A Quasi-Experimental Study of the Effect of the *Leader in Me* on Attendance and Discipline in Missouri Schools

A technical report prepared by Dr. Mark White

Survey Research Center

Institute for Social Research

The University of Michigan

January 24, 2018

A Quasi-Experimental Study of the Effect of the *Leader in Me* on Attendance and Discipline in Missouri Schools

The Leader in Me (LIM) is a whole school intervention that has been adopted by over 3,500 schools and serves more than 1,400,000 students. Despite its wide adoption, very little research has formally explored the impact of LIM on student outcomes. This report provides such evidence, exploring the causal effect of the LIM program on attendance rates and disciplinary outcomes in Missouri schools using a quasi-experimental, interrupted time series analysis. Interrupted time series have been shown to provide strong causal evidence for program effects (Jacob, Somers, Zhu, & Bloom, 2016). Rather than providing an introduction and background information on the LIM program or on the methodological approach, this report focuses on the analyses and the results, providing just enough information to judge the reliability and validity of the analyses. Future reports will expand this paper to provide a more traditional introduction, literature review, and implications of the analyses.

Method

Data

The analyses in this report use school-level state data from Missouri. Missouri was chosen because of the large number of LIM schools in the state and the availability of school level behavioral outcomes. Specifically, the full population of Missouri elementary schools was used in the analysis, including 119 LIM schools and 1134 non-LIM schools. All behavioral data analyzed in this report is publicly available from the Missouri Comprehensive Data System - District and School Information. Data was also obtained from the Common Core of Data 2014 (https://nces.ed.gov/ccd/ccddata.asp) to provide school level covariates for matching schools. Data exists for all relevant outcomes and covariates from 2010 through 2016 and for discipline rates and student demographics through 2017.

Of interest for this report was whether LIM had any impact on two publically available student behavioral outcomes: attendance rates and disciplinary incidents. Attendance rates are the percentage of days in a given school year that a student attends school, averaged across all students. The average attendance for Missouri elementary schools was over 90%. Disciplinary incidents captures reportable disciplinary issues totaled across a number of categories: alcohol possession, drug possession, tobacco Use, violent incidents, incidents involving weapons, in-school suspensions, out-of-school suspensions, and expulsions. Note that minor incidents are not captured, but only serious problems so the incidence rate is quite low: an average of ~0.4 total incidents per 100 students. These two outcomes were deemed relevant to the intended goals of the LIM intervention. As students develop leadership skills emphasized in the LIM program, it is expected that they will make a greater effort to attend school and that their behavior while in school will improve.

The LIM schools adopted the program between 2011 and 2017. Two schools began LIM in 2011, 10 schools began in both 2012 and 2013, 30 began in 2014, 22 began in 2015, 19 began in 2016, and 25 started in 2017. This staggered starting of the program is beneficial from a causal standpoint because no single event co-occurred with adoption of LIM that could provide an alternative explanation of the treatment effect. Ten schools no longer implement LIM and 26 schools that have achieved lighthouse

	Unmatched Controls	Matched Controls	LIM Schools
Attendance Rates			
Mean	90%	92%	92%
SD	0.06	0.04	0.04
Disciplinary Incidents (per 100 students)			
Mean	0.4	0.2	0.2
SD	0.1	0.8	0.5

Table 1: Mean and SD of Outcome Variables in Umatched, Matched, and LIM schools

status, which is a special designation that demonstrates that the school has achieved a high level of fidelity and skill in the implementation of the LIM program. One question of interest is whether schools that have achieved Lighthouse status and those that no longer implement the program have different outcomes than "typical" LIM schools.

The LIM, like most other educational programs, has developed and improved over time. The program began to be offered by FranklinCovey Education in 2006 and since that time has seen two major revisions. The shift from 1.0 to 2.0 occurred before 2011 and does not concern us here as no schools in the analysis use or used 1.0. The shift from 2.0 to 3.0 occurred in January, 2015 and represented a major advance in the training, implementation, quality measurement, and curriculum. Thus, schools in the analysis that started LIM before 2015 primarily use the 2.0 version[1], while schools starting in or after 2015 use the 3.0 approach. Due to the significant improvements from 2.0 to 3.0, one important question of interest is whether treatment effects vary based on which version of the program the schools are using.

Missing Data. Despite the availability of data, there was a fair amount of missing data in the Missouri Department of Education's data records. Roughly 17% of attendance data is missing in treatment and control group and 6% of disciplinary incidents data are missing. Other relevant variables had missing rates of ranging from 3%-16%. Because of the level of missing data, especially in the attendance outcome variable, I chose to use multiple imputation by chained equations using the MICE package in R (van Buuren & Groothuis-Oudshoorn, 2011) to impute 25 complete data sets to use in the analyses. All school level variables were used for the imputation and results reported here represent pooled estimates across the 25 data sets as described by (van Buuren & Groothuis-Oudshoorn, 2011).

Analytic Approach

The basic approach taken in this report combines propensity score matching, prognostic scores, and a continuous interrupted time series analysis. The propensity score matching creates a set of matched comparison schools that are balanced across a range of relevant covariates. The prognostic

[1] All 2.0 Leader in Me schools were encouraged to transition to the 3.0 version which involved training staff in the 3.0 content or professional onsite coaching from former educators now trained in the LIM's 3.0 approach. Schools that engaged in the 3.0 training were counted as implementing the 3.0 program.

score adjusts outcome measures for variables unaffected by treatment status (i.e. school urbanicity), thereby reducing unexplained variance in the outcome, and increasing power. The interrupted time series then models the linear trends in outcomes of the comparison and treatment group, treating as the causal effect of interest, deviations from the baseline trend that occur in the treatment schools, but not the matched control schools. The combination of these two methods should provide strong, quasi-experimental evidence of the effectiveness of LIM.

Matching. The first step in the analyses involved matching the treatment schools to comparable control schools. I used propensity scores with full matching within a 0.2 caliper and with exact matching on highest grade (2-4th, 5-6th, or 7-8th) and lowest grade (Pre-K-K or 1-2nd) (Hansen & Klopfer, 2006). One challenge for the matching was that schools adopted the LIM program across the span of 8 years. Thus, the propensity score varied somewhat based on when schools adopted LIM, balancing matches on variables from the years before schools started LIM. For schools starting in 2011, the propensity model was based on 2009 variables; for schools starting between 2012 and 2015, the propensity model was based on 2011 variables; and for schools starting between 2016 and 2017, the propensity model was based on 2015 variables. Variables used in the propensity score included the following: % White, % Hispanic, % FRL, % IEP, urbanicity, FRL attendance rates, 2009 test score averages, total school size, attendance rate, number of disciplinary incidents, and % teachers with a masters. Matching was done independently across imputed data sets so matches varied across imputed data sets. Because of the restriction that schools match with similar schools that have the same grade range, some treatment schools (average of ~2.5 across imputed data sets) were unable to be matched to a group of comparison schools. One school was K-3rd grade only and did not match across any imputed data sets; one 3-6th grade school matched only in 8 data sets and one P-5th school matched in only 15 data sets. Three other schools, one P-6th, one 2-5th, and one P-1st school have no matches in 6, 2, and 2 imputed data sets, respectively. While restricting the grade ranges led to some schools not matching, this restriction was important since school level outcomes were used in the analyses. This is especially true as a number of LIM schools had unusual grade ranges, such as 4-6th. Matching was done using full matching with at most 4 control schools per treatment school (Hansen & Klopfer, 2006).

Table 2 and 3 below show the results of the matching. Table 2 shows the balance for 2010 data across a number of variables for the unmatched data (left set of columns) and the matched data set (right set of columns). The first two columns in each set show the mean values for the control (Treat_Sch=0) and intervention (Treat_Sch=1) schools, respectively. The third column shows the standardized differences (std. diff) between the two groups using randomization inference. Note that the recommendation is that standardized differences should be less than 0.25 (Ho, Imai, King, & Stuart; 2007). Table 3 provides the same information, but only for the matched sample and for the baseline year (i.e., t=0) and the year before baseline (i.e., t=-1). Note that Table 3 represents a range of calendar years since schools began the LIM program across eight years.

As seen in Table 2, before matching, LIM schools were larger, less poor, more white, had higher attendance rates, lower disciplinary incidents, more teachers with masters degrees, and higher test scores than other schools in the state in 2010 (and generally across years). This is to say, the LIM schools were relatively more advantaged and higher achieving than the typical MO schools. These differences largely disappeared after matching, though some differences remained for total attendance rates, IEP attendance rates, and free-reduced lunch attendance rates. However, all differences remained below the 0.25 recommended cut-off for differences between treatment and control groups and omnibus tests shows no differences between treatment and matched controls. Further, the more relevant question to judge the quality of matching is whether schools were balanced during the baseline year of implementation. As Table 3 shows, there is full balance between control and treatment groups in two years prior to treatment (i.e. baseline year and the year before the baseline year), both in omnibus tests and for each focal variable.

## strata	No.Strat				MatchesGrade			
## stat	Treat_Sch=0	Treat_Sch=1	std.diff		Treat_Sch=0	Treat_Sch=1	std.diff	
## vars								
## NumSt_2010	3.532	3.964	0.256	**	4.077	3.935	-0.084	
## FRL_2010	0.549	0.448	-0.445	***	0.470	0.446	-0.106	
## WHITE_2010	0.745	0.827	0.278	**	0.833	0.834	0.002	
## Atnd_TOTAL_2010	0.898	0.918	0.344	***	0.909	0.919	0.172	*
## Disc_INCIDENT_2010	1.135	0.509	-0.190	.	0.515	0.436	-0.024	
## T_YrsExp_2010	12.483	12.450	-0.012		12.594	12.449	-0.051	
## T_MA_2010	0.534	0.594	0.313	**	0.602	0.596	-0.030	
## SCALE_M_ALL_2010	651.926	657.370	0.375	***	656.544	657.364	0.057	

---Overall Test---

##	chisquare	df	p.value
## No.Strat	26.3	8	0.00092
## Matched	3.4	8	0.90994
## MatchesGrade	6.8	8	0.56192

Signif. codes: 0 '***' 0.001 '** ' 0.01 '* ' 0.05 '. ' 0.1 ' ' 1

## strata	No.Strat				MatchesGrade			
## stat	Treat_Sch=0	Treat_Sch=1	std.diff		Treat_Sch=0	Treat_Sch=1	std.diff	
## vars								
## SCALE_M_FRL_2010	643.744	646.593	0.226	*	646.139	646.940	0.064	
## Disc_10+_DAYS_2010	0.165	0.140	-0.032		0.108	0.116	0.011	
## Disc_VioWeap_2010	0.273	0.061	-0.241	*	0.158	0.063	-0.108	
## Atnd_BLACK_2010	0.880	0.887	0.090		0.883	0.889	0.070	
## Atnd_HISPANIC_2010	0.897	0.896	-0.015		0.902	0.895	-0.089	
## Atnd_IEP_2010	0.834	0.865	0.408	***	0.850	0.866	0.219	*
## Atnd_FRL_2010	0.862	0.873	0.186	.	0.865	0.875	0.166	.
## BLACK_2010	0.185	0.086	-0.358	***	0.098	0.083	-0.054	
## HISPANIC_2010	0.041	0.052	0.145		0.038	0.049	0.141	
## ELL_2010	0.028	0.030	0.023		0.027	0.029	0.034	
## IEP_2010	0.135	0.135	0.018		0.131	0.136	0.109	

---Overall Test---

##	chisquare	df	p.value
## No.Strat	35	11	0.00029
## Matched	16	11	0.15962
## MatchesGrade	17	11	0.11390

Signif. codes: 0 '***' 0.001 '** ' 0.01 '* ' 0.05 '. ' 0.1 ' ' 1

Table 2. 2010 Balance Tables

Baseline Year Balance

```
##               strata MatchesGrade
##               stat          LIM=0     LIM=1  std.diff
## vars
## GRADES_K_12             4.04757   4.01078   -0.02211
## FRL                     0.47975   0.46479   -0.06526
## WHITE                   0.79784   0.79796   0.00056
## Attend_TOTAL            0.92412   0.92586   0.04267
## Disc_INCIDENT           0.82007   0.53201   -0.06715
## T_YrsExp                12.37051  12.35587  -0.00584
## T_MA                    0.60393   0.62759   0.13379
## SCALE_M_ALL             656.98415 657.88429 0.07105
## ---Overall Test---
##              chisquare df p.value
## MatchesGrade       3.8  8    0.87
## ---
## Signif. codes:  0 '***' 0.001 '** ' 0.01 '*  ' 0.05 '.  ' 0.1 '   ' 1
```

Year before Baseline Balance

```
##               strata MatchesGrade
##               stat          LIM=0     LIM=1  std.diff
## vars
## GRADES_K_12             4.06345   4.01783   -0.02702
## FRL                     0.49570   0.46079   -0.15602  .
## WHITE                   0.80237   0.80218   -0.00098
## Attend_TOTAL            0.92046   0.92287   0.05722
## Disc_INCIDENT           0.72261   0.63398   -0.03153
## T_YrsExp                12.39286  12.32419  -0.02653
## T_MA                    0.59813   0.61553   0.09923
## SCALE_M_ALL             657.26728 657.62555 0.02824
## ---Overall Test---
##              chisquare df p.value
## MatchesGrade       5.1  8    0.75
## ---
## Signif. codes:  0 '***' 0.001 '** ' 0.01 '*  ' 0.05 '.  ' 0.1 '   ' 1
```

Table 3. Covariate Balance for Baseline Year and Year before Baseline Year

Prognostic Scores. Prognostic scores represent an approach to controlling for variation in outcome measures that seeks to reduce the likelihood of Type 1 errors and increasing power (Hansen, 2006). Similar to matching, this is achieved by separating the process of creating a prognostic score that predicts the outcome from the process of examining treatment effects. Further, prognostic scores allow for matching to be "doubly robust" (Hansen, 2006). A prognostic score is a prediction of the outcome of interest, usually generated from some left out sample (here the unmatched schools) that is used to predict the expected outcome. This predicted outcome then can be used as a regression coefficient, reducing unexplained variance in the outcome and eliminating the temptation to run several models until finding a model that produces the desired outcome (Hansen, 2006). Here, I used the sample of schools in MO that were neither treatment schools nor matched control schools to train a random forest prognostic score using the variables YEAR, school size, % FRL, % White, teacher experience, and % teachers with master's degrees. Note that all of these prediction variables are unaffected by the LIM program. This random forest explained 51% of the variance in attendance rates and 46% of the variance in discipline rates. The random forest models were used to estimated expected outcomes (i.e., prognostic scores) for each outcome that were used in the interrupted time series models as predictors.

Analyses. Because of the longitudinal nature of the data, I used a short-interrupted time series design to analyze the effectiveness of LIM (described in detail below). This analysis uses the baseline data to create linear trend-line of the outcomes for both the treatment and matched control schools and estimates deviations from these baseline trends that coincide with LIM adoption in treatment schools, but not matched comparison schools as the causal treatment effect. Thus, the analysis controls for both baseline outcome values, trends in baseline outcome values, and global impacts that coincide with treatment (which effect both comparison and treatment schools).

The model being used is:

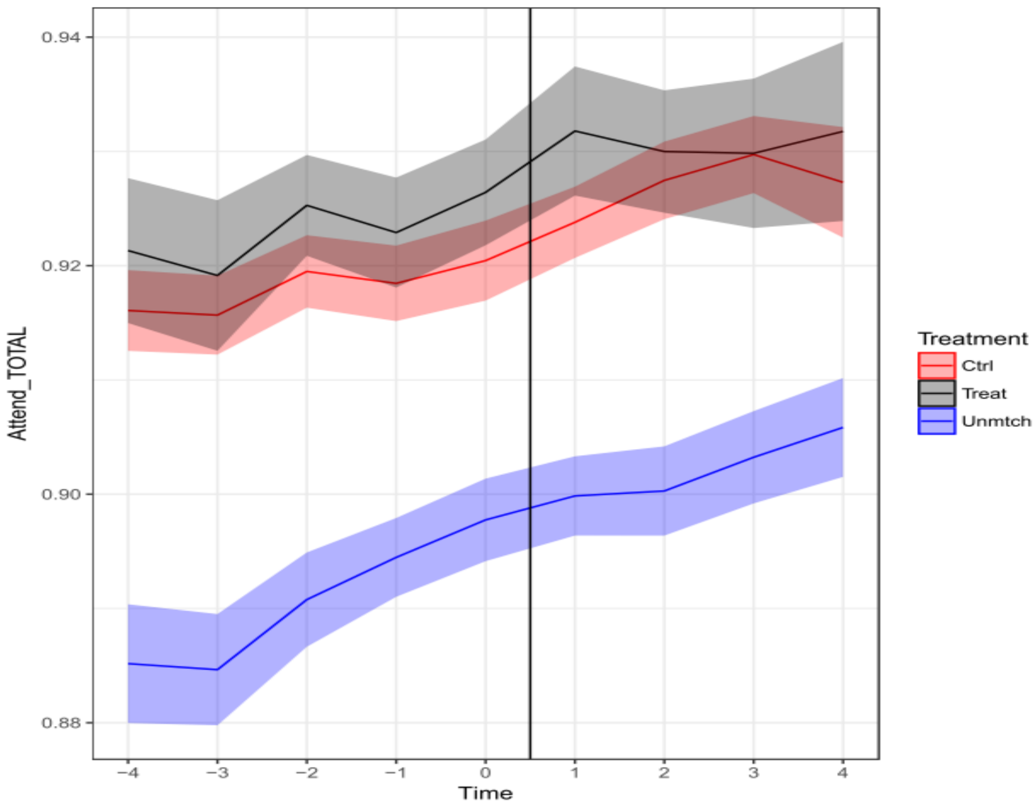
$$out_{tim} = \beta_0 + (MatchT)_{ti} + \left(Att_{Prog}\right)_{ti} + \gamma_m + (LIM)_i + (Time\ After\ Treat)_{ti} + (TreatITT)_{ti}$$

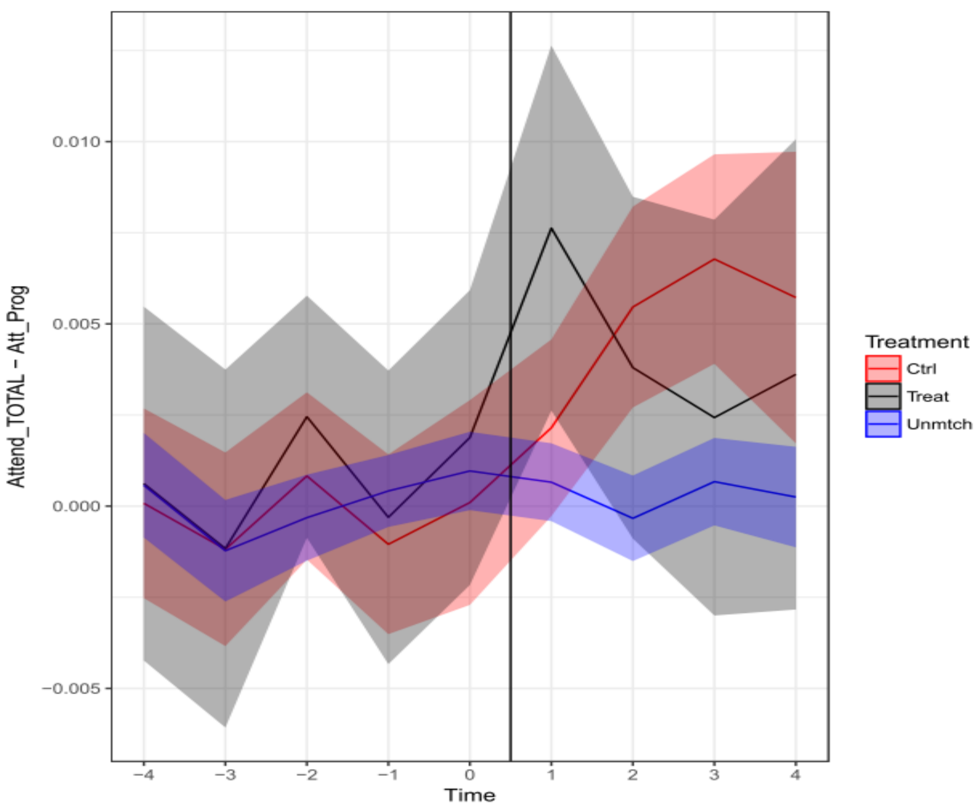
where *t* is time, *i* is school, *m* is match, γ_m are match fixed effects (matches is equations); out_{tim} is the outcome of interest; $\beta_0 + (MatchT)_{ti}$ is the intercept and slope for the baseline time trend, which are allowed to vary randomly across schools; $\left(Att_{Prog}\right)_{ti}$ is the prognostic score that predicts outcome from demographics and school characteristics unaffected by treatment; $(LIM)_i$ allows the time trend intercept for treatment schools to be different for treatment schools (and this term is sometimes interacted with the slope to allow different slopes across treatment and control schools, though this interaction can tend to over-fit based on work by Jacobs and colleagues); $(Time\ After\ Treat)_{ti}$ (labeled I(MatchT>0) below) is an indicator for whether the time is after treatment began, allowing for shocks contemporaneous with the start of treatment that affect both control and treatment schools; and $(TreatITT)_{ti}$ is the causal treatment impact, which is estimated as the deviation from the baseline trend for treatment schools beyond any deviation that occurred in matched control schools. Some models replace$(TreatITT)_{ti}$ with $(Treat)_{ti}$, which provides a Treatment on Treated effect estimate. This estimation approach controls for both baseline scores and trends in baseline scores, as well as shocks contemporaneous with treatment that affect matched controlled schools and treatment schools equally. The number of discipline incidents was modeled using the same linear mean estimate, but using a Poisson regression with log link function and offset for the log of the school size.

Results

Attendance.

The graph below shows the mean attendance rates over time graphically, divided by treatment, control, and unmatched schools with a vertical line indicating the start of LIM and shading to indicate [sqrt(2)**SE*] of the mean, so that overlapping shaded areas indicate non-significant differences. The graph shows the treatment schools have marginally higher attendance rates (this is accounted for in modeling), though the treatment and control schools are far more similar than the unmatched schools. The graph also shows the instability of attendance rates, a positive trend starting in the baseline time point and no clear or obvious deviation from the trends coinciding with treatment. The second graph below shows the Attendance rates minus the prognostic score, showing the raw trends after removing predicted values. Here the three groups are more similar pre-treatment with the treatment and control schools showing a clear upward trajectory coinciding with treatment.


0.94
0.92
0.90
0.88
Attend_TOTAL
−4
−3
−2
−1
0
1
2
3
4
Time
Treatment
Ctrl
Treat
Unmtch

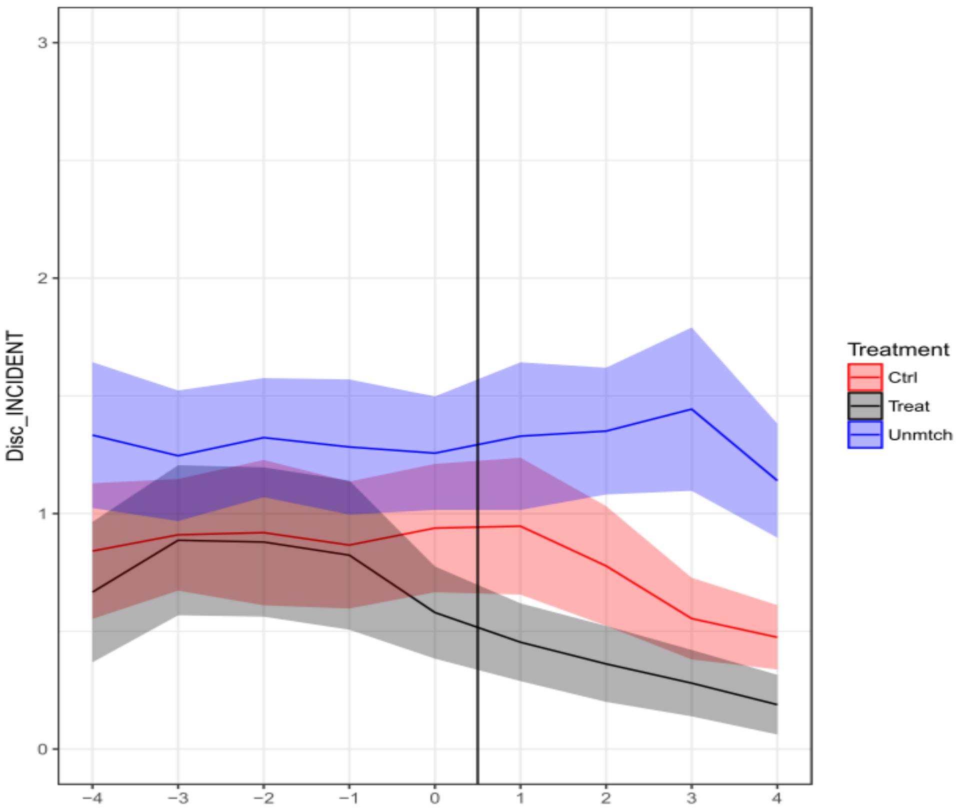

0.010
0.005
0.000
−0.005
Attend_TOTAL - Att_Prog
−4
−3
−2
−1
0
1
2
3
4
Time
Treatment
Ctrl
Treat
Unmtch

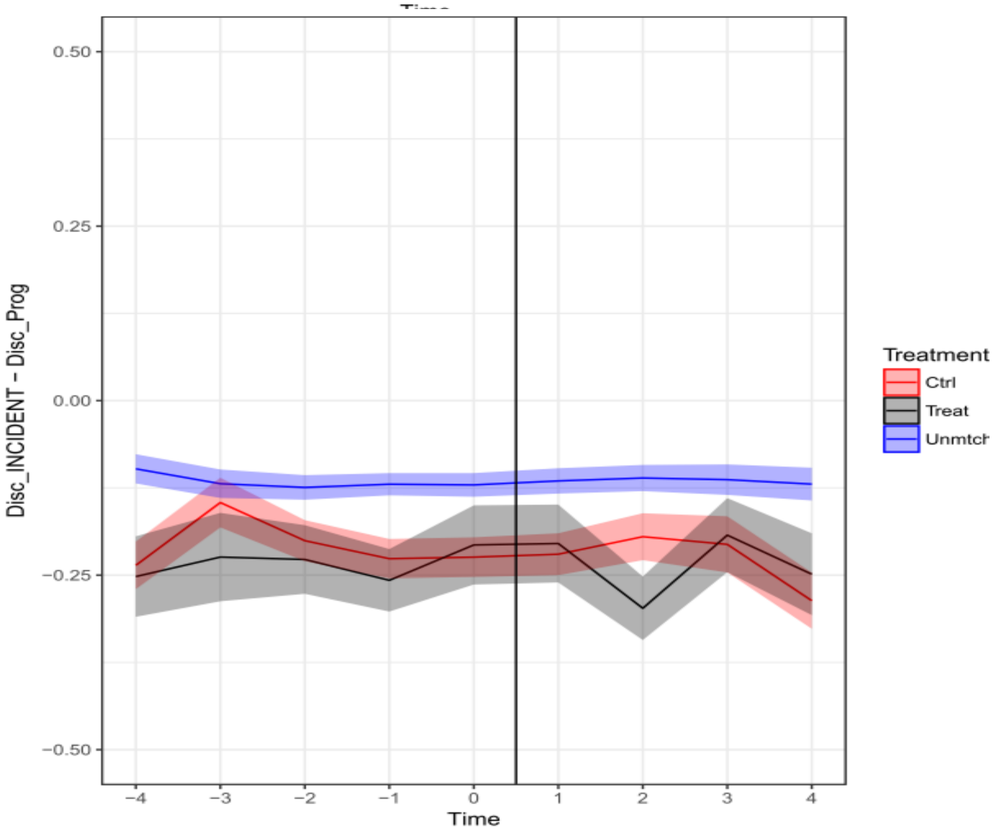
The point estimate of the intent to treat (ITT) effect was near 0 and non-significant. I also ran the two planned exploratory analyses testing for subgroup effects. First, I looked for differential effects on LIM 2.0 and LIM 3.0 schools. The estimate for 2.0 schools was non-significant and slightly negative. The estimate for LIM 3.0 schools was a marginally significant estimate (*beta*=0.006; *p*=0.059; *ES*~0.1). LIM 3.0 schools have about 0.6% higher attendance rates than would be expected due to LIM, which corresponds to an effect size of 0.1, or a small effect. The planned exploration of differential effects by school status (regular, dropped, lighthouse) showed that schools that eventually dropped LIM had a significantly negative lower than expected observed attendance rate post-treatment (*beta*= -0.018; *p*<0.001). The effect for lighthouse schools (*beta*=0.005; *p*=0.15) and non-lighthouse schools (*beta*=0.002; *p*=0.60) was positive, but non-significant.

Given the negative effect of schools that dropped the LIM program, it is of interest to explore the treatment on treated (TOT) effects of LIM. These effects are not causal, but estimate the benefit for a school that adopts and sticks with LIM over time. The overall TOT effect on attendance was not significant (*beta*=0.004; *p*=0.13; *ES*~0.07). The TOT effect on LIM 3.0 schools is significant (*beta*=0.009; *p*=0.007; *ES*~0.15). Thus, schools that adopt the LIM 3.0 program and stick with the program are likely to see an increase of almost 1% in attendance rates (effect size of 0.15). The estimated TOT effect for Lighthouse schools is marginally significant (*beta*=0.007; *p*=0.061; *ES*~0.12). Lighthouse schools that stick with the program are likely to see a 0.7% increase in attendance rates.

Discipline

The graph below shows the mean discipline incidents over time graphically, divided by treatment, control, and unmatched schools with a vertical line indicating the start of LIM and shading to indicate [*sqrt*(2)**SE*] of the mean, so that overlapping shaded areas indicate non-significant differences. The first graph shows similar treatment and control levels in the baseline time period with both groups decreasing in discipline incidents after treatment and the treatment group possibly decreasing in discipline incidents during the baseline year. The second graph shows the discipline incidents minus the prognostic score, showing the control and treatment groups as more similar.


Disc_INCIDENT
Treatment
Ctrl
Treat
Unmtch
3
2
1
0
−4
−3
−2
−1
0
1
2
3
4


Disc_INCIDENT − Disc_Prog
Treatment
Ctrl
Treat
Unmtch
0.50
0.25
0.00
−0.25
−0.50
−4
−3
−2
−1
0
1
2
3
4
Time

Recall that these models used a Poisson regression with log link functions so results reflect multiplicative increases or decreases in discipline incidents. There is a significant treatment effect of LIM on discipline incidence rates (*beta*=-0.551; *exp(beta)*= 0.58; *p*<0.001). Thus, LIM schools are predicted to have 58% of the discipline incidents they would have had had they not adopted LIM. There are a few ways to think about the size of this effect. The first uses the analytic mode to calculate the estimated base rate of the discipline rate as 0.0005 or 0.2 incidents per year for an average school of 400 students. LIM schools then are estimated to have about 0.12 discipline incidents per year in an average size school. However, the means in the interpretation are means for a Poisson distribution which are not wholly intuitive. A more direct way to estimate the size of the treatment effect is to use a differences-in-differences table (see Table 4), which averages the discipline rates per 100 students separately for control and treatment schools and pre-/post-treatment. LIM schools go from 0.193 discipline incidents per 100 students before treatment to 0.107 incidents per student post-treatment or a reduction of 0.086 incidents per 100 students. The control schools, on the other hand have a 0.004 increase in discipline rates per 100 students, giving a causal effect of LIM schools of 0.090 incidents per 100 students or roughly 0.36 incidents per year for an average size school of 400 students. Thus, an average size LIM school is predicted to have 0.36 fewer discipline incidents than if they had not adopted the LIM program. While this effect seems small, it represents a large percentage decrease in discipline incidents.

	Pre-Treatment	**Post-Treatment**
Treatment Schools	0.193	0.107
Control Schools	0.203	0.207

Table 4. Difference-in-Difference Table of LIM Effect on Discipline Rate (per 100 students)

Breaking down LIM schools into LIM 2.0 and LIM 3.0, we see there is only a significant effect for both LIM 2.0 (*beta*=-0.433; exp(beta)=.65; *p*=0.010) and 3.0 schools (*beta*=-0.721; exp(beta)=.49; *p*<0.001). Unlike for attendance rates, however, these effects seem to be driven by "normal" LIM schools, which have a significant effect (*beta*=-0.785; exp(beta)=0.45; *p*<0.001), rather than lighthouse schools or schools that eventually drop LIM for which there is no significant effect.

Conclusion

This report looked at the causal impact of the LIM program on the population of LIM schools in MO using matching and an interrupted time series design. It found that there was no overall effect of the intervention on attendance rates, but LIM led to significant decreases in the rate of disciplinary incidents (by about 60% or 0.36 for an average sized school). Further, the LIM 3.0 program had a marginally significant effect on attendance rates, increasing attendance by 0.6% (*ES*=0.1). The effect of LIM 3.0 schools, when estimated as a treatment on the treated effect was significant, increasing attendance rates by close to 1% (*ES*=0.15). However, schools that eventually dropped the LIM program were found to have lower than expected attendance rates, suggesting possibly negative impacts for schools that do not stick with the intervention.

The purpose of this report was to explore overall impacts of LIM using publically available attendance and disciplinary data. Follow-up analyses will further explore impacts of implementation fidelity, professional coaching, and training content (i.e., 2.0 vs. 3.0). Future studies should focus on more proximal measures of student behavior, using a finer grain size than school average scores and focusing on behaviors more immediately targeted by the LIM program, such as leadership skills or behaviors.

References

Hansen, B. B. (2006). *Bias reduction in observational studies via prognosis scores*. Technical Report 441, University of Michigan, Statistics Department.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627. https://doi.org/10.1198/106186006X137047

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*, 199–236. https://doi.org/10.1093/pan/mpl013

Jacob, R., Somers, M.-A., Zhu, P., & Bloom, H. (2016). The Validity of the Comparative Interrupted Time Series Design for Evaluating the Effect of School-Level Interventions. *Evaluation Review*, *40*(3), 167–198. https://doi.org/10.1177/0193841X16663414

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03