Abstract Title Page
Not included in page count.

Title: Do Interim Assessments Influence Instructional Practice in Year One? Evidence from Indiana Elementary School Teachers

Authors and Affiliations:
Gregory Chojnacki (American Institutes for Research),
Jared Eno (American Institutes for Research),
Feng Liu (American Institutes for Research),
Coby Meyers (American Institutes for Research),
Spyros Konstantopoulos (Michigan State University),
Shazia Miller (American Institutes for Research),
Arie van der Ploeg (American Institutes for Research)

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E090005 to Learning Point Associates, a subsidiary of the American Institutes for Research. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract Body

Background / Context:

Recent work that has examined the impact of what are variously called periodic, interim, benchmark, or diagnostic assessments, typically administered three or four times during a school year, has produced mixed findings. For instance, one study reported small significant effects in mathematics in grades 3-8, but not in reading (Carlson et al., 2011). Other research however, has reported significant effects on both mathematics and reading (Slavin et al., 2011). Finally, a very recent study found no effects on reading achievement in grades 4-5 (Cordray et al., 2012).

The state of Indiana was among the first to implement statewide technology-supported interim assessments in math and English Language Arts (ELA) to be taken by all K–8 students multiple times each school year at volunteering schools. Indiana expects teachers to use assessment information to improve ongoing instruction and increase student achievement. In 2008 the Indiana Department of Education (IDOE) began its roll-out of what it called its "diagnostic assessment tools."

In 2009-10, the American Institutes for Research conducted the first round of a two-cohort randomized controlled trial to evaluate the effectiveness of the interim assessment tool in schools receiving it for the first time (Konstantopoulos et al., in press). Findings suggested a positive but modest treatment effect across all grades. Still, even small positive impacts in the first year of an interim assessment intervention are notable, given evidence suggesting that such interventions may take multiple years to affect student performance (Slavin et al., 2011). Further, observed effect sizes in the range of 0.10 to 0.19 are of substantive policy interest.

The theory of action supporting interim assessments' effectiveness hinges on teachers making changes to their instructional practice (Blanc et al., 2010). In particular, differentiation of content scope and sequence, instructional level and grouping methods are among aspects of instructional practice theorized to improve quality of instruction by drawing on improved information about student needs (Tomlinson 2000). Evidence suggesting small, positive impacts in schools' first year using interim assessments motivates this study's focus on areas of teacher practice hypothesized to be intermediate outcomes of the interim assessment intervention.

Purpose / Objective / Research Question / Focus of Study:

This study compares instructional practices of teachers in schools that were randomly assigned to receive an interim assessment tool with those of teachers in schools that did not receive the tool. Using rich data collected at 16 time points during the school year, we study teachers' self-reported instructional practices to determine whether teachers with access to an interim assessment tool alter each of three facets of instructional practice—scope and sequence of content coverage, instructional level, and instructional grouping—more than those without the tool. Our research questions are:

(1) Do teachers with access to the interim assessment change the scope and sequence of content, and/or vary instructional difficulty level and grouping methods more than those without?
(2) Do variations in these teacher practices respond to variations in student Acuity performance?

Setting:

The data used in this study are drawn from an RCT that took place in Indiana in 2009-2010. Schools were randomly identified from a queue of K-8 public schools that had volunteered

to implement diagnostic assessments in the Spring of 2009. This set of schools was then randomly assigned to treatment or a control (one-year delay in implementation) condition.

Population / Participants / Subjects:

Data on instructional practices were collected for 2nd and 5th grade teachers as part of the RCT described above. This study focuses on 5th grade teachers using the Acuity Predictive assessment, a version of the interim assessment aligned to the statewide ISTEP+ exam administered each spring.[1] Eight students were randomly sampled in each participating teacher's class. Please see Table 1 for details on the study sample in the context of the broader RCT.

Intervention / Program / Practice:

As described by Konstantopoulos and colleagues (2011), the interim assessment tool studied here is a series of 30 – 35 item multiple choice tests in mathematics and ELA, administered three times during the school year. The tests are closely aligned to Indiana's statewide year-end test, and the intervention provided teachers with rapid access (within 24 hours) to a variety of class- and student-level reports on performance, including predicted proficiency on the year-end test.

Significance / Novelty of study:

A small but increasing number of rigorous evaluations of interim assessments exist (May and Robinson, 2007; Henderson et al., 2007; Quint, Sepanek and Smith, 2008; Carlson et al., 2011; Slavin et al., 2011; Cordray et al., 2012). However, the current study draws on considerably richer data on teacher practices than existing impact studies. As described in the data collection section, we analyze data from detailed checklists (or "logs") completed by teachers for each of eight students at sixteen time points during the school year. These data provide a nuanced picture of instructional practices utilized by teachers with and without the interim assessment intervention. By applying existing analytic methods to repeated, detailed measurements of teacher practice, we provide new evidence on teacher practices as intermediate outcomes responding to interim assessment information in the first year of implementation.

Research Design:

We employ treatment vs. control comparisons to explore whether teachers with the interim assessment intervention engage in expected instructional practices more than those without it. Treatment-group-only analyses of the association between teacher practices and student assessment performance provide evidence on the extent to which teachers target instruction to student performance. In this analysis, each testing window is considered as a juncture at which teachers potentially acquire new information about students. Accordingly, teacher change in instructional level is estimated at each assessment window.

Statistical, Measurement, or Econometric Model:[2]

When comparing Acuity teachers with comparison teachers, we employ hierarchical generalized linear models that account for the data's nested structure, with instructional logs (multiple time observations) nested within students, who are nested within teachers, who are in

[1] An additional rationale for this focus is that subgroup analyses in the original impact study indicate that measured impacts were largest in fifth grade and among Acuity Predictive users.

[2] Please note that outcome measures are described at the end of the Data Collection section.

turn nested within schools. We use a logistic model to estimate differences between treatment and control teachers on binary measures of instructional practice (described in the Data Collection and Analysis section below). We consider teacher decisions such as instructional grouping and level of instruction as student-level outcomes measured at each log, because the theory of differentiated instruction implies that teachers make distinct instructional decisions for each student. In contrast, we model curricular decisions, such as whether a given topic is covered on a given day, as class-level phenomena, since these are made for the whole class. We model the parameter of interest, the treatment-control contrast, as a fixed coefficient and account for the nested data structure using random effects at the student, teacher and school level. The resulting model can be expressed as follows:

$$\ln\left[\frac{\Pr(Y_{ijkt}=1)}{\Pr(Y_{ijkt}=0)}\right] = \beta_0 + \beta_1 D_k + u_k + v_{jk} + w_{ijk} \text{ , (1)}$$

where Y_{ijkt} is the student-level observation of teacher practice, D_k indicates whether a school received the Acuity Predictive interim assessment tool, β_1 is the parameter of interest measuring the treatment-control contrast, u_k, v_{jk}, and w_{ijk}, represent school, teacher, and student random effects, respectively.[3] This model is estimated for the full sample including all logs for each student, as well as for subsamples including only the logs following each interim assessment window.

When analyzing associations between student Acuity performance and teacher practices, we consider a "differences in differences" specification in which teacher practice prior to each test acts as a counterfactual for practice following the test window. The first difference in the model is an average difference between students in the top and bottom half of their class sample's performance on the Acuity test (for example, a difference in share of students experiencing remedial instruction). If the contrast between top- and bottom-half performers grows following the test window, this is consistent with the hypothesis that teachers change their instructional practices based on new information from the Acuity assessment. This model can be expressed as follows:

$$\ln\left[\frac{\Pr(Y_{ijkt}=1)}{\Pr(Y_{ijkt}=0)}\right] = \beta_0 + \beta_1 B_i + \beta_2 P_t + \delta(T_i \times P_t) + u_k + v_{jk} + w_{ijk} \text{ , (2)}$$

where Y_{ijkt} is the student-level observation of teacher practice, B_i is an indicator taking "1" when a student is in the bottom half of his class-sample's Acuity performance for a given test window, P_t is an indicator taking "1" in the period following the test window and "0" before, δ is the parameter of interest measuring the difference in differences, and the last three terms are random effects as described in Equation 1. This model is estimated for two subsamples corresponding to the two Acuity assessment periods for which there are pre- and post-test data. Each subsample includes four instructional log dates, two before the test window and two after.

Usefulness / Applicability of Method:

The usefulness of the methods applied in this study is described in the significance section.

Data Collection and Analysis:

[3] In analysis of topics covered by teachers as an outcome (described briefly in the "Data Collection and Analysis" section), the teacher-date, rather than the student-date, is considered as an observation, because topic-level content coverage decisions are conceived as applying to the whole class. Accordingly, a three-level model is used, with no teacher-level random effect specified.

Teachers in control and treatment schools in grade 5 were asked to complete 16 instructional checklists throughout the school year, roughly one every two weeks. Our staff developed a separate checklist for math and ELA. The ELA checklists were based on Rowan and Correnti's checklist (2009).The mathematics checklists were developed by content experts, following the ELA model and guided by the Indiana mathematics standards. In each checklist, items were categorized by topic area. The math checklist had seven topic areas: number sense; computation; algebra and function; geometry; measurement; problem solving; and data analysis and probability. Each topic area contained items related to teacher instruction, concepts and skills, and student activities. The ELA checklist contained nine topic areas, and collected the same detailed instructional information as the math logs on five of these: comprehension, writing, word analysis, reading fluency, and vocabulary. Teachers completed checklists online and results were stored on servers.

Following procedures described by Rowan and Correnti (2009), eight students were randomly selected by each teacher to focus on while completing the checklist. These same eight students were used for the entire year. For each checklist date, teachers indicated whether each student was instructed in each topic and whether they used a particular instructional grouping method with each student. If they had taught particular content, they indicated whether they had taught that student at the remedial, regular, or enriched level.

Using data collected by teacher checklists, we developed binary measures indicating whether a student experienced relevant instructional practices on a given day. A series of binary variables - one for each topic area - indicate whether a student received instruction on each topic on that day. A second binary variable indicates whether a student received any remedial or enriched instruction that day. A third binary variable, used in the difference in differences model described above, measures whether a student received remedial instruction on a given day. Finally, two binary variables indicate whether a student received instruction in a small-group or individual format that day.

Findings / Results:

Initial findings suggest little evidence of strong impacts on teacher practice as a result of access to the Acuity Predictive interim assessment tool.[4] The time series presented in Figures 1-4 show some periods where Acuity teachers sustain higher levels of engagement in specific instructional practices than comparison teachers. However, these selective periods do not cohere into a broader pattern of Acuity teachers using expected practices more widely than comparison teachers. Panel A of Table 2 presents estimates of the treatment control contrast estimated using a three-level adaptation of Equation 1 over all logs; statistically significant contrasts do not emerge in any of the seven content areas. While estimates of the treatment vs. control contrast in use of individual and small group instruction (Panel B) are both positive and of substantial magnitude, they are also not statistically significant. The estimated difference in levels of targeted (enriched or remedial) instruction in Panel C similarly indicates that a lack of significant difference between the groups, although the estimate's sign and magnitude suggest that Acuity teachers may increase levels of targeted instruction. Results from the difference in differences models (Equation 2 above) are not presented here but are broadly confirmatory, characterized by mixed signs and few significant estimates.

Conclusions:

[4] Results discussed in this abstract are summary in nature and only consider mathematics instruction.

We report results from rich data on teacher instructional practices generated at sixteen intervals by teachers with and without access to a specific interim assessment tool. Estimates provide no strong evidence that teachers change the instructional practices measured here in response to Acuity performance data. One possible reason for these findings is that Acuity is not a unique intervention, and a significant number of control teachers reported using other interim assessment tools. Another possible explanation for these results is that the relatively small sample of teachers completing checklists harms the study's power. Finally, these results pertain to the first year of the intervention, when teachers are likely still learning how to use the assessment tool and integrate it into their instructional practice. Future research should explore the hypothesis that impacts on teacher practice grow over time as teachers learn to use the assessment tool.

Appendices

Not included in page count.

Appendix A. References

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities.*Peabody Journal of Education*, *85*(2), 205-225.

Carlson, D., Borman, G.D., & Robinson, M. (2011). A multi-state district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33,* 378-398.

Cordray, D., Pion, G., Brandt, C., and Molefe, A. (2012). *The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement* (Issues & Answers Report, REL 2012–No. 2013–4000).

Henderson, S., Petrosino, A. Guckenburg, S., & Hamilton, S. (2007a). *Measuring how benchmark assessments affect student achievement* (Issues and Answers Report, REL 2007 No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

Konstantopoulos, S., Miller, S., & van der Ploeg, A. (In press). The Impact of Indiana's System of Benchmark Assessments on Mathematics and Reading Achievement. *Educational Evaluation and Policy Analysis*.

May, H., & Robinson, M. A. (2007). *A randomized evaluation of Ohio's Personalized Assessment Reporting System (PARS).* Philadelphia: University of Pennsylvania Consortium for Policy Research in Education.

Quint, J., Sepanik, S., & Smith, J., (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Student Thinking in Reading (FAST-R) Program in Boston Elementary Schools.* New York: MDRC.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the Study of Instructional Improvement. *Educational Researcher*, *38*(2), 120.

Slavin, R. E., Cheung, A., Holmes, G. C., Madden, N. A., Chamberlain, A. (2011). Effects of a data-driven district reform model. Baltimore, MD: Johns Hopkins University's Center for Research and Reform in Education.

Tomlinson, C. A. (2000). *Differentiation of instruction in the elementary grades*. Champaign, IL: ERIC Clearinghouse on Elementary and Early Childhood Education, University of Illinois.

Appendix B. Tables and Figures

Not included in page count.

Table 1. Sample Size, Full RCT Sample and Subsample with Instructional Log Data

	Study Samples						
Grade 5 Observations	**All Acuity, Full RCT**	**Acuity Predictive Users, Full RCT**			**Acuity Predictive Users, Log Data Collected**		
	All	**All**	**T**	**C**	**All**	**T**	**C**
Schools	56	29	19	10	22	12	10
Teachers	148	87	57	30	52	27	25
Students	3,711	1,962	1,233	729	416	216	200

The right-most three columns report schools, teachers and students in the present study sample.

Table 2. Treatment vs. Comparison Contrast in Three Measures of Instructional Practice

Area of Instructional Practice	Treatment-Control Contrast (Odds Ratio)	Treatment-Control Contrast (Logit coefficient)	Standard Error	N
Panel A. Content Coverage				736
Number Sense	1.01	0.01	(0.37)	
Computation	1.06	0.06	(0.27)	
Algebra and Functions	1.25	0.23	(0.41)	
Geometry	0.93	-0.07	(0.24)	
Measurement	1.09	0.08	(0.30)	
Problem Solving	1.05	0.05	(0.28)	
Data Analysis and Probability	1.03	0.03	(0.32)	
Panel B. Instructional Grouping Methods				5450
Small Group Instruction	1.42	0.35	(0.52)	
Individual Instruction	1.29	0.25	(0.94)	
Panel C. Instructional Difficulty Level				4662
Received at least One Concept at Enriched or Remedial	1.41	0.35	(0.80)	

** $p < 0.05$*

Figure 1. Average Levels of Math Content Coverage in Seven Content Areas, 9/2009-5/2010

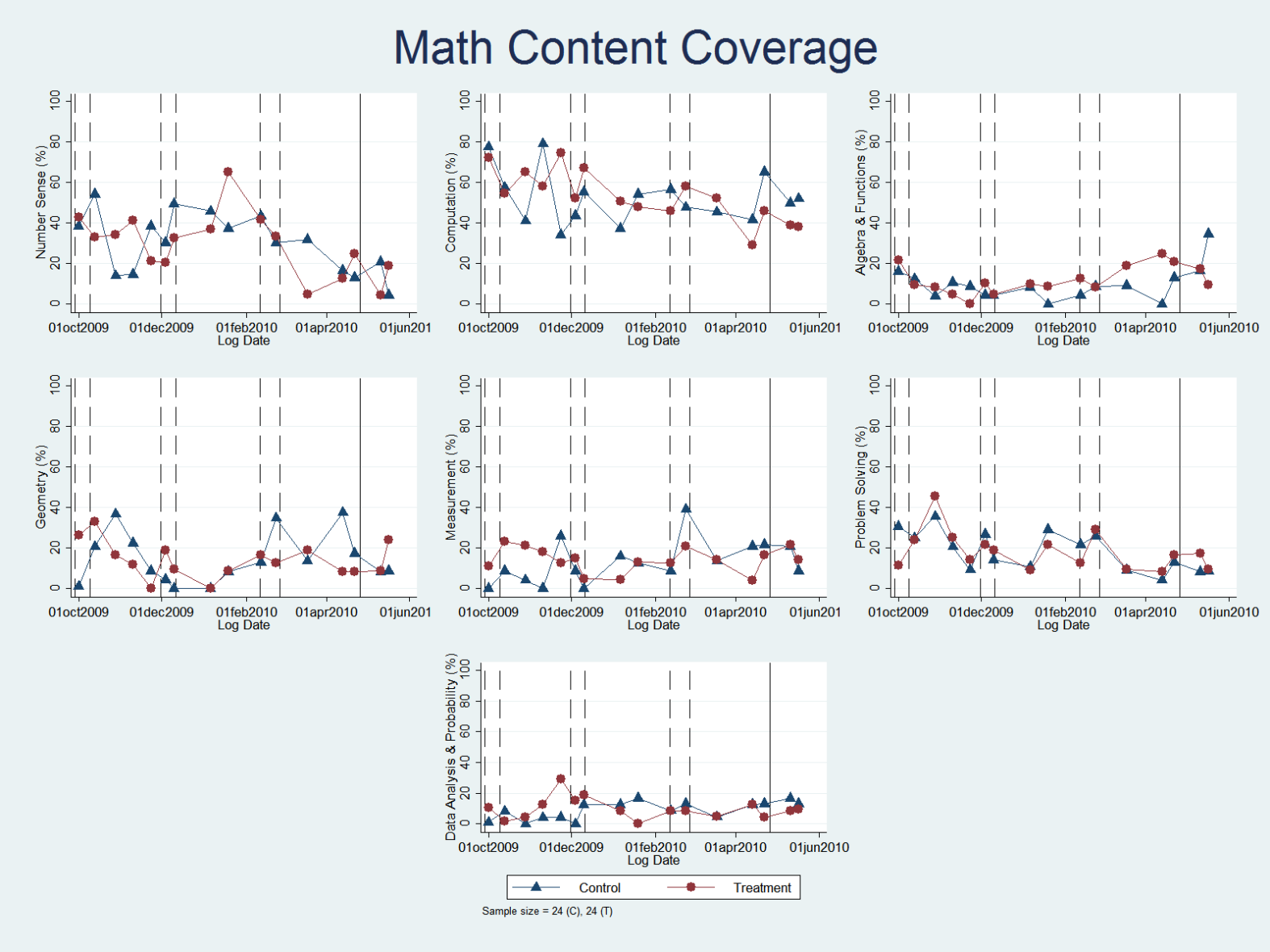

Figure 2. Average Levels of Small Group Instruction in Math, 9/2009 – 5/2010

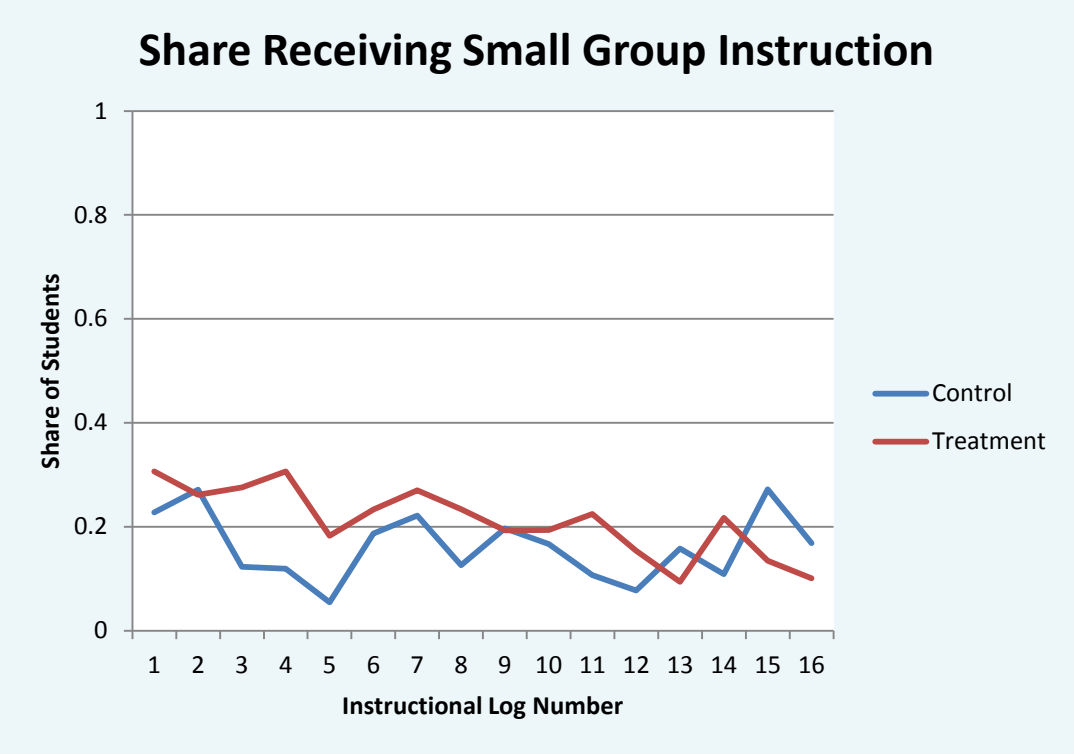

Figure 3. Average Levels of Individual Instruction in Math, 9/2009 – 5/2010

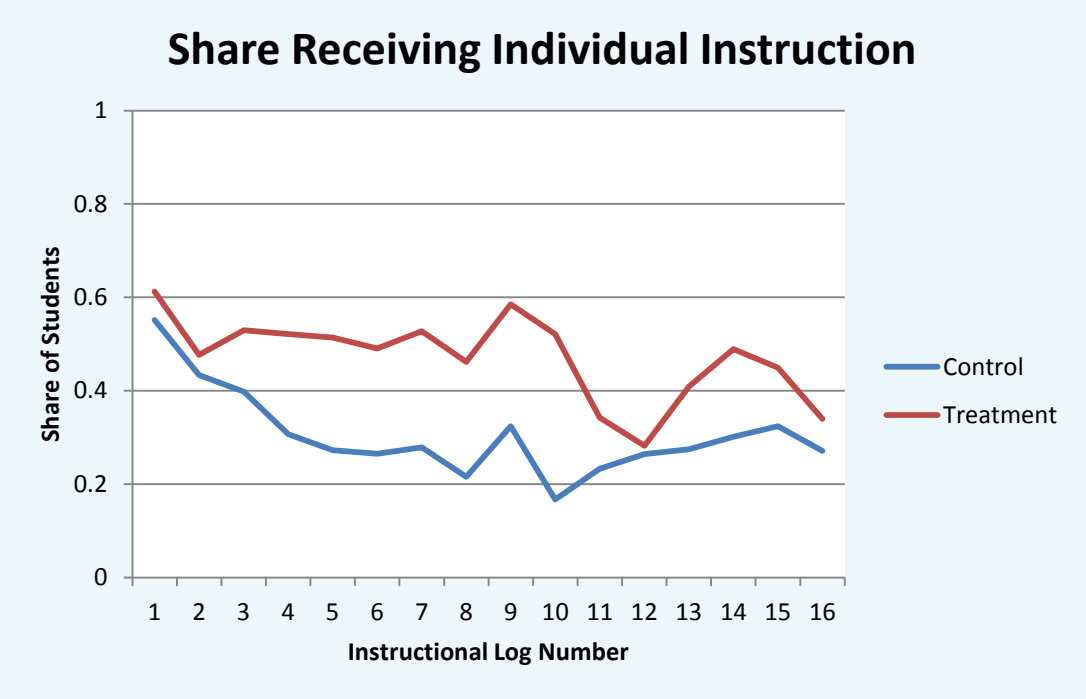

Figure 4. Average Student Share of Topics at Remedial or Enriched Level, 9/2009 – 5/2010