Abstract Title Page
Not included in page count.

Title: Approaches to Incorporating Late Pretests in Experiments: Evaluation of Two Early Mathematics and Self-Regulation Interventions

Authors:
Fatih Unlu, Abt Associates
Carolyn Layzer, Abt Associates
Douglas Clements, University of Denver
Julie Sarama, University of Denver
David Cook, Abt Associates

Abstract Body

Limit 4 pages single-spaced.

Background / Context: Many educational Randomized Controlled Trials (RCTs) collect baseline versions of outcome measures (pretests) to be used in the estimation of impacts at posttest. Although pretest measures are not necessary for unbiased impact estimates in well-executed experimental studies, using them increases the precision of impact estimates and reduces sample size requirements to detect desired effect sizes (e.g., Raudenbush and Lui, 2000; Bloom, Richburg-Hayes, & Black, 2005; Hedges and Hedberg, 2007; and Schochet, 2008b). Ideally, baseline data collection should occur before the start of program/intervention implementation but unforeseen factors may delay it, leading to "contaminated" baseline measures (henceforth: "late" pretests).

Incorporating late pretests in impact regressions has two consequences (Schochet, 2008a). First, resulting impact estimates may be biased, reflecting early intervention effects. Magnitude and direction of this bias depends on the size and direction of the early effects (e.g., large and positive early effects cause impact estimates with large downward bias). Second, late pretests tend to be correlated with the treatment indicator(s) included in impact regressions due to the early treatment effect, yielding larger standard errors than uncontaminated pretests. Despite these adverse effects, using late pretests in impact analyses may still be preferable because they may explain a significant portion of the outcome variance and help with precision, offsetting the bias they introduce. This bias-precision tradeoff depends on the size of early treatment effects, growth trajectory of treatment effects, and how well late pretests explain posttest measures.

Purpose / Objective / Research Question / Focus of Study: This paper provides a theoretical overview of the late pretest issue and empirically assesses the bias-precision tradeoff in the experimental evaluation of the efficacy of adding a curriculum designed to support the development of children's self-regulation skills (Scaffolded Self-Regulation, SSR) to an established early mathematics curriculum (Building Blocks, BB) to form a synthesized curriculum, BBSSR.

Setting: The RCT described in this paper is conducted in three large districts in Southern California.

Population / Participants / Subjects: Our analytic sample includes 807 students who participated in both baseline and posttest data collection. These students were in 84 4-year-old classrooms across the three districts. A large proportion of their classrooms were multi-racial/multi-ethnic. In the three districts Hispanic children were the majority minority at on average 39%, Asian Pacific Islander 18%, African-American 11%, and non-Hispanic White 31%. On average, 27% of the students were English Language Learners (with roughly 20% having Spanish as the primary language).

Intervention / Program / Practice: Both of the evaluated interventions are theoretically and empirically grounded. The NSF-supported Building Blocks (BB) project produced a research-based math curriculum that addresses geometric and spatial ideas and skills and quantitative ideas and skills. The approach of BB is finding the math in, and developing math from, children's activity. Funded by the NSF and IES, three RCT evaluations have documented BB's positive

effects on young children's math achievement (e.g., Authors, 2007, 2008). Increasing math proficiency is significant—using each of six longitudinal data sets, the strongest predictors of later achievement are early math skills, followed by reading skills and then attention (Duncan et al., 2007). The self-regulation approach combines current research on the development of self-regulation and executive function with Lev Vygotsky's cultural-historical theory of child development to design optimal ways to support the development of self-regulation in young children (Bodrova & Leong, 2007). Studies indicate that scaffolding that promotes self-regulation improves mathematics learning (Barnett et al., 2006). The BBSSR intervention is a theoretically-grounded synthesis of this scaffolding and BB.

Significance / Novelty of study: Many recent educational evaluations collect and control for baseline outcome measures thanks to the vast literature recommending them for boosting statistical power. The dynamic nature of school districts and schools coupled with unforeseen logistical factors often delay collection of such data yielding contaminated pretest measures, which in turn produce biased impact estimates and diminish gains in statistical power. Building on Schochet (2008a), we examine this important issue within the context of a cluster RCT analyzing the effects of two preschool interventions. We also propose a novel approach that aims to decontaminate late pretest measures and can be applied in many similar situations.

Statistical, Measurement, or Econometric Model: Schochet (2008a) theoretically and empirically examines the implications of using late pretests in RCTs. Specifically; he compares the performance of the following four impact estimators with respect to their mean squared error and minimum detectable effect sizes:

- Posttest-only estimator: does not use any pretest measures,
- Difference-in-differences estimator (DID): uses gains scores as the outcome variable,
- ANCOVA estimator: controls for pretest measure(s), and
- Unbiased ANCOVA (UANCOVA) estimator: controls for uncontaminated pretest measure(s) obtained from extant sources (e.g., school-level test scores for prior student cohorts)

Schochet conducts simulations for typical RCTs and data generating parameters such as intra-class correlations, pretest and posttest correlations, and impact growth trajectories. He finds that:

- Both the DID and ANCOVA estimator is preferred to the posttest-only estimator unless early treatment effects are large (e.g., larger than 0.10 standard deviations for the ANCOVA estimator in a cluster RCT);
- The ANCOVA estimator is preferred to the UANCOVA estimator unless predictive power of alternative pretests and early treatment effects are large; and
- The DID estimator typically has larger bias and variance than the ANCOVA estimator.

Following Schochet, we assess the extent of early treatment effects and compare estimated impacts from the posttest-only and ANCOVA estimators for the BB and BBSSR conditions. In addition, we implement an alternative ANCOVA estimator that removes the contaminated portion of pretests, yielding unbiased impact estimates while still explaining some portion of the outcome. This approach entails: (i) building a model that uses the contaminated pretest as the outcome and all available exogenous covariates and the time between school start and baseline

testing as covariates, (ii) estimating this model using only the control students' pretest measures, and (iii) creating predicted pretests for all students using the estimated model but setting the time between school start and pretest date to zero. Note that this process yields predicted pretest scores that are free of early treatment effects because the prediction model is estimated using covariates that were collected before the treatment and only control students' scores.

Usefulness / Applicability of Method: Numerous educational studies rely on pretest measures to boost statistical power but unanticipated factors lead to contamination of these measures in many cases. The approach described above aims to address this issue and can be applied in all such cases with exogenous predictors of the pretest measures.

Research Design: We analyze a three-armed cluster RCT in which classrooms in early childhood centers or schools were randomized to the two intervention conditions and a business-as-usual control condition. Randomization was conducted separately for schools with one or two participating classrooms (groups A and B). Schools/centers in group A were placed into five randomization blocks such that each block consisted of all half-day or full-day PreK classrooms in one of the three study districts. Within each block, schools/centers were sorted with respect to prior math achievement, % reduced price lunch eligible, and %ELL and randomly assigned to the three conditions three at a time, starting with a randomly chosen point in the sorted list. This process was used to ensure balanced experimental groups. In each group B center/school, the two study classrooms were randomly assigned to two conditions that were randomly determined.

To examine the late pretest issue, we present separate impact estimates for BB and BBSSR conditions contrasting their outcomes with the control condition from three different models. The first model excludes pretest controls (posttest-only estimator). The second model includes four pretest measures described subsequently (ANCOVA with late pretests). The third model uses decontaminated versions of pretests aggregated to the school-level because most of the available predictors used in the correction process were at the school-level: age and gender of students and school-level covariates including size, average class size, percent minority and low income, and prior achievement (ANCOVA with corrected pretests). All three models (i) are two-level hierarchical linear models (HLM; Raudenbush & Bryk, 2002) that nest students within classrooms; (ii) include two indicator variables for the BB and BBSR conditions as primary predictors yielding impacts of BB and BBSSR; and (iii) include students' age at posttest and gender and indicators for randomization blocks (fixed effects) as covariates.

Data Collection and Analysis: Exhibit 1 describes pretest and posttest measures used in this paper. Measures of self-regulation included Pencil Tapping task (Diamond & Taylor, 1996), Head-Toes-Knees-Shoulders (Ponitz & McClelland, 2008), Self-Ordered Pointing (Blair & Willoughby, 2006), and Forward and Backward Digit Span (Gathercole & Pickering, 2000; Wechsler, 1986). Expressive Vocabulary Test, Second Edition (EVT-2, Williams, 2007) is the language measure analyzed. Measures of early mathematics ability were the Tools for Elementary Assessment in Mathematics (TEAM; Clements, Sarama, & Wolfe, 2011), and the mathematics battery from the Early Childhood Longitudinal Study-Birth Cohort (ECLS-B; NCES). In addition, we are currently collecting scores from the Desired Results Developmental Profile (DRDP), an early childhood assessment administered by teachers in the study districts, to be used in further iterations of the late pretest-correction process described above.

In most schools, baseline data was collected in fall 2010, between 30 and 90 days after the start of the school day as seen in Figure 1.[1] Factors contributed to this delay included delays in school district approvals and parent permission, and delays in security clearance for data collectors. Posttest measures were collected in spring 2011 without any major issues.

Findings / Results: Exhibit 2 presents various impact estimates. Columns 1-4 show impacts of BB while Columns 5-8 present BBSSR impacts. Specifically, Column 1 presents the estimated early impact of BB on three pretest measures while Columns 2-4 show BB impact estimates at posttest from the three estimators described above (no-pretest, ANCOVA with actual pretests, and ANCOVA with corrected pretests estimators), respectively. Column 1 suggests that the early impacts of BB were sizeable: 0.24 standard deviations (sds) and statistically significant on Head-Toes-Knees-Shoulders score and 0.17 sds on TEAM score. Consequently, controlling for these pretest measures in impact regressions for posttest measures dramatically reduces impact estimates for almost all outcomes, as seen in Columns 2 and 3. For example, for ECLS-B, the impact is 0.29 sds and statistically significant from the posttest-only model while it decreases to 0.13 sds and is not significant when late pretests are included as covariates. Also note that adding contaminated pretests hardly improves the precision of estimated impacts (standard errors of impact estimates decrease by 10-20%), which may be due to sizeable early treatment effects. These results suggest that when estimating impacts for of BB, controlling for contaminated pretests is not worthwhile due to large early treatment effects and the low explanatory power of late pretests. Comparing Columns 2 and 4 shows that using decontaminated pretests changes neither the magnitude of the impact estimates nor their standard errors. While the former result is expected, the latter suggests that predictors used in the decontamination process (student demographics and school characteristics) did not predict pretest measures well.

Columns 5-8 of Exhibit 2 display results for BBSSR. Column 5 shows that early treatment effects are much smaller. Hence, controlling for late pretests does not change the estimated impacts much while the boost in precision is somewhat larger than for BB – around 20-30%. Examining Column 8 suggests that corrected pretest measures do not change the impact estimates or standard errors. These results suggest that controlling for late pretests in this case may not be as problematic as it was in the previous case.

Conclusions: Analyses conducted thus far show that controlling for contaminated pretest measures in impact regressions can lead to substantially biased impact estimates while their effect on the precision of impact estimates is less profound, causing one to question the overall merit of using them. The preliminary application of our decontamination approach with a limited set of predictors yields unbiased estimates, but it does not seem to help much with precision. These analyses will be extended and improved in time for the conference with additional data collection currently underway. It will be interesting to observe how corrected pretests perform when DRDP scores —presumably better predictors of study-administered pretests— are used in the correction process.[2]

[1] There were no statistically significant differences in the timing of baseline testing across the three conditions.
[2] We will also compare results from our approach with those from the unbiased ANCOVA estimator that will directly control for these scores.

Appendices

Not included in page count.

Appendix A. References

Barnett, W. S., Yarosz, D. J., Thomas, J., & Hornbeck, A. (2006). Educational effectiveness of a Vygotskian approach to preschool education: A randomized trial: National Institute of Early Education Research.

Blair, C.B. & Willoughby, M.T. (2006). Measuring executive function in young children: Self-Ordered pointing. Chapel Hill, NC: The Pennsylvania State University and The University of North Carolina at Chapel Hill.

Bloom, H.S., Richburg-Hayes, L., and Black, A. R. 2005. Using Covariates to Improve Precision: Empirical Guidance for Studies that Randomize Schools to Measure the Impacts of Educational Interventions. *MDRC Working Paper*.

Bodrova, E., & Leong, D. J. (2007b). Play and early literacy: A Vygotskian approach. In K. A. Roskos & J. F. Christie (Eds.), Play and literacy in early childhood (2nd ed) (pp. 185200). Mahwah, NJ: Lawrence Erlbaum Associates.

California Department of Education. (2007). Desired Results Developmental Profile-Revised (DRDP-R).

Clements, D. H., Sarama, J., & Wolfe, C. B. (2011). *TEAM—Tools for early assessment in mathematics*. Columbus, OH: McGraw-Hill Education.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do". *Developmental Psychobiology, 29*, 315–334.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., …, Japel, C. (2007). School readiness and later achievement. Developmental Psychology, 43(6), 1428–1446.

Gathercole, S.E. & Pickering, S.J. (2000). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *British Journal of Educational Psychology, 70(2)*, 177-194.

Hedges, L. V. & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60-87.

Ponitz, C.E.C, McClelland, M. M., Jewkes, A.M., Connor, C. M., Farris, C.L., Morrison, F.J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly 23*, 141–158.

Raudenbush, S. W., & Lui, X. F. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods (Second ed.)*. Newbury Park: Sage.

Schochet, P. Z. (2008a). The Late Pretest Problem in Randomized Control Trials of Education Interventions (NCEE 2009-4033). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P. Z. (2008b). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics, 33(1)*, 62-87.

U.S. Department of Education, National Center for Education Statistics. Early Childhood Longitudinal Study-Birth (ECLS-B) Cohort. Washington, D.C.: Institute of Education Sciences.
Wechsler, D. (1986). Wechsler Adult Intelligence Scale, Revised (WAIS-R). New York, NY: The Psychological Corporation.

Williams, (2007). Expressive Vocabulary Test, Second Edition (EVT-2). Minneapolis, MN: Pearson.

Appendix B. Tables and Figures

Not included in page count.

Figure 1: Days Between School Start and Baseline Testing

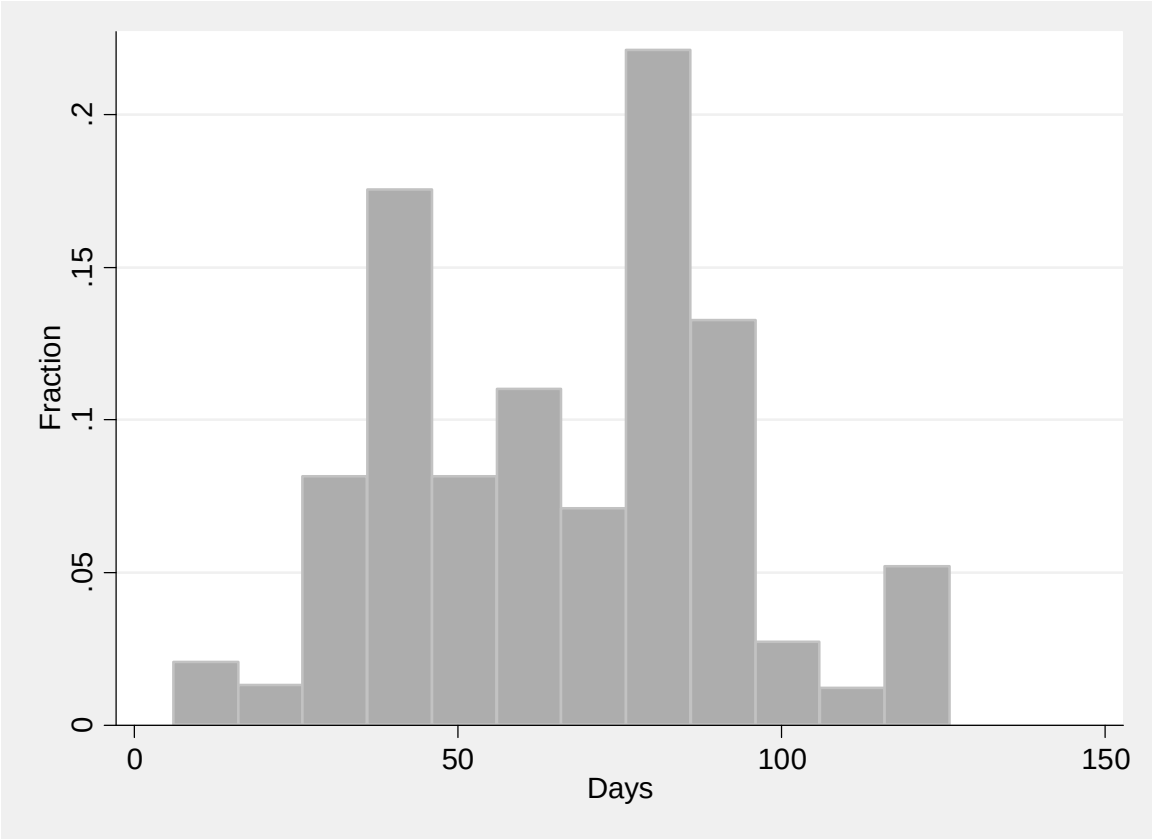

Exhibit 1: Outcome Measures			
Measure	Pre-test	Post-test	Description
Pencil Tapping ("Pencil")	X	X	In this measure of inhibitory control, children are asked to tap a pencil once if the assessor taps twice and tap twice if the assessor taps once.
Head, Toes, Knees, and Shoulders Score	X	X	In this measure of behavioral regulation (specifically, inhibitory control and working memory), the child must do an action that is systematically different from the oral instruction given by the assessor (and not follow the instruction given by the assessor).
TEAM Scaled Score	X	X	This is a measure of developmental progressions in number (e.g., verbal counting, object counting, subitizing, number comparison and sequencing, number composition and decomposition, adding and subtracting, and place value) and geometry (e.g., shape recognition, shape composition and decomposition, congruence, construction of shaped, spatial imagery, geometric measurement and patterning).
PPVT-III	X		The Peabody Picture Vocabulary Test, 3rd Edition, is a test of receptive vocabulary.
ECLS-B Math Score		X	This is the math battery from the ECLS-B test and is comprised of a collection of items from other widely-used tests of early mathematics.
Forward Digit Span Score		X	This is a measure of working memory, testing a child's ability to reproduce an increasingly long string of numbers that the assessor says. The assessor presents (orally) the subject with a series of digits (e.g., '8, 3, 4'), and the subject must immediately repeat them back. If s/he does this successfully, s/he is given a longer list (e.g., '9, 2, 4, 6'). The length of the longest list a person can remember is that person's digit span.
Backward Digit Span Score		X	This is a measure of working memory, testing a child's ability to reproduce an increasingly long string of numbers that the assessor says, in reverse order. While the participant is asked to repeat the digits in the given order in the forward digit-span task, s/he is asked to repeat them in reverse order in the backward digit-span task (e.g., if presented with '8, 3, 4', the

			subject must produce ‘4, 3, 8’). As in Forward Digit Span, the length of the longest list the subject is able to repeat is the subject’s digit span.
EVT-2 Score		X	This is a measure of expressive oral language ability (English). The subject is shown a picture and asked standardized prompts to elicit an English word for the object’s label, depicted action, or use.
Self-Ordered Pointing Score		X	This is a measure of working memory in which the child is shown multiple pages that contain the same pictures in different arrangements. The child is asked to identify a picture which s/he has not selected on previous pages so must recall which item(s) s/he has already selected.

Exhibit 2: Estimated impacts of BB and BBSSR (presented in effect sizes)								
Outcome Measures	BB vs. CTRL					BBSSR vs. CTRL		
	Impact on Pretest	Impact on Posttest				Impact on Pretest	Impact on Posttest		
		Posttest only	ANCOVA w/ actual pretests	ANCOVA w/ corrected pretests			Posttest only	ANCOVA w/ actual pretests	ANCOVA w/ corrected pretests
	(1)	(2)	(3)	(4)		(5)	(6)	(7)	(8)
Forward Digit Span Score	-	0.15 (0.11)	0.05 (0.09)	0.13 (0.11)		-	0.01 (0.11)	0.01 (0.09)	0.00 (0.10)
Backward Digit Span Score	-	0.20* (0.10)	0.13 (0.09)	0.24* (0.10)		-	0.00 (0.09)	0.01 (0.09)	0.00 (0.09)
ECLS-B Math Score	-	0.29* (0.11)	0.13 (0.09)	0.29* (0.11)		-	0.05 (0.11)	0.09 (0.08)	0.05 (0.11)
Pencil Tapping ("Pencil") Score	0.14 (0.11)	0.08 (0.09)	0.01 (0.09)	0.09 (0.10)		-0.02 (0.11)	0.01 (0.09)	0.07 (0.08)	0.00 (0.09)
Head, Toes, Knees, and Shoulders	0.24* (0.11)	0.18 (0.11)	0.04 (0.09)	0.19 (0.11)		-0.14 (0.10)	0.02 (0.11)	0.08 (0.09)	0.01 (0.10)
TEAM Scaled Score	0.17 (0.12)	0.10 (0.12)	-0.02 (0.09)	0.10 (0.12)		-0.04 (0.11)	0.08 (0.12)	0.11 (0.08)	0.05 (0.12)
EVT-2 Score	-	0.07 (0.16)	-0.06 (0.10)	0.08 (0.16)		-	0.03 (0.16)	0.03 (0.10)	0.00 (0.15)
Self-Ordered Pointing Score	-	-0.11 (0.13)	-0.15 (0.11)	-0.13 (0.13)		-	-0.07 (0.12)	-0.02 (0.11)	-0.06 (0.12)

Notes: Standard errors of impact estimates are presented in parentheses. * denotes statistically significant impacts at the $p<0.05$ level. Standard deviation of the outcome measures in the control group are used to translate impact estimates and standard errors into effect sizes.