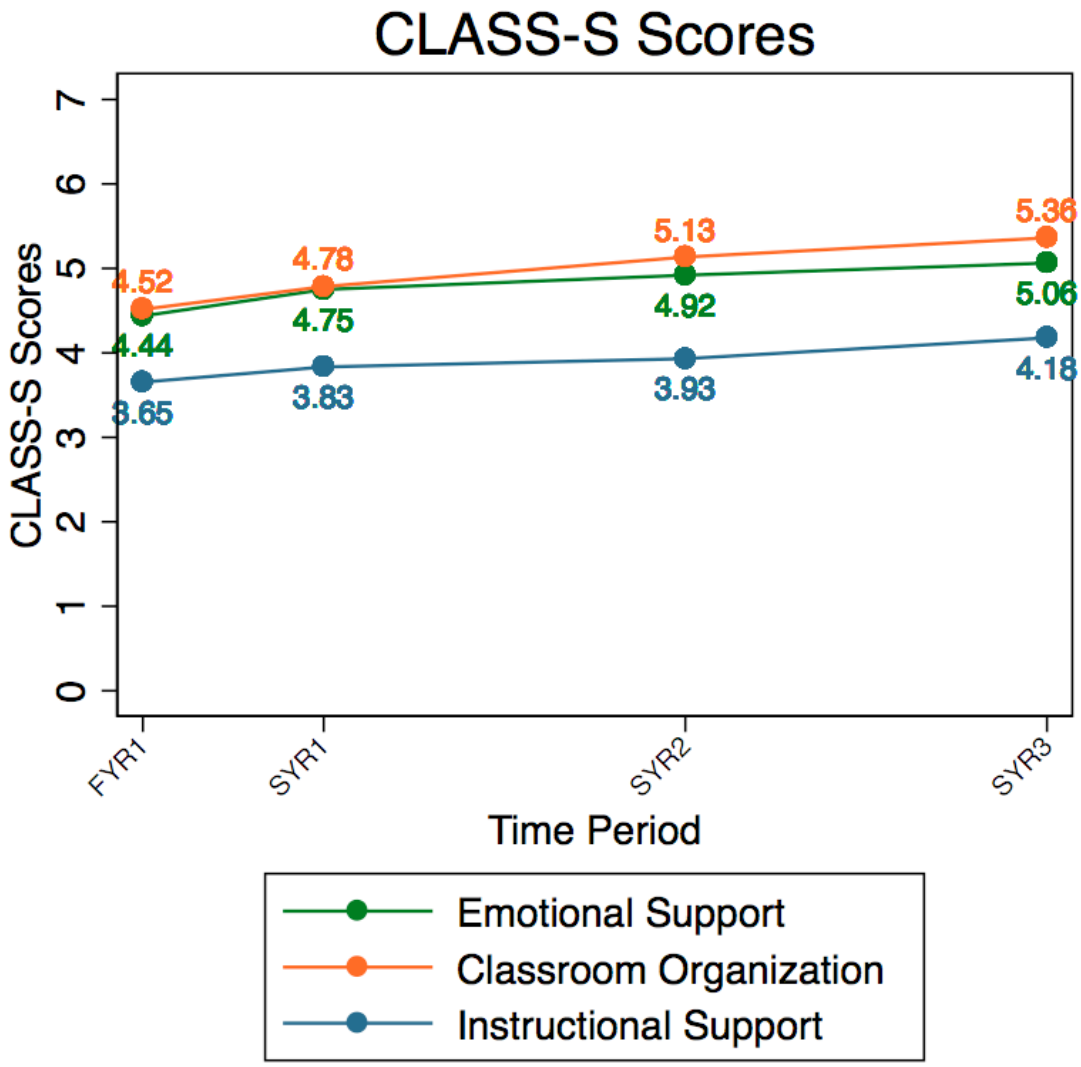

Figure 1. CLASS-S Scores Over Time

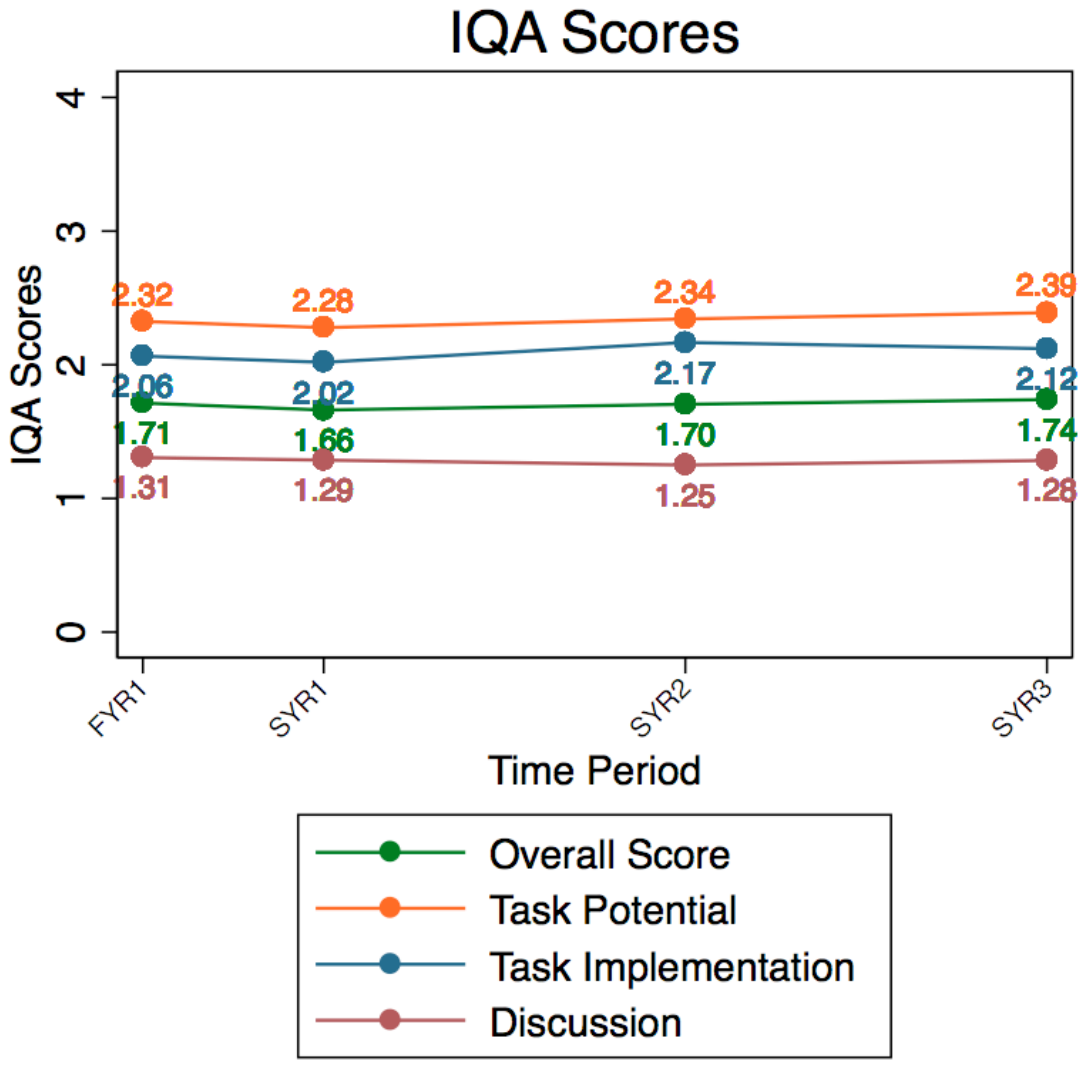

Figure 2. IQA Scores Over Time

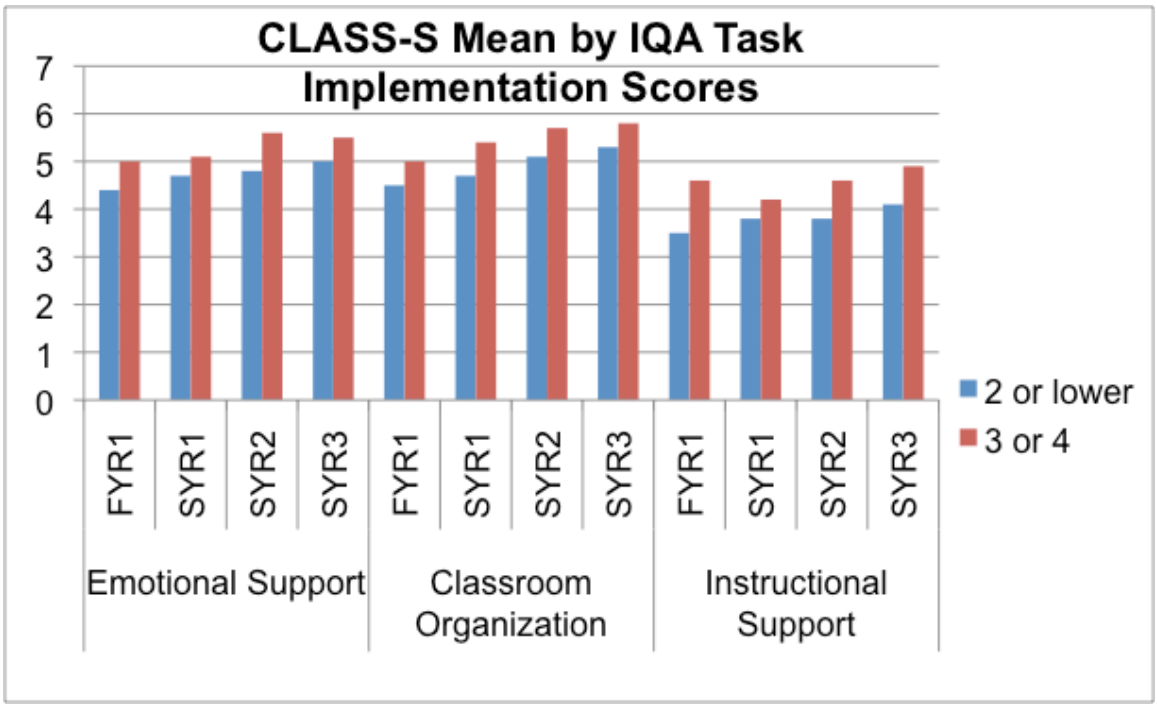

Figure 3. CLASS-S Means by IQA Task Implementation Scores