WORKING PAPER

Controlling for Changes in Test Conditions when Estimating Education Intervention Effects

Matthew D. Baird and John F. Pane

RAND Education

WR-1245-BMGF

May 2018

Funded in part by the Bill and Melinda Gates Foundation

For more information on this publication, visit www.rand.org/pubs/working_papers/WR1245.html

Published by the RAND Corporation, Santa Monica, Calif.

Abstract

Education evaluations that analyze student achievement typically use baseline scores to control for unobserved ability. These analyses rely on the assumption that the conditions under which the test was administered are consistent from pretest to posttest. We caution evaluators to examine this assumption, provide a framework for evaluation under changes in test conditions, and demonstrate an application of this framework. We examine a case in which treatment status is correlated with the changes in test durations. We test three estimation strategies to account for this, each finding bias in estimation of treatment effects when not accounting for the changes in test conditions. We also discuss what to do in absence of such information.

Keywords: Evaluation, Economics of Education, Program Evaluation, Testing

1. Introduction

Evaluations of education interventions, teacher effectiveness, and school performance assessments make use of student examinations to measure student achievement growth. In addition to using a posttest as a study outcome, a pretest is typically used as a control for unobserved heterogeneity in student ability. The conditions under which the test is administered are implicitly assumed to stay constant, both from pretest to posttest and between treatment and control students. However, in most studies, these testing conditions are rarely if ever examined, let alone accounted for in the estimation procedure. This may be because many tests, such as state accountability tests, are administered under centrally imposed, highly controlled conditions, such as strict time limits or prohibitions against access to reference materials or calculators. However, there is an increased prevalence of other tests researchers might desire to use for their studies, and these may be administered under more flexible conditions. If conditions change from pretest to posttest in a way that is correlated with the treatment status of the students, any empirical analysis that does not account for these changes will yield biased results. In this paper, we caution evaluators to question the assumption of constant test conditions, provide a framework for evaluation under changes in test conditions, and demonstrate an application of this framework.

To do so, we examine a setting where there was an educational intervention in certain schools, and an adaptive online assessment is used in the treated schools and a collection of control schools across the United States. The tests were administered without centrally imposed time limits, although the actual test duration is documented for each test event. The assessments also lack strong central guidance on other aspects of testing conditions (e.g., the use of calculators) but those conditions are not documented in test records. We use test duration as an

example of variation in testing conditions, to demonstrate situations where ignoring such variation can yield biased results in empirical analysis. We provide a framework under which researchers can consider the problem and possible solutions. The primary challenge is in separating growth in test duration related to ability gains and to growth related to changed testing conditions. Our strategies, and in particular the 2SLS estimators, address this issue.

In our setting, we find strong reasons to be concerned about the role duration plays in the estimated treatment effects. This paper focuses on duration as symptomatic of changes in test conditions because it is observed in our data; however, potential solutions we offer are more generally applicable to any problematic test condition that can be measured. We examine the following research questions. (1) Do test conditions affect impact evaluations? In particular, do changes in conditions from pretest to posttest that are correlated with treatment status influence the impact estimates? (2) If changes in conditions do matter, how can researchers account for the systematic differences across study conditions in order to find unbiased program effect estimates? (3) Can the treatment effects be decomposed into the various elements of interest, such as the direct effect of treatment on achievement versus the effect of testing conditions on apparent achievement? Can indirect effects, where the intervention affects how students capitalize on flexibility in conditions, also be separately estimated? (4) What lessons can be drawn for program evaluation, both when measures of test conditions are available and when they are not? We explore three methodologies for addressing the problem: filtering out students or schools with anomalous test conditions, investigating alternative spans, and implementing instrumental variables to solve a system of equations that describe the relationship between student achievement and test conditions.

In order to investigate these research questions, we use data from an educational intervention that occurred in nearly 100 schools across the United States. The intervention, described in detail in Pane et al. 2015, involved schools implementing personalized learning (PL). The data comes from computer-administered evaluations using the Measures of Academic Progress (MAP) assessment from the Northwest Evaluation Association (NWEA) in the fall and spring of two academic years. The purpose of the evaluation is to examine whether PL changed achievement trajectories of students. MAP is an adaptive online assessment that can efficiently determine accurate scores across a wide range of abilities and place them on a continuous development scale from kindergarten through 10th grade. MAP may be administered up to three times each school year (fall, winter, spring); in our analysis, the test is administered at least in the fall (baseline or pretest) and spring (outcome or posttest). There is no control group in our study, so NWEA provided us with data from its national database of MAP assessment records of comparison students that are similar on observables, forming a Virtual Comparison Group (VCG) of up to 51 comparison students for each treated student. Pane et al. 2015 provide further details regarding the data and empirical strategy for using the MAP data.

The properties of the MAP assessment and the availability of NWEA's VCG service make the data appealing for studies in which researchers do not have more direct access to sites or common measures for a comparison group. Moreover, researchers may increasingly rely on this and similar facilities as changes to federal regulations and state testing requirements, along with increasing student opt-out of state tests, make state tests a less viable data source for research and evaluation.

While the availability of fall pretest scores and spring posttest scores are attractive for many reasons—including the ability to abstract from possible summer learning loss—it does

present a challenge if schools use the fall and spring tests for different purposes and that leads to differences in testing conditions. Generally speaking, schools elect to use MAP for a variety of purposes. Formative purposes may include start-of-year placement to guide teachers' instruction, and mid or end of year testing for measurement of progress towards meeting standards. Summative reasons may include using end-of-year testing to evaluate school, teacher, or program performance, or possibly as part of course grades for students. These varying purposes may lead to variation in testing conditions both within and across schools. These may include implicit or explicit pressures on students or educators, i.e. pressures to do well on the spring administrations of MAP that are not present in the fall. NWEA does not provide strong guidance on testing time (possibly to accommodate for the varying uses of MAP), but do say measures of growth should use pretests and posttest taken under similar conditions, and provide rough guidance on typical durations. However, in investigating the duration trends and in discussions with some treated schools, it became clear that at least in certain cases test conditions varied substantially from the fall to spring.

We are not the first to suggest that manipulation of student test conditions may bias the measure of student achievement on tests. Haladyna, Nolen, and Haas (1991) discuss this issue with several potential forms of "polluting practices", and encourage standardization of the test conditions. Other more recent studies have noted this potential phenomenon as well, including Pearson, Valencia, and Wixson (2014) and Kesler (2013). Welsh, Eastwood, and D'Agostino (2014) explore the relationship between test preparation instruction and outcomes, and find no relationship. However, to our knowledge, no research has explored how to account for variation in test conditions in estimating treatment effects, nor has any focused on test duration in

particular. We contribute to the literature by presenting methodologies to account for changes in test conditions, applying them on real-world data, and discussing the implications for analysis.

The paper proceeds as follows. Section 2 presents the underlying model as well as three empirical strategies for accounting for the change in testing conditions to yield estimators of interest. Section 3 discusses the MAP data in more detail and presents the overall duration trends that encouraged this analysis. Section 4 presents the results of the three strategies, Section 5 discusses these results, and Section 6 concludes.

2. Model

We consider the case of evaluating the effects of an educational intervention on student outcomes, when changes in the testing conditions themselves might have independent effects on achievement scores unassociated with learning. We desire to separate out these testing condition effects from the direct effects (and indirect effects, as will be explained) in order to understand the impact of the intervention holding testing conditions constant (unless changing the testing condition is the intervention itself). To that end, we consider the following model. For this exercise, we will consider everything in growth models instead of including the baseline score as a regressor. We will also, for simplicity, assume linear functions. We assume two kinds of ability: content understanding (U), and perseverance on difficult tasks, or grit (G). The two types of ability have differing effects on scores and duration. The change in content understanding for student i in VCG group v is a function of treatment status T; student and school characteristics X; and (for our model) VCG fixed effects ψ_v, one for each treated student (and all but one included in the regression), which control for other student and school factors that might affect growth in scores. Equation 1 presents this model.

$$\Delta U_{iv} = \alpha_U T_i + X_i^U \beta_U + \psi_v^U + \varepsilon_i^U \quad (1)$$

The change in grit has a similar format. Here we assume that, while students have differing innate levels of grit, it can further be developed (declined) through schooling; grit is assumed to be a function of the same factors, although with coefficients we would expect to differ from equation 1.

$$\Delta G_{iv} = \alpha_G T_i + X_i^G \beta_G + \psi_v^G + \varepsilon_i^G \quad (2)$$

Understanding is not directly observed; instead, noisy measures are observed through testing, which yields observed test scores S_{iv}. Thus, change in test scores has the potential to shed light on the change in ability. But the change in test scores may also be driven by changes in test conditions, such as the duration of the test. Equation 3 defines this relationship.

$$\Delta S_{iv} = \Delta U_{iv} + \gamma_D \Delta D_{iv} + \varepsilon_i^S \quad (3)$$

In addition to the change in understanding and other unobserved factors contained in ε_i^S (weather, how well the student slept, random distracting events, etc.), there is an additional factor *D* that affects score growth (but *not* growth in understanding) and which may be directly affected by treatment (i.e., through changes in testing conditions). In our paper, we will focus on changes in test duration. Our underlying hypothesis is that students that take their time or are given more time, even if they have the same change in ability, will observe larger increases in test scores, although the model does not impose this restriction.

Combining equations 1 and 3 into equation 4, we have the change in scores as a function of treatment, changes in duration, student and school characteristics, and unobservable factors.

$$\Delta S_{iv} = \alpha_U T_i + X_i^U \beta_U + \gamma_D \Delta D_{iv} + \psi_v^U + \varepsilon_i^U + \varepsilon_i^S \quad (4)$$

We then need to model the change in duration. We do so in equation 5.

$$\Delta D_{iv} = \alpha_D T_i + X_i^D \beta_D + \gamma_U \Delta U_{iv} + \gamma_G \Delta G_{iv} + \psi_v^D + \varepsilon_i^D \quad (5)$$

The change in duration is allowed to depend on treatment status, as well as by other student and school observable characteristics in *X*. In addition, changes in grit and understanding may affect change in test duration. Increased grit may cause longer durations as students focus and endure better, while increases in understanding may lead to a faster ability to comprehend and answer the problems, leading to shorter durations (though, again, the model does not impose a direction on these effects).

In the current form, we are unable to estimate equation 5 because of the presence of the unobserved changes in understanding and grit. We may substitute this out in more than one way. For our purposes, we will pursue the following substitution from equations 1 and 2:

$$\begin{aligned} \Delta D_{iv} = {} & \alpha_D T_i + X_D \beta_D + \gamma_U \left(\alpha_U T_i + X_i^U \beta_U + \psi_v^U\right) + \gamma_G (\alpha_G T + X_G \beta_G + \psi_v^G) \\ & + \varepsilon_i^D + \gamma_U \varepsilon_i^U + \gamma_G \varepsilon_i^G \end{aligned} \quad (6)$$

The system of equations of interest is given by equations 4 and 6. The full treatment effect on test scores is the expected change in score growth for being in the treatment group, and is given by equation 7. For our purposes, we are interested in four different effects that compose the full effect of treatment status: the direct effect, the indirect effect through changes in grit, the indirect effect through changes in understanding, and the test conditions effect. The direct

treatment effect, given in equation 8, is the change in score growth holding the duration effect on score growth constant. It represents our typical concept of an education treatment effect, where a change in policy affects understanding directly, which then affects scores. The indirect treatment effect through grit, given by equation 9, is the effect of treatment on grit, which changes duration, which then affects score growth. The indirect treatment effect through understanding (equation 10) measures how an increase in understanding affects scores, holding the direct effect of understanding constant. In particular, increased comprehension may decrease duration, which may then (holding understanding constant) actually decrease scores. Practically speaking, this could arise if a student with strong understanding of the content moves quickly through the test, becoming more susceptible to oversights or mistakes. The test conditions effect (equation 11) shows how differences in testing conditions between the treatment and control groups can affect score growth, holding constant the effects of treatment on ability.

Full Effect on Test Scores

$$E[\Delta S|T=1] - E[\Delta S|T=0] = \alpha_U + \gamma_D(\alpha_D + \gamma_G\alpha_G + \gamma_U\alpha_U) \tag{7}$$

Direct Treatment Effect on Test Scores:

$$E[\Delta S|T=1, \Delta D=d] - E[\Delta S|T=0, \Delta D=d] = \alpha_U \tag{8}$$

Indirect Grit Treatment Effect:

$$E[\Delta S|T=1, \alpha_D=0, \alpha_U=0] - E[\Delta S|T=0, \alpha_D=0, \alpha_U=0] = \gamma_D\gamma_G\alpha_G \tag{9}$$

Indirect Understanding Treatment Effect:

$$E[\Delta S|T=1, \alpha_D=0, \alpha_G=0] - E[\Delta S|T=0, \alpha_D=0, \alpha_G=0] = \gamma_D\gamma_U\alpha_U \tag{10}$$

Test Conditions Effect:

$$E[\Delta S|T=1, \alpha_U=0, \alpha_G=0] - E[\Delta S|T=0, \alpha_U=0, \alpha_G=0] = \gamma_D\alpha_D \tag{11}$$

We primarily want to estimate a treatment effect that is the direct effect on test scores, uncontaminated by a test conditions effect. Even better would be to estimate the full effect on test scores, and decompose the overall treatment effect into all four components.

Typical evaluations would estimate a simplified version of equation 4, omitting the contribution of duration change on score growth. If treatment status is independent of changes in duration conditional on the changes on grit and understanding (i.e., the effect of the treatment on duration is realized only through the effect of the treatment on grit and understanding), then such a procedure would estimate a net effect of treatment on achievement growth that combines the direct effect as well as the two indirect effects through grit and understanding. In this case, duration can be viewed as an orthogonal omitted variable, which would then not cause bias. In other words, effectively, α_D would be equal to zero, and the full effect (equation 7) would be composed of only the direct effect of changes in understanding on changes in scores and the two indirect effects of treatment on duration through changes in grit and understanding. However, there may be cases when the treatment status is correlated with changes in test conditions. Such a case may happen if the treated schools are more likely to change testing conditions from the pretest to the posttest than control schools. For example, treated schools may have different incentive structures that make the pretest not consequential but the posttest highly consequential in a way not relevant to the control schools. We observe a correlation between change in test durations and treatment. In such a case, a simple regression of score growth on treatment status and controls while ignoring duration will lead to a net effect which also includes the test conditions effect, which should not be attributed as part of the intervention effect. Including the change in duration in the regression does not resolve the issue either, as there is endogeneity in

the regressors due to the correlation between change in duration and ε_i^U and ε_i^G. Larger shocks to the change in ability will feedback to increase duration (equation 4).

We will consider three approaches to estimating the net treatment effect that is the direct plus indirect effects, stripped of the test conditions effect: filtering, alternative spans, and instrumental variables.

2.1. Filtering Methodology

With filtering, we are making the assumption that α_D is large (giving us reason to be concerned with the typical estimator), and significantly larger than the indirect grit and understanding treatment effects. If those assumptions are true, then we can impose various filters to select out students that have much higher than normal growth in duration. Given change in test conditions will likely affect all students within a classroom, we also consider filtering out classrooms that have anomalous growth in duration. Unfortunately, in our data we are unable to observe classroom assignments, so we instead consider school-level filters.

Student Filter 1: Drop if fall or spring test durations are below 5th percentile or above 95th percentile for grade and subject (national duration, provided in personal communication by NWEA)

Student Filter 2: Drop if the change in test duration from fall to spring exceeds the national 90th percentile of change in test duration for grade and subject.[1]

Student Filter 3: Drop if the durations meet the criteria of both filter 1 and filter 2.

[1] The 90th percentile change in duration for fall to spring in the same academic year is available at https://public.tableau.com/profile/john.cronin.nwea#!/vizhome/testdurationworkbook/Dashboard1.

Given the specific nature of our data described above, if a treated student met a filter's criteria, all of the VCG records for that student were also filtered out. However, if a VCG student was filtered we did not drop the corresponding treated student, nor other VCG records that did not meet filter criteria.

We use two methods to filter out schools, for each of the three filters:

School Filter Aggregation 1: Calculate average durations by subject and grade for all students in the school and filter out the school if a given filter criteria are met.

School Filter Aggregation 2: Filter out a school if over 40% of students in that school meet filter criteria.

2.2. Alternative Spans Methodology

Many of our concerns arise from the fact that the pretest is in the fall and the posttest in the spring, and that testing conditions vary between those two administrations. However, year-to-year conditions may be stable for a particular time of year (e.g., fall). With multi-year data, we can estimate treatment effects using time spans other than fall to spring. For example, with a two-year span data of Fall 2013 to Spring 2015, we can compare estimates of Fall 2014-Spring 2015 to Spring 2014-Spring 2015. If the treatment effect is specific to fall-to-spring duration changes, then for these alternative spans, α_D should be equal to zero and we can perform our typical regression. This is reinforced by the fact that we find large differences in test duration between fall and spring for the treated students, with fall durations typically shorter than spring durations, but such a relationship does not persist fall-to-fall or spring-to-spring. Therefore, using fall-to-fall or spring-to-spring timespans alleviates the issue.

However, there are potential problems with these alternatives. First, they include summer, and students often experience test score declines over the summer. If summer declines are an outgrowth of differences in testing conditions and not related to actual learning, then including summer may result in a more accurate measure of learning during the school year because the pretest and posttest are administered under more similar conditions. However, it may be that some of this summer decline is true loss of achievement that accrued the prior school year. Depending on the span used, these losses are attributed to the schools and practices prior to summer (fall-to-fall) or to the schools and practices after the summer (spring-to-spring), which may be more problematic. Moreover, if we believe that the fall or spring test durations are so short or so long as to result in invalid scores, these alternative durations may also suffer from the same problem.

For spring-to-spring, an additional potential complication is if most of the growth and effect happens in the first year of exposure to the school or to the intervention, then this will be missed by not starting from a baseline fall score. Some more technical complications also arise.[2]

2.3. Instrumental Variables (IV) Methodology

In order to consistently estimate the parameters and do a full decomposition, we must find an instrument for duration that is correlated with change in duration and only affects change in scores through change in duration. It may be correlated with treatment, but cannot be

[2] When we use spring pretests, the students are not matched to their VCGs on this pseudo-baseline. To account for this, we also evaluate a treatment effect where we drop all VCGs not within 5.5 points on the RIT scale (approximately 95% of VCGs are within +/- 5.5 points of the PL student's score on the interim spring test, while an even higher proportion of VCGs are within +/- 5.5 for the true baselines on which they were matched). Moreover, for the analysis of the 2014-15 score growth, the alternative span data comes from 2013-2015, with the VCGs matched on fall 2013 baselines. Thus, the VCGs for this dataset are different than for the 2014-15 dataset, although given they still match at baseline they should be similar.

correlated with ε_i^D or ε_i^S. Effectively, we are looking for variables that are in X_i^D but not in X_i^U. In this paper, we consider the time of day they begin taking the test as such an instrument. The instrument works on the theory that students taking the test later in the day will have less time to take the test. However, we are concerned with major differences in time of day having a direct effect on scores (for example, becoming drowsy after lunch). For that reason, we only consider students that start taking the test between 7 and 11:30 AM, which assumes that students taking the test closer to lunch will face shortened time in which they can take the test. The endogenous variable is the change in duration, and we instrument with two variables, the times of day the pretest and posttest start.

After estimating equation 4 using instrumental variables and recovering the parameters, we then proceed to estimate equation 5. We create an estimator for ΔU given by $\widehat{\alpha_U} T_i + X_i^U \widehat{\beta_U} + \widehat{\psi_v^U}$, where $\widehat{\psi_v^U} = \frac{1}{N_v} \sum_v \left(\Delta S_i - \widehat{\alpha_U} T_i - X_i^U \widehat{\beta_U} - \widehat{\gamma_D} \Delta D_i \right)$.

We then estimate equation 6. Unfortunately, we currently do not have data that measures the grit of the student. As such, we are unable to separate out the test conditions effect from the grit effect. They are observationally identical in our data, as they both capture how duration changes, holding understanding constant (captured by the change in scores holding duration constant). In order to separate them out, a researcher would need a measure of grit from which they can estimate equation 2, after which they can follow a similar approach to the above paragraph. Note that our use of the predicted understanding growth based on treatment status and observables that affect understanding, including variables that only affect understanding and not duration, serves as a type of two stage least squares regression. In this paper, for observed variables that affect changes in understanding and grit, but not duration, we use the number of

elapsed days between the testing periods (more days to acquire human capital), and the fraction of the school that is free or reduced lunch program-eligible. This assumes that the fraction of students eligible for free or reduced lunch affects growth in understanding (perhaps, for example, but having fewer educational resources or worse teachers), but does not affect how long they had available on the test, and thus only affects changes in test duration through changes in ability. This assumption is one we are unable to directly test.

With these models estimated, we may calculate the direct effect of treatment on understanding, the indirect effect through understanding, and the sum of the indirect treatment effect through grit and the test conditions effect.

3. Data

We use data from the MAP assessments. We focus on the 2014-15 academic year, which has all of the available data that we need. Each of the methodologies requires different data and thus provides a different sample of students for whom we have the necessary data and that meet the relevant restrictions. For consistency of analysis, we conduct the analysis on the intersection samples of the three methodologies. Table 1 provides the sample sizes of each methodology as well as the average number of VCG control students per treated student.

Table 1: Sample Sizes

	Number of treated students		Average number of control (VCG) students per treated student	
	Math	Reading	Math	Reading
All available students	24,610	24,076	44.5	45.0
Filter analysis	24,610	24,076	44.5	45.0
Span analysis*	11,339	11,057	14.1	13.4
2SLS (IV) analysis**	8,845	8,903	12.1	13.6
Intersection of sets	4,345	4,373	12.1	13.6

*Primary source of dropped obs.: not being tested in the school in the 2013-2014 school year

**Primary source of dropped obs.: not taking the test in the morning in the fall and spring

The span analysis has fewer students because of the necessity to have the student also in our data the previous year; the 2SLS (or IV) analysis has fewer students because of our restriction to only take students who start their exam between 7 and 11:30 AM. The final empirical data set, which is the intersection of the sets, has over 4,000 treated students with an average of over 12 control students for each treated student, which will give us enough students to evaluate while still having a consistent sample across methodologies.

Table 2 presents the summary statistics for the intersection set, which is our empirical sample. There is good balance across most variables, with the exception of the test durations, and in particular the spring test duration. While our analysis uses standardized test scores, here we report the raw test scores. The treated students score almost identical to the VCGs in the fall and outperform them in the spring.

Figure 1 presents the duration statistics graphically to emphasize the differences between the treated students and the VCG students. Treatment and VCG students have similar average fall durations, and test duration increased in both groups on average from pretest to posttest, but treatment students have noticeably longer test durations in the posttest, aligned with them having significantly larger changes in duration.

This relationship is driven by a subset of schools. Figure 2 presents the school-level average change in fall to spring score durations on the x-axis and the same for the VCG of those students in each school. Schools near the 45-degree line have growth in duration for the treated students similar to that for the control students. The schools far to the right of the 45-degree line are the treated schools driving the differences in Figure 1.

Schools offer various reasons for the anomalous changes in test duration. Some claim that part of their instruction is teaching students how to be more persistent in taking tests. This would be accounted as a change in ability that then affects duration, and we would want to include this as part of the treatment effect. Other schools explained that the purpose of their fall testing is for a quick reading of ability for formative purposes, while spring testing is used for evaluative purposes so that the instructions given to students for taking the assessments varies from the fall to the spring. In one school, students were asked to write out all of their answers, but only in the spring. Some schools also claim that the spring tests are administered in close proximity to the high-stakes state assessments, for which students are given a lot of instruction on how to do their best. These schools argued that this behavior carried over to the MAP assessments. These reasons would not be considered treatment effects of interest, and would contaminate estimates with the test condition effect.

Table 2: Summary Statistics for Empirical Sample

	Math		Reading	
	Treated	VCG	Treated	VCG
Grade	4.62	4.69	4.63	4.65
	(2.54)	(2.52)	(2.60)	(2.55)
Male	0.49	0.49	0.49	0.49
	(0.50)	(0.50)	(0.50)	(0.50)
Fraction of School Free/Reduced Lunch	83.23	84.14	85.06	85.69
	(15.60)	(13.81)	(12.75)	(11.68)
Fall MAP RIT Score	203.38	203.41	196.06	196.01
	(26.28)	(25.43)	(24.15)	(23.48)
Spring MAP RIT Score	216.24	214.20	207.21	204.31
	(22.54)	(22.28)	(19.82)	(19.55)
Fall Test Duration	64.30	55.48	62.91	55.51
	(34.09)	(27.28)	(36.85)	(27.62)
Spring Test Duration	87.74	68.66	94.17	66.53
	(46.41)	(39.46)	(56.75)	(35.35)
Days Elapsed from Fall to Spring Exam	260.69	258.07	259.88	257.46
	(18.35)	(16.30)	(16.48)	(14.43)
Charter School	0.97	0.98	0.98	0.98
	(0.16)	(0.14)	(0.15)	(0.13)
Race: White	0.05	0.05	0.04	0.04
	(0.22)	(0.21)	(0.20)	(0.19)
Race: Black	0.20	0.18	0.18	0.15
	(0.40)	(0.38)	(0.38)	(0.36)
Race: Hispanic	0.50	0.53	0.52	0.55
	(0.50)	(0.50)	(0.50)	(0.50)
Race: Other/Missing	0.24	0.24	0.26	0.26
	(0.43)	(0.43)	(0.44)	(0.44)

Mean and standard deviation; standard deviation in parentheses
Grade and gender are exactly matched; differences reported here result from differences in number of VCGs across treated students of different genders and grades

Figure 1: Average test duration for pretest and posttest by treatment status

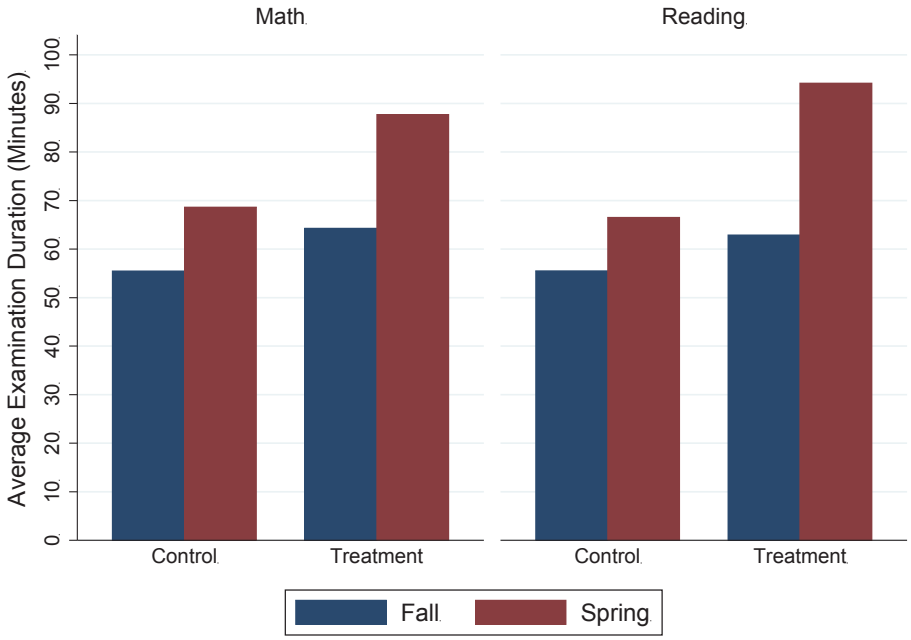
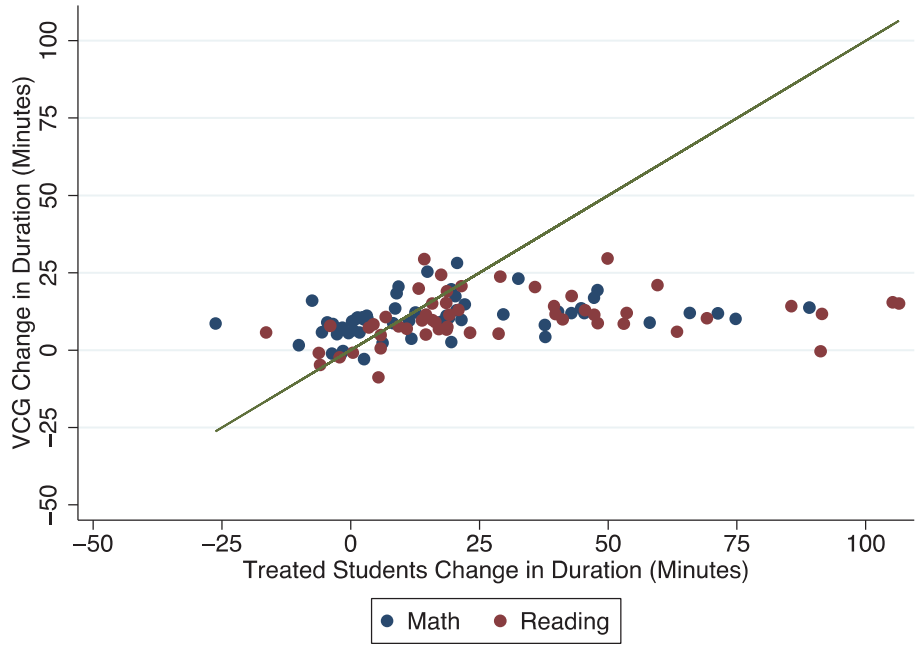
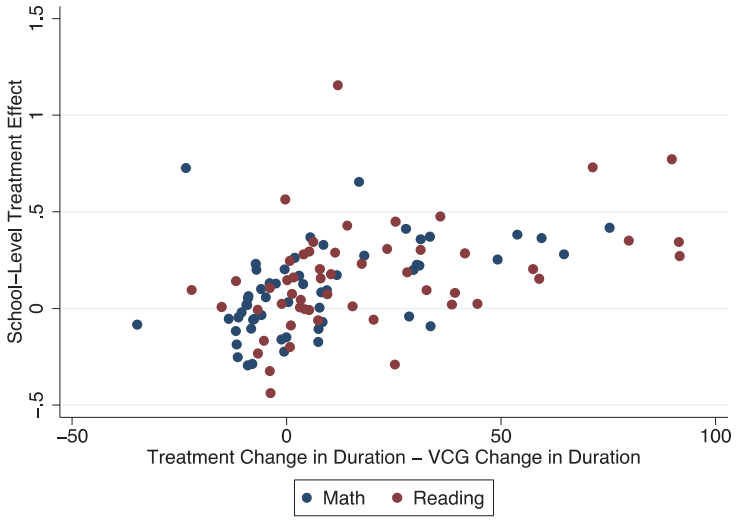


Figure 2: Change in duration, treatment vs. control



The difference in duration is correlated with treatment effects. As Figure 3 demonstrates, the gap between duration growth of treatment and control schools is also correlated with the estimated size of the school-level treatment effect. Schools that have larger difference in duration growth compared to their VCG students tend to also have larger estimated treatment effects. A simple regression of treatment effect size on relative change in duration has a coefficient of 0.005 (t-statistic 4.00) for math and 0.004 (t-statistic 3.34) for reading. This means that about a two-minute increase in the relative growth in duration, treatment vs. control, is associated with a 0.01 student standard deviation increase in the treatment effect. Given the mean treatment effect is around 0.1, this is a large change for only a few additional minutes. Based on this alone, it is unclear whether this is because these effective schools increases their students' ability and grit which translates to both higher change in duration as well as higher achievement, which we would want to include in the treatment effect measures, or whether it is driven by these schools changing testing conditions in a way that increases scores but not ability or grit, which we would not want to include in the treatment effect estimate which would be used to assess the effectiveness of the intervention.

Figure 3: Change in duration compared to treatment effect by school



4. Results

Table 3 presents the results from the regressions of standardized score growth on the treatment status, fixed effects for each treated student and their VCG control students, and some controls including elapsed days between the pretest and the posttest, school-level fraction of students eligible for free or reduced price lunch (FRPL), and switch test type which is an indicator for whether the student had a Common Core test for one examination and a non-Common Core test for the other. We find relatively large treatment effects of personalized learning on this subsample of students. We also find that the school-level FRPL fraction has a positive relationship with score growth in math but not in reading.

Table 3: Overall Results

	Math	Reading
Treatment	0.131***	0.176***
	(0.00874)	(0.00875)
Fraction School FRL	0.00284***	0.000126
	(0.000629)	(0.000609)
Elapsed Days	0.00157***	0.00268***
	(0.000498)	(0.000472)
Switch Test Type (CC)	0.0538	-0.170***
	(0.0662)	(0.0384)
Constant	0.112	-0.158
	(0.135)	(0.131)
Observations	56,323	63,118
R-squared	0.006	0.009
Number of Treated Students	4,316	4,354

Includes VCG group fixed effects, which in our sample exactly match on school grade, gender, race, charter status, urbanicity, and approximately match on school FRL (within 15 points), fall score (within 5 points on raw scale) and days elapsed (within 18 days)
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

4.1. Filtering Results

Table 4 presents the percent of students retained after each filter. Table 5 presents the percent of schools retained with the school-level filters. Recall that control (VCG) students are retained only if they are not filtered and their associated treatment student is not filtered. As anticipated, a greater fraction of treated students are filtered than control students, and by construction, filter 3 is the least restrictive.

Table 4: Fraction of Students Retained After Filter

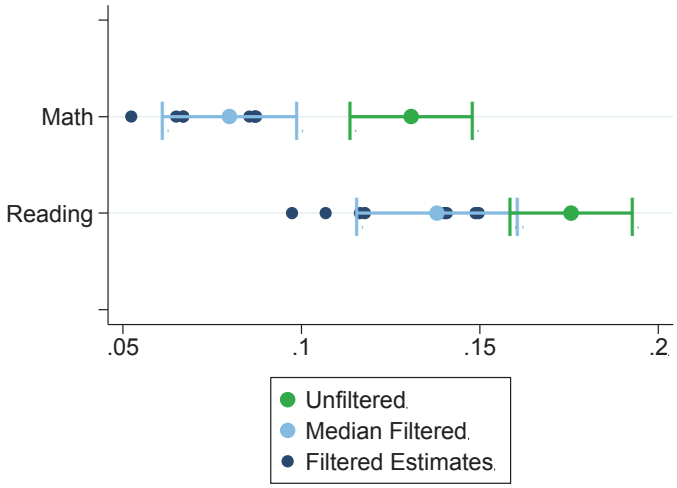
		Filter 1	Filter 2	Filter 3
Math	Control	67%	82%	85%
	Treatment	45%	67%	71%
Reading	Control	69%	82%	66%
	Treatment	46%	61%	66%

Table 5: Fraction of Schools Retained After Filter

	Average Student			Over 40% of Students		
	Filter 1	Filter 2	Filter 3	Filter 1	Filter 2	Filter 3
Math	73%	77%	80%	64%	73%	82%
Reading	72%	70%	75%	60%	66%	79%

Given we estimate effects with several filters, we decide here to report the range of estimates against the unfiltered estimate, along with the median filtered estimate and its confidence interval. Figure 4 reports these estimates. Every filtered estimate is smaller than the corresponding unfiltered estimate, but there is a wide range of estimates. None of the filtered estimates are negative and the median filtered estimate is still significantly different from zero and positive. Although filtering tends to decrease the estimates, they are still positive, suggesting that while some test duration anomalies may be biasing the effect size upwards, the direction and significance of the findings remain the same. Students exposed to personalized learning schooling environments are succeeding in test performance at a higher rate than similar students not exposed to PL. Across the two subjects, the average decrease in the treatment effect estimates from filtering is around 30%.

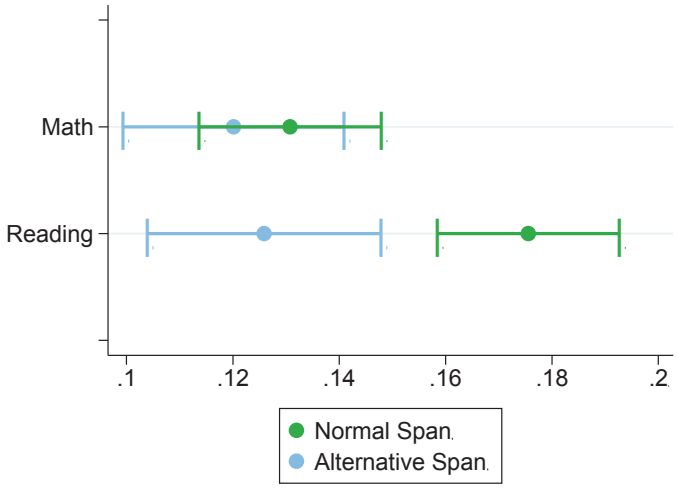
Figure 4: Comparison of Unfiltered and Filtered Treatment Effects



4.2. Alternative Spans Results

Figure 5 reports the treatment effects for the alternative spans.[3] We find decreases in the treatment effect estimates from the alternative span, averaging 18%.

Figure 5: Alternative Span Analysis



[3] For the alternative span of spring-to-spring, we need to subset the VCGs to those that are within 5.5 RIT points of the PL student on the spring (2014) pretest, in order to mimic the matching that NWEA provides for fall-to-spring.

4.3. Instrumental Variables Results

We next estimate the model using the instrumental variables strategy described above. The first stage F-stats are larger than 100 in each case, so we are unlikely to have weak instruments. Table 6 presents the parameter estimates along with 95% confidence intervals from bootstrapping. A few things are of note. First, α_U, the direct effect, is still positive and significantly different from zero for both subjects. $\alpha_D + \gamma_G \alpha_G$ is also universally positive and significantly different from zero for reading; for math, it is not statistically different from zero. However, both are of similar magnitude. The estimated γ_D is very small and statistically insignificant; suggesting that the effects of duration growth on score growth is near zero. This would imply that, even if there are large and systematic differences in the change in test duration between the treated students and the control students, these would not introduce bias in the estimated indirect effects because the seeming advantage of increased time to take the tests is not translating into larger gains in test scores. While the estimates of γ_U are not statistically significant for either subject, their direction suggest that increases in understanding increase duration (independent of grit), although the effect is small (a 0.1 standard deviation increase in understanding is associated with 40 seconds more growth in duration for math and about 3 minutes more growth in duration for reading).

Table 6: Model Parameter Estimates

Subject	α_U	$\alpha_D + \gamma_G \alpha_G$	γ_D	γ_U
Math	0.12	8.40	0.001	6.90
	(0.08,0.15)	(-2.30,17.34)	(-0.00,0.00)	(-79.03,81.64)
Reading	0.17	8.82	-0.0004	29.05
	(0.08,0.27)	(3.29,14.23)	(-0.01,0.01)	(-1.64,52.30)

Bootstrapped 95% Confidence intervals in parentheses

From Table 6 we can derive the treatment effects represented by equations 8-11, as shown in Table 7. The direct effects are a larger proportion of the overall treatment effects than we observe from the filtering or alternative spans, at least for reading. This is not substantially changed when we include the near-zero indirect effects of understanding. Finally, the combined effect of test condition and grit is small and close to zero, and not statistically significant.

Table 7: 2SLS Estimates of Effects

	Full Effect	Direct Effect of Understanding	Indirect Effect of Understanding	Direct + Indirect Understanding	Test Condition + Indirect Effect of Grit
Math	0.13 (0.11,0.14)	0.12 (0.08,0.15)	0.001 (-0.03,0.01)	0.12 (0.07,0.14)	0.01 (-0.01,0.06)
Reading	0.17 (0.15,0.18)	0.17 (0.08,0.27)	-0.002 (-0.07,0.02)	0.17 (0.09,0.21)	-0.003 (-0.05,0.07)

95% Confidence intervals in parentheses

We can more closely examine these results in one way by applying the same instrumental variable strategy on two subsets of schools: those that are filtered out above and those that are not filtered out. Of the many school-level filters, we use student filter 1 (fall and spring durations being larger than the national 5th percentile and smaller than the national 95th percentile) for school aggregation choice 2 (the average duration change in the school). We choose this because it splits the sample roughly in half, students in filtered schools and students in non-filtered schools. Tables 8 and 9 report these results. The first thing to note is that the test condition + grit effect is 2 to 4 times larger in the filtered schools, as we would anticipate. However, the filtered schools also have relatively strong direct effects. These together seem to suggest that the instrumental variables strategy is working properly, even if we are unable to separate out the grit

and test conditions effect. Note that there are cases where the derived effect estimates for the overall sample is not between the estimates for the two subgroups, filtered schools and non-filtered schools.

Table 8: 2SLS Parameters Subgroup Analysis by Filtered or Unfiltered School

	Filtered	α_U	$\alpha_D + \gamma_G \alpha_G$	γ_D	γ_U
Math	Yes	0.30 (0.21,0.41)	-13.63 (-45.20,25.34)	-0.004 (-0.01,0.00)	109.74 (-26.20,228.26)
	No	0.05 (0.02,0.07)	4.41 (1.05,7.54)	0.007 (0.00,0.01)	-51.62 (-96.17,4.80)
Reading	Yes	0.16 (-0.02,0.32)	25.35 (18.32,29.33)	0.003 (-0.00,0.01)	0.23 (-25.48,25.60)
	No	0.12 (0.02,0.23)	-2.66 (-10.53,6.62)	-0.006 (-0.04,0.02)	51.74 (-74.62,177.91)

Table 9: 2SLS Derived Effects Estimates Subgroup Analysis by Filtered or Unfiltered School

	Filtered	Full Effect	Direct Effect of Understanding	Indirect Effect of Understanding	Direct + Indirect Understanding	Test Condition + Indirect Effect of Grit
Math	Yes	0.22 (0.19,0.25)	0.30 (0.21,0.41)	-0.14 (-0.33,0.00)	0.16 (0.02,0.26)	0.06 (-0.04,0.19)
	No	0.06 (0.04,0.08)	0.05 (0.02,0.07)	-0.017 (-0.04,0.00)	0.03 (0.01,0.06)	0.03 (0.00,0.06)
Reading	Yes	0.24 (0.22,0.27)	0.16 (-0.02,0.32)	0.0001 (-0.02,0.01)	0.16 (-0.02,0.30)	0.08 (-0.06,0.26)
	No	0.10 (0.07,0.12)	0.12 (0.02,0.23)	-0.036 (-0.19,0.11)	0.08 (-0.00,0.16)	0.02 (-0.06,0.10)

5. Discussion

We explore three methodologies to account for the potential correlation between treatment and duration that is unassociated with actual ability changes, but is a result of changes in testing conditions. The filtering approach builds on the assumption that anomalous changes in duration being driven by test condition changes and not changes in ability. This assumption is supported by the instrumental variables approach, which yields larger direct effects of test conditions than indirect effects through understanding and grit. The alternative span approach

relies on the assumption that using the same time frame of testing (fall or spring) for both pretest and posttest has less of a likelihood of a test condition effect. The instrumental variable approach depends on finding excluded variables in the system of equations. Table 10 summarizes the different treatment effect estimates across our methodologies. The estimates are smaller once we control for potential changes in test condition, although the extent of decrease varies across methodologies.

Table 10: Comparison of Estimated Treatment Effects

	Math	Reading
Base	0.131***	0.176***
	(0.009)	(0.009)
Alternative Spans	0.120***	0.126***
	(0.011)	(0.011)
Filter Anomalous Durations	0.080***	0.138***
	(0.010)	(0.011)
2SLS	0.116***	0.169***
	(0.015)	(0.053)

We suspect that many education interventions are unlikely to suffer from the specific problems addressed in this paper. These problems emerge if test conditions change from pretest to posttest and that change is systematically correlated with treatment status. The first condition is less likely if the pretest and posttest are at the same time of year or where the tests are administered under strongly controlled conditions. Both of these are common in many interventions that use standardized tests administered each spring. The second condition may also be unlikely, if there is no connection between treatment and test conditions/duration. However, education researchers should not dismiss out-of-hand changes in test conditions, but should determine if there is any reason to suspect changes in the test conditions, and if so, how strategies such as those suggested in this paper can help address the sensitivity of the results to any such issue.

When duration data is unavailable, the direct effect of treatment on outcomes can be captured if the researcher can find an instrumental variable correlated with treatment but not with changes in duration. We are unable in our data to find such an instrument and contrast the findings.

6. Conclusion

Education evaluations that analyze student achievement typically rely on contrasts of pretests and posttests. For valid estimation, test conditions must be consistent and uncorrelated with the likelihood of receiving treatment. While this may typically be the case, it certainly will not always be true, and careful researchers should take time to evaluate whether there is reason for suspicion that this assumption is violated. In settings where changes in test conditions from pretest to posttest are correlated with treatment status, ignoring this potential issue can have a large effect on the treatment effect estimates. We find a decrease in the treatment effects when accounting for the change in test conditions, averaging about a 19% decrease for math and 18% for reading. We are unable to separate out the grit and test conditions effect, although we have described the methodology by which it can be done if the researcher collect a measure of grit. While the need for the methodology presented in this paper is limited only to certain settings, evaluations of educational interventions that use pretests and posttests should consider whether or not there is evidence for consistent testing conditions. If not, the methodology in this paper will help uncover the estimates of interest.

Future research may find a grit metric that allows for a fully decomposed treatment effect in order to understand how important the role grit plays in raising student achievement. In the context of other changes in test condition, the methodology suggested here may in fact simplify. For examples, some test conditions may not have any indirect effects (e.g., the use of a calculator

in some schools and not in others), so that the full effect would be decomposed only into the direct effect on treatment and the test condition effect. Understanding the context under which tests are administered is necessary for reliable estimates of educational intervention effect sizes, and our methodology allows for additional information of potential value besides the treatment effect, such as the relationship between learning and duration, and a separation of ability growth into grit and comprehension.

References

Haladyna, Thomas M., Susan Bobbit Nolen, and Nancy S. Haas. "Raising standardized achievement test scores and the origins of test score pollution."*Educational Researcher* 20, no. 5 (1991): 2-7.

Pane, John F., Elizabeth D. Steiner, Matthew D. Baird and Laura S. Hamilton. Continued Progress: Promising Evidence on Personalized Learning. Santa Monica, CA: RAND Corporation, 2015. http://www.rand.org/pubs/research_reports/RR1365.html.

Pearson, P. David, Sheila W. Valencia, and Karen Wixson. "Complicating the world of reading assessment: Toward better assessments for better teaching." *Theory Into Practice* 53, no. 3 (2014): 236-246.

Welsh, Megan E., Melissa Eastwood, and Jerome V. D'Agostino. "Conceptualizing teaching to the test under standards-based reform." *Applied Measurement in Education* 27, no. 2 (2014): 98-114.

Kesler, Ted. "Unstandardized Measures." *The Elementary School Journal,* 113, no. 4 (2013): 488-516.