The Effects of the Payne School Model on Student Achievement

Submitted by Dr. Joseph A. Taylor

Executive Summary

This study was commissioned by *Teaching Tolerance* to synthesize evidence of effectiveness for the Payne School Model (PSM). The following synthesis report summarizes findings from 33 study reports that seek to demonstrate the effectiveness of the PSM. These reports can be found on the following web sites:

http://www.ahaprocess.com/solutions/k-12-schools/results-best-practices/
http://www.ahaprocess.com/solutions/k-12-schools/results-best-practices/scientific-research/.

Among the 33 study reports, there are 10 comprehensive studies that were conducted by Dr. William Swan, or in some cases, Swan and colleagues. The author(s) of the remaining *Data Speaks* reports are unknown to me. The larger body of studies range from quasi-experimental designs to studies of longitudinal student achievement trends. The approach to this report was to assess each study for its ability to attribute student outcomes to PSM exposure, as well as to make global assessments of what the collective evidence suggests about the effectiveness of the PSM. I had no contact with any of the researchers who conducted the studies described in the 33 reports so my assessments are based solely on the information provided in the reports.

The conclusion from my synthesis of these reports is that the current evidence base for PSM does not support confident conclusions that it is effective. This assessment is based on evidence criteria that draw strongly from those of the What Works Clearinghouse, the preeminent federal entity for vetting the validity of studies of education interventions. Each of the studies reported had one or more design limitations that severely limited its ability to isolate the effects of PSM from other factors that might affect outcomes or from pre-existing differences between groups. Specifically, each study either lacked a valid comparison group, was seriously confounded (i.e., PSM effect cannot be disentangled from that of specific teachers, schools, or districts), did not demonstrate that groups being compared were equivalent prior to the PSM intervention, or suffered from a combination of these limitations. Finally, attempts to synthesize effects were severely inhibited by the statistical reporting practices in the reports. That is, effect

sizes could rarely be extracted from the studies and it is these effect size metrics, not p-values, that can be defensibly aggregated across studies.

This said, the researchers did make prudent choices in electing to use state test scores that tend to be reliable and valid measures and employed sophisticated analysis techniques such as Analysis of Covariance (ANCOVA) to estimate treatment effects. However, these steps are not sufficient to overcome the limitations above. In a similar vein, findings from selected studies suggest that the PSM could be a promising intervention under certain conditions. However, the collective evidence stops short of demonstrating effectiveness with the confidence needed for policy decisions.

My suggestion is that the next step in the research agenda be a more controlled study of the PSM with a larger sample of students and schools and that the effects from these and subsequent rigorous studies be reported in such a way (i.e., report effect sizes) that future synthesis efforts have more information to draw upon when forming conclusions.

Full Synthesis Report

Synthesis Study Objectives

The primary objectives of this study were to: a) assess the evidence base for the effectiveness of the Payne School Model (PSM) on student achievement in multiple domains, and b) synthesize the effects of the PSM when those effects can be extracted from rigorous studies.

Method

Overview of Procedure

The synthesis study was conducted in three steps:

1. Code each study/substudy on key characteristics related to the strength of the evidence it provides.
2. Make an assessment the overall strength of the evidence a study/substudy provides as well as its compliance with What Works Clearinghouse (WWC) criteria:(http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf)
3. For studies/substudies that Meet WWC Evidence Standards with or without Reservations, and for which a standardized mean-difference effect size could be extracted, synthesize the effects of the PSM using meta-analytic techniques.

Sample of Studies Reviewed

A total of 33 studies (and their eligible substudies) were reviewed for this synthesis. The titles of these studies are listed in the leftmost column of Table 1. This set of studies contains full research reports as well as shorter research briefs.

Limitations in Study Scope

There are several limitations to the study scope to note:

1. The synthesis study researcher did not communicate with any of the researchers who conducted the effectiveness studies of PSM. Therefore no additional information, besides what is in the 33 reports, was available to the synthesis

researcher. Thus, all assessments and conclusions are based solely on what is written in each of the reports.

2. Similarly, the synthesis researcher assumes that all of the effectiveness evidence for the PSM resides in these reports. Thus, any conclusions about PSM effectiveness or the evidence base around it are based solely on these 33 reports.
3. This synthesis only examined the effects of the PSM on student outcomes.
4. The synthesis researcher focused on assessing findings related to the "main effect of treatment." That is the synthesis focuses on effects that involved outcomes from all students in the treatment group and all students in the comparison group (not subgroup effects, e.g., effects for economically disadvantaged students only).

Study Coding Criteria

Outcome Domain(s). Each study/substudy was coded for the outcome domain of the test used to document impact (often a state standardized test in mathematics or literacy)

Design. Each study/substudy was coded as either a Quasi-Experimental Design (QED) where treatment and comparison groups were formed using a non-random process, or a TREND study that examines changes in achievement levels over time [Note: one study used time series (TS) design]. TREND studies cannot meet WWC standards (See Table 1 Key below for rationale).

Confounds. Each study/substudy was coded for the existence of a "n=1" or bundled intervention confound. Bundled intervention confounds occur when the program of interest is known to be occurring at the same time as another intervention that might also affect outcomes. N=1 confounds occur when an intervention and/or comparison group includes just one teacher, school, or district. In such a situation, the effect of the intervention cannot be disentangled from that of the specific teacher, school, or district. Presence of either type of confound will not allow a study to meet WWC standards.

Sample Size. For each study/substudy that provided a sample size, this information was recorded.

Baseline Equivalence (BLE). Each study/substudy was assessed for whether the study can demonstrate that groups whose achievement is being compared post-intervention were equivalent on achievement measures prior to the intervention. Specifically, studies much show that the exact set of students being compared post-intervention (i.e., the analytic sample) were equivalent at baseline. The WWC cutoff for equivalence is a baseline equivalence effect size smaller than 0.25 standard deviations. This cutoff exists even when researchers use Analysis of Covariance (ANCOVA), as often done in these study reports, to correct for baseline differences. ANCOVA is a statistical technique that adjusts an estimated treatment effect on an outcome measure (e.g., post-PSM achievement) for pre-existing differences across groups on that very same outcome (e.g., pre-existing differences in pre-PSM achievement). Demonstrating baseline equivalence over and above conducting ANCOVA is necessary because the ANCOVA adjustment does not function optimally when baseline (pre-existing) differences are too large.

Direction of Findings. The direction of the effect, positive (favors PSM), negative (favors comparison), or null (essentially zero effect) was gathered from each study/substudy.

Type I Error Probability. The study p-value of the treatment effect was recorded for each study/substudy.

Effect Size. A standardized mean-difference effect size was extracted for studies where sufficient statistical information was reported. **Background on effect sizes**: Consider a simple study with 50 students (25 treatment, 25 control). If the treatment students outperform the control students by 5 points, on average, on the post test, and both groups have a standard deviation of 10 points in their post test scores, the standardized mean-effect size would be 5/10 or 0.50. Had all else been the same and the groups had 50 students each, the effect size would be exactly the same, 0.50. Thus the effect size is a sample-size independent measure of the magnitude of an intervention effect and allows for "apples and apples" comparisons of effects across studies of different sample sizes. On the other hand, p-values cannot be defensibly synthesized across studies. Study p-values are not only a function of the size of the effect, but also the study sample size and this mutual influence cannot be easily disentangled. Thus counting significant and

insignificant treatment effects is not a valid approach to assessing overall effectiveness as large effects could be insignificant if the sample size is small and small effects could be significant if the sample size is large. This is the primary limitation of Dr. Swan's two-page synthesis of PSM effects titled: *The Payne School Model's Impact on Student Achievement— A National Study (*https://www.ahaprocess.com/wp-content/uploads/2014/01/Payne-School-Model-Impact-National-Study.pdf*)*.Consider again the example above. In the case where there were 25 students in each group, the p-value associated with the 5-point mean difference is 0.08 (a non-significant result) and the p-value associated with the same 5-point mean difference with 50 students in each group is 0.01 (a significant result). As a final illustration, consider a mean difference of 2 IQ points between two groups of students. Clearly, this would be an imperceptible difference in intelligence quotient and has a correspondingly small effect size (0.13). However, this small difference in IQ could be statistically significant with as few as 425 students per group.

Study Strengths. Each study/substudy was coded on the strength of its design, baseline, equivalence, and reliability of measures.

Study Limitations. Each study/substudy was coded for design flaws (e.g., confounds), evidence of baseline equivalence, validity of comparisons, and the reliability of measures.

Strength of Evidence. Each study/substudy was coded as providing either strong, moderate or weak evidence of effectiveness based on the following criteria:

> **Strong**: Design is a Randomized Control Trial (RCT) with low attrition, no evidence of a confound, reliable measures
>
> **Moderate:** Design is a QED with established baseline equivalence, reliable measures and no evidence of confounds
>
> **Weak:** Design is a QED with confounds OR no evidence of baseline equivalence OR unreliable measures. Design is longitudinal (e.g., pre-post, trend) without a valid comparison group.

WWC Rating. Each study/substudy was assessed using WWC standards (3.0) for design, baseline equivalence, and reliability of measures. These are the gold standards of *intervention research*, but not all research, in education. The WWC standards were put in place by the US Department of Education to help policy makers and school district personnel make more informed decisions about education programs by helping them wade through the sea of intervention research available in education, some rigorous, some not. The WWC assigns three different ratings to studies of education interventions. These ratings pertain to the level of evidence provided by a study that observed changes in student outcomes can be confidently attributed to the intervention that was provided. The ratings are:

> **Meets WWC Evidence Standards**. This rating can only be given to a randomized control study (RCT) with a sound design (e.g., no confounds), reliable measures, and low attrition.
>
> **Meets WWC Evidence Standards with Reservations.** This rating is given to quasi-experiments (treatment groups formed by non-random processes) and high attrition RCTs that can demonstrate baseline equivalence on outcome measures, use reliable measures, and are free from research design issues (e.g., n=1 confounds). This is the highest rating that a quasi-experiment can receive. This is because the non-random process of group assignment in a QED can cause groups to be dissimilar in ways that affect outcomes and that cannot be adjusted for statistically. Shadish, Cook, and Campbell (2002) refer to this problem as *selection bias*.
>
> **Does Not Meet WWC Evidence Standards**. This rating is assigned to any intervention study that has a confound or uses unreliable measures. This rating also applies to quasi-experimental studies that cannot demonstrate baseline equivalence and randomized control studies that have high attrition and cannot demonstrate baseline equivalence.

Results

Global Observations

Coded variables for all 33 studies are in Appendix A (Table 1). Global observations across the studies are summarized below:

1. All studies provide weak evidence of effectiveness and none meet WWC standards. This is the case because all studies had one or more of the following characteristics that severely limited its internal validity:
    a. N=1 confound: many studies compared the PSM model in just one school to business as usual instruction in just one other school. This makes it impossible to disentangle the effects of the PSM model from the individual school effects. Other studies had similar confounds at the teacher or district levels.
    b. No evidence of baseline equivalence (BLE): it is extremely important, especially in quasi-experiments that form groups in non-random ways, to demonstrate that the treatment groups were equivalent at baseline. No quasi-experiments provided such evidence and the statistical adjustments conducted may not have been sufficient if the BLE effect size was larger than 0.25SDs.
    c. Designs without valid comparison groups. The trend studies mainly focus on comparisons of different cohorts of students who might have been systematically different regardless of the intervention. Similarly, student mobility alone could explain the reported longitudinal changes in percent proficient (for example). **Note**: see the Key to Table 1 for further description of the internal validity threats associated with confounds, lack of baseline equivalence, and trend studies.

Synthesis of Effects and Comparison to Empirical Benchmarks

In the end, it was my opinion that a meta-analytic synthesis of effect sizes and subsequent comparison to empirical benchmarks was not a valid undertaking. The rationale for this is two-fold: a) none of the quasi-experimental substudies provided even a moderate level of evidence so the magnitude of these effects is suspect. Further, I was only able to extract effect sizes for 7 of the 38 quasi-experimental substudies in the set of reports. I would have no grounds for assuming

that these seven effects are representative of the larger sample of 38 effects that I would have calculated had sufficient statistical information been reported. Effect sizes could not be extracted for the remaining 31 quasi-experimental substudies, either directly or indirectly. For example, ANCOVA results can be converted to effect sizes but the necessary correlation between pretest and posttest was not reported. Further, effect sizes could have been computed using the reported adjusted means but the necessary standard deviations for the treatment and comparison groups were not reported.

As such, computing a suspect summary effect for the PSM would, in turn, create suspect comparisons to established empirical benchmarks. This said, should the reader be interested in comparing the raw effects from Table 1 to empirical benchmarks from rigorous meta-analyses, these can be found in Hill, Bloom, Rebeck-Black, & Lipsey (2008) – see References.

Conclusion

Many of the quasi-experimental studies of the intervention are confounded and lack evidence of baseline equivalence so the evidence from these designs is inconclusive. The trend data is consistently positive in the years that the schools engaged with the PSM intervention. However, most if not all of the studies make achievement comparisons where the effect of the intervention cannot be fully isolated from other influences. Given the consistent positive trends suggested by the longitudinal data, the notion that the PSM is a promising intervention seems quite plausible to me. This issue is that the current evidence base for the PSM, as provided in the reports and as reviewed here, simply cannot support confident causal claims about its effectiveness. This said, should new or additional information become available about the effectiveness of PSM, I would be amenable to revisiting my assessments.

I recommend that future research employ one or more rigorous random experiments (or quasi-experiments with matching) that test the intervention with multiple schools in the treatment (PSM) condition and multiple schools in the comparison condition (to avoid confounds). In addition, researchers should track attrition to make sure that the students in the sample used to compute treatment effects have the same equivalence of characteristics as was likely produced by the random assignment process. This would entail computing treatment effects with only those students who were randomly assigned to treatment conditions (do not include “joiners” in the analysis) and demonstrating that the sample of students used to estimate treatment effects was equivalent on achievement outcomes as baseline (pre-intervention). Further, reporting of full

descriptive statistics for the outcome measure (means, SDs, sample sizes, by treatment group) will facilitate effect size calculations that constitute more useful estimates of the practical significance of PSM effects. Finally, it would be optimal to have an independent, third-party evaluator perform the random assignment, collect the outcome data, and conduct the impact analysis.

References

Hill, C.J., Bloom, H.S., Rebeck-Black, A., & Lipsey, M.W. (2008). Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2(3), 172-177.

Shadish, W., Cook, T. & Campbell, D (2002) *Experimental & Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Table 1: Study Data

Report Title	Outcome Domain	Design	N=1 Confound?	Sample Size	BLE	Direction of Effect	Study *p*	Effect Parameter	Strengths	Limitations	Strength of Evidence	WWC Rating
Research Reports												
1. Report-Lowndes-03-09-2012	Math, ELA	Trend, QED	Yes	*NR*[3]	No	Positive	NR	*1*	*1*	*1,2*	Weak	Does not Meet WWC Evidence Standards
2. Report-Ridgeroad 04-01-2011	Math, Literacy	Trend, QED	Yes	*NR*[3]	No	Positive	NR	*1*	*1*	*1,2*	Weak	Does not Meet WWC Evidence Standards
3. Research-Reports-Arkansas -03-04	Math, Literacy	QED	Yes	231	No	Positive for Literacy Null for Math	0.586 (Math) .058 (Literacy)	ES=0.00 (Math) ES=0.14 (Literacy)	*1*	*2*	Weak	Does not Meet WWC Evidence Standards
4. Research-Reports-Arkansas -04-05	Math, Literacy	QED	Yes (8th grade) Unknown (7th grade)	232 (8th grade Math) 233 (8th grade literacy) *NR*[3]*(7th grade Math)*	No	8th grade Math - negative 8th grade literacy-positive *7th grade Math-positive*	8th grade Math - *NR* 8th grade literacy-0.04 *7th grade Math-0.029*	8th grade Math - 2 8th grade literacy- 2 *7th grade Math- 1*	*1,2*	*2*	Weak	Does not Meet WWC Evidence Standards
5. Research-Reports-Arkansas -05-06	Math, Literacy	QED	Presumably no	293 (6th grade Math) 273 (7th grade Math) 110 (7th grade literacy) *163 (8th grade literacy)*	No	6th grade Math - positive 7th grade Math - negative 7th grade literacy - positive *8th grade literacy - positive*	6th grade Math - <.001 7th grade Math 0.308 7th grade literacy 0.303 *8th grade literacy 0.046*	*2*	*1,2*	*2*	Weak	Does not Meet WWC Evidence Standards
6. Research-Reports-Indiana-01-03	ELA, Math	TS	No	NR	No	8 positive comparisons, 6 null, 4 negative	<0.01, <0.001	*4*	*1,2*	*1*	Weak	Does not Meet WWC Evidence Standards
7. Research-Reports-	Math,	Trend,	Likely for 5th grade math,	365 (3rd grade	no	3rd grade math -	3rd grade math -	3rd grade math:	*1, 2*	*1,2*	Weak	Does not Meet

Report Title	Outcome Domain	Design	N=1 Confound?	Sample Size	BLE	Direction of Effect	Study *p*	Effect Parameter	Strengths	Limitations	Strength of Evidence	WWC Rating
Kansas-05-06	Reading	QED	5th and 6th grade reading	math) 330 (4th grade math) 258 (5th grade math) 269 (6th grade math) 364 (3rd grade reading) 332 (4th grade reading) 298 (5th grade reading) 280 (6th grade reading)		negative 4th grade math - negative 5th grade math- negative 6th grade math - positive 3rd grade reading - negative 4th grade reading - negative 5th grade reading - positive 6th grade reading - positive	0.091 4th grade math – 0.266 5th grade math- 0.487 6th grade math - <.001 3rd grade reading – 0.341 4th grade reading – 0.140 5th grade reading – 0.024 6th grade reading – 0.008	ES=-0.29[3] 4th grade math: ES=-0.12[3] 5th grade math[2] 6th grade math[2] 3rd grade reading: ES=-0.05[3] 4th grade reading: ES= -0.16[3] 5th grade reading: ES= 0.47[3] 6th grade reading[2]				WWC Evidence Standards
8. Research-Reports-New York-04-05	Math, ELA	QED	Yes	111 (ELA) 113 (Math)	no	ELA – positive Math - positive	ELA: 0.116 Math: 0.271	ELA[2] Math[2]	*1,2*	*2*	Weak	Does not Meet WWC Evidence Standards
9. Research-Reports-Tennessee-04-05	Math, Reading	QED	Likely for Math and Reading 2nd-6th grade	129 (2rd grade math) 142 (4th grade math) 103 (5th grade math) 139 (6th grade math)	no	2nd grade math - positive 4th grade math - negative 5th grade math- positive 6th grade math – positive	2nd grade math: <0.001 4th grade math: 0.085 5th grade math: 0.167 6th grade math: 0.134	2nd grade math[2] 4th grade math[2] 5th grade math[2] 6th grade math[2] 7th grade math[2] 8th grade	*1,2*	*2*	Weak	Does not Meet WWC Evidence Standards

Report Title	Outcome Domain	Design	N=1 Confound?	Sample Size	BLE	Direction of Effect	Study *p*	Effect Parameter	Strengths	Limitations	Strength of Evidence	WWC Rating
				133 (7th grade math) 140 (8th grade math) 127 (2nd grade reading) 142 (4th grade reading) 103 (5th grade reading) 138 (6th grade reading) 132 (7th grade reading) 139 (8th grade reading)		7th grade math - positive 8th grade math - negative 2nd grade reading - positive 4th grade reading - negative 5th grade reading - positive 6th grade reading - positive 7th grade reading - positive 8th grade reading - negative	7th grade math: 0.778 8th grade math: 0.642 2nd grade reading: <0.001 4th grade reading: 0.081 5th grade reading: 0.512 6th grade reading: 0.187 7th grade reading: 0.836 8th grade reading: 0.046	math[2] 2nd grade reading[2] 4th grade reading[2] 5th grade reading[2] 6th grade reading[2] 7th grade reading[2] 8th grade reading[2]				
10. Research-Reports-Wisconsin-04-05	Math, Reading	QED	Yes	62 (5th grade math) 72 (6th grade math) 87 (8th grade math) 65 (10th grade math)	no	5th grade math- positive 6th grade math – positive 8th grade math - positive 10th grade math - positive	5th grade math: 0.857 6th grade math: 0.011 8th grade math: 0.008 10th grade math: 0.023	5th grade math[2] 6th grade math[2] 8th grade math[2] 10th grade math[2] 5th grade reading[2] 6th grade	*1,2*	*2*	Weak	Does not Meet WWC Evidence Standards

Report Title	Outcome Domain	Design	N=1 Confound?	Sample Size	BLE	Direction of Effect	Study *p*	Effect Parameter	Strengths	Limitations	Strength of Evidence	WWC Rating
				62 (5th grade reading) 72 (6th grade reading) 87 (8th grade reading) 65 (10th grade reading)		5th grade reading - positive 6th grade reading - positive 8th grade reading - positive 10th grade reading - positive	5th grade reading: 0.052 6th grade reading: 0.781 8th grade reading: <0.001 10th grade reading: 0.472	reading[2] 7th grade reading[2] 8th grade reading[2]				
Data Speaks												
11. Issue #1 – RRMCS 2008	Math, Literacy	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
12. Issue #2 – MISD 2009	Office Referrals	Trend	No	NR	NR	Positive	NR	*4*		*1*	Weak	Does not Meet WWC Evidence Standards
13. Issue #3 – NLR Math 2009	Math	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
14. Issue #4 – MISD CFK Year 1	Truancy	Trend	No	NR	NR	Positive	NR	*4*		*1*	Weak	Does not Meet WWC Evidence Standards
15. Issue #5 – MISD Read By 5	Reading (words, sounds)	QED	Presumably not	NR	NR	Positive	NR	*4*		*2*	Weak	Does not Meet WWC Evidence Standards
16. Issue #6 – Kepner 2009	Math, Writing, Science	Trend	No	NR	NR	Mixed	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
17. Issue #7 – South Brandywine	Math, Reading	Trend	Yes (bundled intervention)	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
18. Issue #8 – MISD CFK Year 2	Truancy, behavior	Trend	No	18	NR	Positive	NR	*4*		*1*	Weak	Does not Meet WWC Evidence Standards
19. Issue #9 – 7th Street Elementary	Math	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
20. Issue #10 – NLR Math 2010	Math	Trend	No	NR	NR	Mixed	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence

Report Title	Outcome Domain	Design	N=1 Confound?	Sample Size	BLE	Direction of Effect	Study *p*	Effect Parameter	Strengths	Limitations	Strength of Evidence	WWC Rating
												Standards
21. Issue #11 – North Brandywine	Math, Reading	Trend	No	530	NR	Mixed	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
22. Issue #12 – Allen ISD	ELA, Math, Science	Trend	No	No	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
23. Issue #13 – Winfield Elementary – Tucker	Signing	Trend	No	No	NR	Positive	NR	*4*		*1*	Weak	Does not Meet WWC Evidence Standards
24. Issue #14 – RRMCS 2010	Math, Literacy	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
25. Issue #15 – Cabot JH North	Math, Literacy	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
26. Issue #16 – MISD Science 2010	Science	Trend	No	~60	NR	Mixed	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
27. Issue #17 – West Lowndes 2010	Math	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
28. Issue #18 – Allen ISD	ELA, Math, Science	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
29. Issue #19 – Fort Worth CAN! Academy	Math, ELA, Science, Social Studies	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
30. Issue #20 – Goose Creek Memorial HS	Math	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
31. Issue #21 – Bedford Elementary	Math, Reading	Trend	No	NR	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
32. Issue #23 – Bedford Middle School	English, Math	Trend	No	472	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards
33. Issue #24 – Elephant's Fork Elementary	Reading, Writing, Math, Science, History	Trend	No	335	NR	Positive	NR	*4*	*1*	*1*	Weak	Does not Meet WWC Evidence Standards

Key for Study Data Table (Table 1)

Strengths

1 = This analysis uses state achievement test data. State tests normally have strong psychometric properties (i.e., reliability and validity) and test content based on the state and/or national standards that students and teachers are working toward.

2 = This study quantifies model fidelity with an observation protocol where protocol users were highly consistent in their ratings (IRR>0.50).

Limitations

1 = This study makes claims about program effectiveness based upon upward longitudinal trends in achievement (e.g., percent proficient) on the state test. Longitudinal analyses without a comparison group or that compare achievement gains to those expected by chance do not meet WWC evidence standards. The rationale is that program effects estimated within this design cannot be disentangled from those of other interventions or school-level initiatives happening concurrently with the intervention (i.e., The PSM), that could also affect state test scores in the relevant time frame (see *history effects* in Shadish, Cook, & Campbell, 2002). In addition, it is impossible to know from the report the influence of student mobility. For example, an exodus of lower achieving students or an influx of higher achieving students, over time, could alone produce the same effects shown in this study. This confounding influence of mobility could be especially problematic given the small sample size and the use of percent proficient as the outcome.

Further, very few of these claims are based on longitudinal comparisons of the same cohort of students (e.g., percent proficient in third grade in 2009 to percent proficient in 5^{th} grade in 2011). A few such comparisons can be deduced from the bar graphs and most of the effects from these comparisons appear to be modest. The bulk of the claims entail comparisons of different cohorts of students who could have different extant levels of proficiency before ever receiving the intervention. Thus, one has no way of knowing whether the reported increases in proficiency within a given grade level are due to the intervention or whether they are due to the fact that the various cohorts of students are systematically different with or without the intervention. Finally, it is impossible to know whether the trends reported in this study are unique to the timeframe of the intervention or are noteworthy when compared to the natural variation in state test scores. For example, evidence is not provided that the upward trends were not occurring before the intervention and that the observed upward trend is not a continuation of a prior (extant) upward trend. Also, the significance of the upward trend data would be easier to interpret in the context of the magnitude of natural variation in test scores for the schools.

2 = This study uses a quasi-experimental design to compare the student outcomes of students in the treatment school with the outcomes of students in a demographically similar school. Demographic equivalence is not the same as baseline equivalence (BLE) on achievement measures and this study does not demonstrate achievement equivalence prior to the onset of the intervention. Thus, one does not know whether the observed treatment effect is a genuine effect or just a reflection of extant achievement differences across the two schools. Further, the fact that there is just one district, school or teacher in each treatment group introduces a "n=1" confound to the quasi-experiment. That is, the effect of the intervention cannot be disentangled from the effects of being associated with that particular district, school or teacher. This confound and/or the inability to demonstrate baseline achievement equivalence prohibits this quasi-experiment from meeting WWC evidence standards.

NR = Not Reported

NR[3]= Sample size only reported for the treatment group

Effect Parameter

1 = Effect size cannot be computed as the comparison group sample size is unknown

2 = An effect size could not be computed using the F statistic from the ANCOVA without knowledge of the correlation between outcome measure and baseline measure OR the effect size could not be computed using the adjusted means as the treatment group-specific standard deviations were not reported. For some studies, an effect size could be calculated if raw (unadjusted) means, SDs, and sample sizes were reported, but this effect could be biased based on the size and direction of any baseline achievement differences.

3= Effect size computed from samples sizes and F statistic from ANOVA.

4 = No comparable effect size can be calculated from the reported design and data.