Incorporating End-of-Couse Exam Timing into Educational Performance Evaluations

Eric Parsons
Cory Koedel
Michael Podgursky
Mark Ehlert
P. Brett Xiang

Background and Motivation

- District and state education agencies are increasingly interested in incorporating test-based measures of performance into evaluation systems for districts, schools and teachers.
  - Educational administrators cannot cut out segments of the education population for convenience – e.g., evaluation systems in practice cannot simply focus on "tested grades and subjects."
    - Most of the research literature that has served as the impetus for creating these systems uses data from math and reading standardized tests for students in grades 3-8.

- Expanding the scope of evaluations outside of traditionally tested grades and subjects and into higher grades is challenging.
  - The degree of within school sorting of students and teachers likely increases (Anderson and Harris, 2013; Jackson, forthcoming)
  - **Unlike standardized tests in grades 3-8, tests in higher grades are not universally administered the timing of the test is endogenous.**

Contributions

- We develop a complete procedure for incorporating EOCs into educational evaluations.
  - We (try to) fully internalize the regulator's problem.
    - This work originated from a request from the Missouri Department of Elementary and Secondary Education (DESE) to help them incorporate EOCs into the district evaluation system in Missouri.
- The primary contribution is to take direct account of the endogeniety of course-timing decisions, with the goal of incorporating EOCs into educational evaluation systems.
- Parents, students and district/school staff can all play a role in determining when students take EOCs.
  - District and school policies regarding when students take particular courses can meaningfully affect student achievement (Clotfelter et al., 2012a/b).
    - This makes accounting for course timing of independent interest to regulators as well as to educational administrators looking to improve performance.
    - Given the importance of course-timing in determining student achievement, course-timing effects must also be accounted for so as to not confound efforts to measure instructional effectiveness.

Contributions

- Our procedure can be adopted "off the shelf," but also allows for considerable flexibility in its implementation depending on regulator preferences/constraints.
  - In addition to providing useful performance measures, the procedure also offers diagnostic value.
    - It explicitly separates out the influence of course-timing policies and instructional effectiveness on student achievement.

- We apply the procedure to construct district-level performance measures, and it is straightforward to extend the framework to produce school-level performance measures as well. The procedure is indirectly informative about some aspects of teacher evaluations (although the within-school sorting problem – the key problem for teacher-level evaluations based on EOCs – is beyond the scope of our study).

- We use algebra-I EOCs in the empirical work, but the procedure is constructed to be generalizable for use with other EOCs (subject to some adjustments).

Procedural Overview

- Step 1: Estimate a model of student EOC performance, conditional on the grade-level in which the EOC is administered.
  - This step produces measures of "instructional effectiveness" for school districts (more generally, educational units) unconfounded by course-timing effects.

- Step 2: Identify the effects of differences in course-timing policies on student achievement, then adjust the performance measures from step-1 to penalize districts for students who take EOCs at the wrong times (i.e., too early or too late).

- Step 3: Introduce "thoughtful" flexibility in terms of the extent to which the course-timing penalties are imposed.
  - Offsets available for high-achieving students who take the EOC at a time that would be too early for most students; also for low-achieving students who take the EOC too late.
    - Our recommended offsets are based on (admittedly thin) available research (Clotfelter et al., 2012a/b).

- The procedure can also be modified to build in adjustments for students who never take the exam (typically not an issue for standardized tests in grades 3-8, at least not to the same extent as for EOCs).
  - For some EOCs we expect nearly all students to take the exam (e.g., algebra-I, English-I); for others this is not the case (e.g., algebra-II).

Step 1:
Measures of Instructional Effectiveness

- We use a "proportional" two-step model (following Ehlert et a., 2013) to produce initial performance measures for school districts. We estimate the following model of EOC performance separately by grade-level:

- Next we regress the residuals on district indicator variables to obtain proportional district performance measures:

- We have a separate paper that talks about the benefits of proportionality in educational evaluations (Ehlert et al., 2013), but whether proportionality is a desirable property in the first-step of our procedure is not the focus of this paper.
  - The first-step model can take whatever form the researcher/regulator would like.
  - The key substantive feature of the model is that it is estimated separately by grade level so as to not confound course-timing effects with "instructional effectiveness" effects.

Step 2:
Incorporating the Effects of Course Timing

- Clotfelter et al. (2012a, 2012b) show that district policies regarding the grade-level placement of students into algebra-I can significantly affect exam performance and longer-term outcomes such as future course taking.
  - Course timing effects may also be important for other EOCs – there is no empirical evidence on this question to the best of our knowledge.
- Accounting for course-timing effects is important for two reasons:
  - It is of independent interest given that regulators will generally want to promote policies that are good for students, including course-timing policies.
  - We do not want course-timing effects to confound our ability to measure other aspects of educational performance (e.g., effective teaching).
- Course timing is, of course, endogenous.
  - Clotfelter et al. (2012a) show that simple OLS regressions of algebra-I test scores on the usual controls (including lagged performance on standardized tests) and indicators for course-timing produce a positive estimate for the "effect" of taking the course early. Using a natural experiment to provide exogenous variation, they show that the actual effect is negative and quite large.

Step 2:
Incorporating the Effects of Course Timing

- We use an instrumental variables strategy akin to the strategy used by Clotfelter et al. (2012b) to identify the effects of course-timing on achievement (details on next slide).

- We then adjust the student residuals from the first-step regression *ex post* to explicitly account for course-timing effects in the construction of districts' total evaluation measures.
  - Penalize districts for suboptimal course-timing policies.

- We do not need to produce the precise causal effects of course timing for our approach to be useful for educational evaluations.
  - It is important that our course-timing effects are properly signed.
  - Based on numerous interactions with policymakers, they do not seem to place equal weight on the costs associated with mistakes in different directions.
    - If our course-timing effect estimates are going to be wrong, we should error on the side of producing attenuated penalty parameters (i.e., give the benefit of the doubt to practitioners on the ground).

Step 2:
Incorporating the Effects of Course Timing

- We use a simple IV approach.
  - In the first stage we regress indicators for course timing (grade<=8, grade=9, grade=10, grade>=11) on exogenous regressors.
    - The instruments for individuals' course-timing outcomes are district-level enrollment shares by grade level.
- The second stage estimates of $\delta 5$, under some conditions, can be interpreted as the causal effects of course-timing on achievement.

Step 2:
Incorporating the Effects of Course Timing

Table 2: Grade Distribution of the 2012 Algebra I EOC Exam

Grade	Missing	< 7	7	8	9	10	11	12
No. of Students	13	57	694	14488	32124	11634	5090	4521
Percent of Students	**0.02**	**0.08**	**1.01**	**21.11**	**46.81**	**16.95**	**7.42**	**6.59**

Step 2:
Incorporating the Effects of Course Timing

Table 3: Grade Level Coefficients from Pooled Grade-Level Models

	OLS	IV
Grades 7 and 8	**0.134****	**-0.220****
	(0.008)	**(0.026)**
Grade 10	**-0.174****	**0.040**
	(0.008)	**(0.032)**
Grades 11 and 12	**-0.491****	**-0.175****
	(0.011)	**(0.039)**
Student-Level Controls		
Grade-4, 5, and 6 Exam Scores (Both Subjects)	X	X
Missing Exam Score Indicator Variables	X	X
Demographics	X	X
District-Level Aggregates of Student-Level Controls	X	X

Note: † represents significance at the 0.1 level, * at the 0.05 level, and ** at the 0.01 level.

Step 2:
Incorporating the Effects of Course Timing

- Our instrument may not pass muster if we need to convince you that we have identified the precise, causal effect of course-timing on achievement. But that is not the goal – the goal is to provide useful information for performance evaluations.
  - For accelerated algebra, our estimate is same-signed (negative) but smaller than the estimates from Clotfelter et al. (2012a/b). Difference could be due to:
    1. Some lingering endogeneity (e.g., unobservably better districts may accelerate algebra conditional on the observable information in the model)
    2. The difference between identifying the effect of an abrupt policy shift to accelerate algebra-I course-taking (Clotflelter et al.) and identifying the effect of "steady-state" policy differences
  - Given policymakers concerns about over-penalizing districts, the fact that our estimate is smaller is palatable, and it will still be useful for providing proper incentives (directionally).
    - The process of evaluating the usefulness of IV methods when exact exogeneity of the instruments is violated is discussed in Conley, Hansen and Rossi (2012). "Plausible exogenous" instruments can still be useful.

Step 2:
Incorporating the Effects of Course Timing

- For the penalties associated with delayed algebra-I course taking, our approach is less compelling for two reasons:

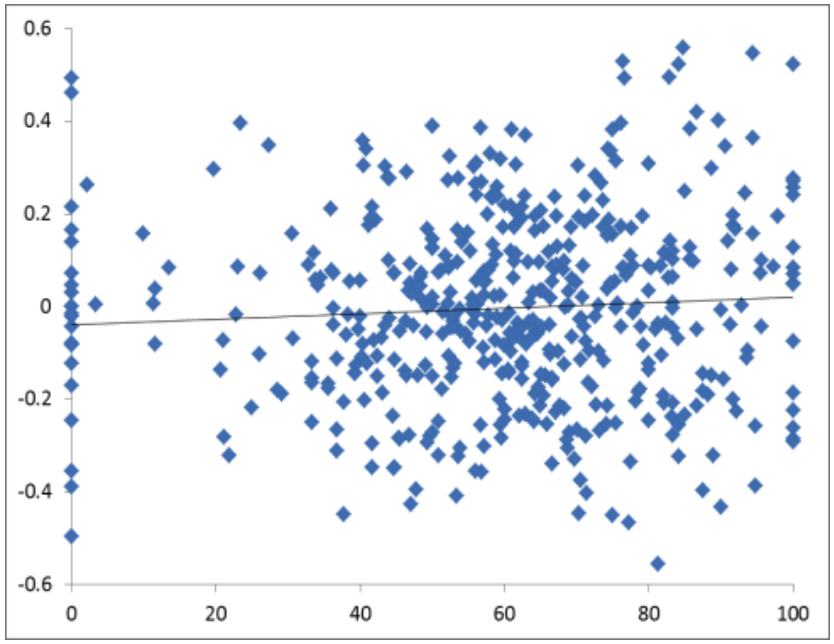
1. If less effective districts along unobserved dimensions have more delayed course taking, this will bias our penalty parameters *away* from zero (regulators may be uncomfortable with this).
2. Unlike the case for accelerated algebra, we are not aware of any research evidence that we can use to externally validate our findings.

- Nonetheless, in the paper we impose the delayed course taking penalty as estimated by the IV model.
  - This is the right thing to do if the instruments are valid, but we cannot verify validity.
  - Regulators could easily modify the approach if this is deemed undesirable.
    - Could impose no penalties for late course taking (lack of research evidence may imply no course of action at present).
    - Could moderate penalties for late course taking (perhaps considerably).

- The penalty for delay is less important empirically because accelerated course taking (at least for algebra-I) is much more prevalent

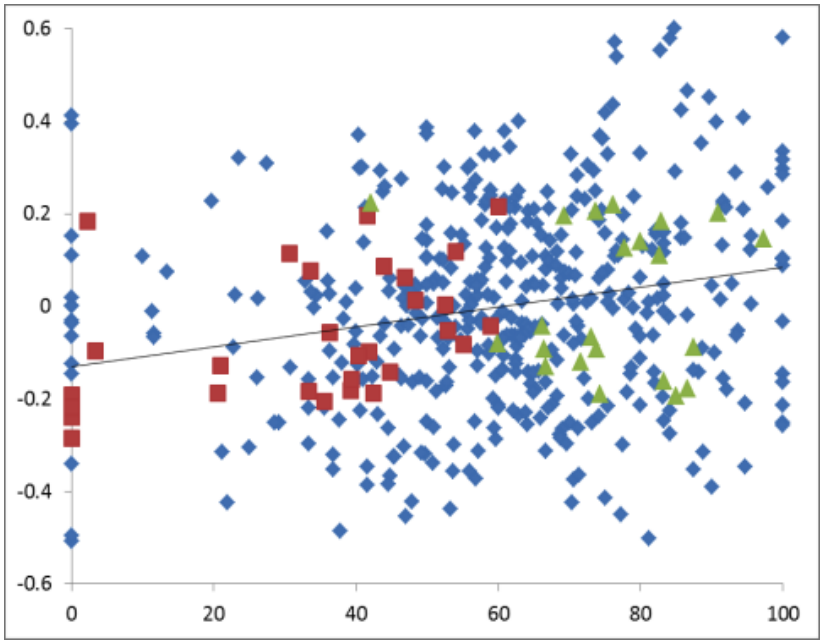
Step 2:
Incorporating the Effects of Course Timing

Unadjusted



Correlation: 0.065 (p-value ≈ 0.16)

After Adjustments Based on Course Timing



Correlation: 0.223 (p-value < 0.01)

Horizontal axis: Districts ordered by the share of students who take the algebra-I EOC in grades 9 and 10

Vertical axis: District performance measures

Step 3:
Allowing for Practitioner Discretion

- The achievement and longer-term effects of course timing may vary across the student population.
  - Clotfelter et al. (2012b) find that although students from every quintile of the prior-achievement distribution have lower algebra-I EOC scores if they take the course in grade-8 or before, students in the top quintile are more likely to pass geometry by grade-11 if they accelerate their algebra coursework.

- Ideally the IV approach could be extended to account for heterogeneous effects, but in practice this seems to be too demanding of the data. Noisy, non-monotonic "penalty" parameters (across the baseline achievement distribution) are identified if we try to build in heterogeneity in course-timing effects into the IV model.
  - Practical alternative: Based on the evidence from Clotfelter et al. (2012b), we impose "penalty forgiveness" for high-achieving students who take algebra-I early, and for low-achieving students who take algebra-I late.
    - There is considerable room for flexibility in our approach – in the paper we impose symmetric forgiveness for students in the top and bottom 20 percent of the achievement distribution.

Step 4:
Students Who Never Take the Test

- A non-negligible number of students do not take EOCs, even in "required" subjects like algebra-I.
  - In the appendix we discuss a procedure by which these students can be included into the evaluation. Including these students can help to align incentives for districts so that they will not discourage students from taking the exam.
  - This can get trickier for EOCs that are not required; but again, our approach can be used as a general point of departure for dealing with students who never take the exam.

Concluding Remarks

- Taking the concept of performance-based educational evaluations from research to policy application is challenging.
  - Researchers (often for good reason) typically circumvent difficult decisions that regulators must make.

- We come at the problem of incorporating EOC performance into district performance evaluations from the perspective of the regulator.
  - Our approach can be readily applied to school-level evaluations, which we discuss in an extension section.
  - It also helps to clarify the value of separating out course-timing effects in teacher evaluations, although we do not take on what is perhaps the most challenging aspect of teacher evaluations in higher grades (sorting)

- We develop a complete procedure that can be used by regulators "off the shelf" or, more realistically, can be used as a point of departure for intelligent discussions about how this can be done.

- Our approach is illustrated using algebra-I EOCs and district evaluations, but is constructed to be generalizable to other EOCs with proper background work.