considerably and significantly in the treatment schools, both in absolute terms and relative to the control schools. Thus, the differences-in-differences estimate based on 1994-to-1996 changes is .62 for reading scores and .46 for math scores. This implies the program increased test scores by roughly half a standard deviation for both reading and math. The estimates are smaller when 1994-to-1995 changes are used, especially for math scores. However, math scores in treatment schools increased between 1994 and 1995 and between 1995 and 1996, while reading scores increased only between 1994 and 1995.

The differences-in-differences estimates for secular schools are significant whether conventional standard errors or Moulton standard errors are used to measure precision. It is important to note, however, that the Moulton standard errors are 2–3 times larger than the conventional standard errors. This reflects the fact that, even though the within-school correlation between pupils' test scores is not very large (on the order of .1), the model being estimated has features that cause a small amount of within-group correlation to have a large effect on regression standard errors. First, the regressor (treatment status) is fixed within groups. Second, there are not many groups (schools), so the group size is large relative to the sample. Moulton (1986) shows that when the regressor of interest is fixed within groups and the groups are large, conventional standard errors may seriously overestimate the precision of regression estimates.

Estimates for religious schools are reported in panel B of table 5. The data for religious schools do not include reading scores for 1996. Between 1994 and 1995, test scores in the religious treated schools fell sharply, both in absolute terms and relative to control schools. This can be seen in columns 1 and 4 of the table. As noted earlier, the treatment in the religious schools started only in September of 1995 (i.e., the beginning of the 1996 school year), so the 1994–95 score decline in treated schools is probably unrelated to the intervention. A consequence of the decline in pretreatment years is that the differences-in-differences estimates of effects on math scores are positive when 1995 is taken as the base year and negative when 1994 is taken as the base year (see cols. 5 and 6). Sensitivity to the choice of base year implies that the assumptions underlying the differences-in-differences strategy are not satisfied (Ashenfelter and Card 1985). In fact, the results for religious schools seem generally less reliable and harder to interpret than those for secular schools. The treated religious sample includes fewer than 100 pupils, from only two schools. Moreover, as noted earlier, we do not have data on inputs in religious control schools.

\* Since most religious schools combine primary and middle grades, they do not use the reading test that is used in secular schools for placing students in the transition from primary to middle school.
Teacher Training and Pupil Learning

while the data on treated schools actually show a decline in teacher training between 1994 and 1995.

B. Regression Estimates

Comparisons of means by treatment status show that pupils in treated schools had lower test scores than pupils in control schools in 1994, before treatment began. Thus, pupils in the two sets of schools are not directly comparable. The differences-in-differences strategy is motivated by a model where the difference between pupils in treatment and control schools is captured by the fixed effect in equation (1a), \( \phi \). Differences-in-differences estimates do not have a causal interpretation, however, if the lower pretreatment scores in treatment schools are due to temporarily low scores as opposed to permanent differences. In such cases, test-score growth in the treatment schools will tend to be larger than test-score growth in the control schools regardless of the program effect, generating spurious positive estimates of treatment effects.\(^{15}\)

To avoid this sort of bias, the regression approach discussed in Section III replaces the fixed-effects assumption with the assumption that pupils in treatment and control schools are comparable conditional on past test scores and other observed covariates. The resulting estimates for secular schools are reported in Table 6. The first column shows results from a specification that includes only the treatment dummy as an explanatory variable. The estimates reported in column 2 are from a model where a limited set of student characteristics (year of birth, sex, immigrant status) were added as controls. The estimates in column 3 are from a model where additional controls (parental schooling, family status, and dummy variables for continent of origin) were added to the list of regressors. Column 4 shows the results of adding lagged test scores, and column 5 shows the results from a model where lagged test scores are the only control variable. Columns 1–5 are for reading scores, and columns 6–10 repeat this sequence for math scores. As before, conventional standard errors are reported in parentheses, and standard errors corrected for within-school correlation are reported in brackets.

The regression results are reported in separate rows for test scores in 1994, 1995, and 1996. The 1994 results are shown because it is of interest to know whether the pretreatment differences between pupils in treatment and control schools are explained by differences in observed covariates. The raw difference in 1994 reading scores by treatment status is \(-.74\), and the regression-adjusted difference in column 3 is \(-.563\). The raw difference in 1994 math scores is \(-.37\), and the regression-adjusted difference in column 8 is \(-.28\). Thus, most of the difference in 1994 scores

\(^{15}\) See, e.g., Ashenfelter and Card (1985).


<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Reading</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Math</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Control (1)</td>
<td>Basic Control (2)</td>
<td>Extended Control (3)</td>
<td>Extended and 1994 Score Only (4)</td>
<td>1994 Score Only (5)</td>
<td>No Control (6)</td>
<td>Basic Control (7)</td>
<td>Extended Control (8)</td>
<td>Extended and 1994 Score Only (9)</td>
</tr>
<tr>
<td>A. All secular schools:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994 score</td>
<td>-.744</td>
<td>.709</td>
<td>.563</td>
<td>...</td>
<td>...</td>
<td>-.370</td>
<td>-.362</td>
<td>-.284</td>
<td>...</td>
</tr>
<tr>
<td>(0.072)</td>
<td>(0.072)</td>
<td>(0.074)</td>
<td></td>
<td></td>
<td></td>
<td>(0.080)</td>
<td>(0.079)</td>
<td>(0.086)</td>
<td></td>
</tr>
<tr>
<td>(0.248)</td>
<td>(0.242)</td>
<td>(0.204)</td>
<td></td>
<td></td>
<td></td>
<td>(0.150)</td>
<td>(0.150)</td>
<td>(0.162)</td>
<td></td>
</tr>
<tr>
<td>1995 score</td>
<td>-2.231</td>
<td>-1.95</td>
<td>-0.49</td>
<td>2.21</td>
<td>.173</td>
<td>-1.124</td>
<td>-0.91</td>
<td>-0.02</td>
<td>0.57</td>
</tr>
<tr>
<td>(0.072)</td>
<td>(0.071)</td>
<td>(0.076)</td>
<td>(0.076)</td>
<td>(0.071)</td>
<td></td>
<td>(0.076)</td>
<td>(0.075)</td>
<td>(0.088)</td>
<td>(0.076)</td>
</tr>
<tr>
<td>(0.229)</td>
<td>(0.197)</td>
<td>(0.201)</td>
<td>(0.204)</td>
<td>(0.217)</td>
<td></td>
<td>(0.243)</td>
<td>(0.222)</td>
<td>(0.256)</td>
<td>(0.243)</td>
</tr>
<tr>
<td>1996 score</td>
<td>.128</td>
<td>-1.16</td>
<td>.069</td>
<td>.315</td>
<td>.281</td>
<td>.089</td>
<td>.082</td>
<td>.138</td>
<td>.260</td>
</tr>
<tr>
<td>(0.081)</td>
<td>(0.082)</td>
<td>(0.084)</td>
<td>(0.082)</td>
<td>(0.078)</td>
<td></td>
<td>(0.072)</td>
<td>(0.072)</td>
<td>(0.076)</td>
<td>(0.067)</td>
</tr>
<tr>
<td>(0.144)</td>
<td>(0.139)</td>
<td>(0.095)</td>
<td>(0.130)</td>
<td>(0.152)</td>
<td></td>
<td>(0.158)</td>
<td>(0.151)</td>
<td>(0.145)</td>
<td>(0.146)</td>
</tr>
<tr>
<td>B. Secular schools, matched sample:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994 score</td>
<td>-.468</td>
<td>-.399</td>
<td>-.373</td>
<td>...</td>
<td>...</td>
<td>-.086</td>
<td>-.083</td>
<td>.035</td>
<td>...</td>
</tr>
<tr>
<td>(0.087)</td>
<td>(0.088)</td>
<td>(0.092)</td>
<td></td>
<td></td>
<td></td>
<td>(0.097)</td>
<td>(0.096)</td>
<td>(0.105)</td>
<td></td>
</tr>
<tr>
<td>(0.321)</td>
<td>(0.299)</td>
<td>(0.275)</td>
<td></td>
<td></td>
<td></td>
<td>(0.122)</td>
<td>(0.122)</td>
<td>(0.153)</td>
<td></td>
</tr>
<tr>
<td>1995 score</td>
<td>-.411</td>
<td>-.341</td>
<td>-.284</td>
<td>-.045</td>
<td>-.102</td>
<td>-.016</td>
<td>-.012</td>
<td>.053</td>
<td>.030</td>
</tr>
<tr>
<td>(0.096)</td>
<td>(0.096)</td>
<td>(0.103)</td>
<td>(0.096)</td>
<td>(0.087)</td>
<td></td>
<td>(0.084)</td>
<td>(0.083)</td>
<td>(0.096)</td>
<td>(0.082)</td>
</tr>
<tr>
<td>(0.215)</td>
<td>(0.178)</td>
<td>(0.152)</td>
<td>(0.139)</td>
<td>(0.175)</td>
<td></td>
<td>(0.284)</td>
<td>(0.298)</td>
<td>(0.310)</td>
<td>(0.290)</td>
</tr>
<tr>
<td>1996 score</td>
<td>-.063</td>
<td>-.063</td>
<td>.008</td>
<td>.178</td>
<td>.173</td>
<td>.147</td>
<td>.148</td>
<td>.241</td>
<td>.228</td>
</tr>
<tr>
<td>(0.098)</td>
<td>(0.100)</td>
<td>(0.102)</td>
<td>(0.096)</td>
<td>(0.091)</td>
<td></td>
<td>(0.085)</td>
<td>(0.084)</td>
<td>(0.091)</td>
<td>(0.082)</td>
</tr>
<tr>
<td>(0.192)</td>
<td>(0.183)</td>
<td>(0.111)</td>
<td>(0.162)</td>
<td>(0.206)</td>
<td></td>
<td>(0.195)</td>
<td>(0.189)</td>
<td>(0.188)</td>
<td>(0.200)</td>
</tr>
</tbody>
</table>

Note.—Conventional standard errors are reported in parentheses. Standard errors corrected for within-school correlation are reported in square brackets. Standardized scores were used in this analysis.
between treatment and control pupils remains even after accounting for treatment-control differences in the observed covariates.

Both 1995 and 1996 are posttreatment years for secular schools. The estimated treated effects on 1995 and 1996 scores in models that do not control for pretreatment test scores, reported in columns 1–3, are negative. These negative estimates are not surprising since pupils in treated schools had lower test scores before the treatment began, and since this difference is not accounted for by the included covariates. Controlling for 1994 scores, however, the treatment effects become positive in both 1995 and 1996, and at least marginally significant in 1996, whether or not covariates other than the 1994 scores are included. For example, the effect of the treatment on 1996 reading scores is .32 with a (corrected) standard error of .13 in a model with lagged scores and the extended set of control variables (reported in col. 4). Dropping covariates reduces the estimate to .25. The estimate for math scores controlling for lagged test scores and the full set of covariates is .26 with a standard error of .15. For both math and reading scores, this is about half as large as the corresponding differences-in-differences estimate, suggesting that the differences-in-differences estimates are biased upward.

In addition to results for the full sample of secular schools, we also computed regression results for a matched sample with similar 1994 school-level average scores in treatment and control schools. As noted earlier, this is a nonparametric procedure for making the treatment and control schools as comparable as possible. An attractive feature of this approach is that it reduces the likelihood of bias from school effects that are correlated with treatment status. The matched sample was constructed by selecting groups or pairs of treatment and control schools where the difference in average math scores in 1994 is less than .1 (in standard deviation units). Treatment schools for which no control school could be found with a comparable average score were discarded and vice versa. The matched subsample includes five treatment and three control schools. The results of analyzing this subsample are reported in panel B of table 6. Columns 6–8 of the table show that matching effectively eliminates treatment-control differences in 1994 math scores, though not in reading scores (we were not able to match schools very closely on 1994 reading scores and therefore use the same matched sample for analysis of both math and reading scores).

Although they are less precise, the results for math scores in the matched subsample are generally similar to those in the full sample. An important difference, however, is that since the schools have already been chosen to have similar math scores in 1994, the estimated effects on 1996 math scores are positive even without controlling for 1994 scores. For example, the estimate in column 8, from a model with the extended set of covariates but no lagged test score, is .241 with a corrected standard error of .19.
For reading scores, the results in the matched sample show smaller treatment gaps in 1994 than in the full sample, but the difference by treatment status for 1994 reading scores is not completely eliminated. Controlling for 1994 reading scores, the estimated effect on 1996 reading scores is positive as before, though not significantly different from zero and smaller than in the full sample.

The last set of regression estimates, for religious schools, is reported in table 7. These results are for math scores only, since there are no reading scores in 1996 (the only posttreatment year in religious schools). Columns 1-3 show that test scores in treated schools were higher in 1994 but much lower in 1995, regardless of which covariates are included in the regression. In contrast with the estimates for secular schools, adding the 1994 score to the equation for 1996 scores actually makes the estimated treatment effect more negative. However, models that include the 1995 score as a control variable, the results of which are reported in column 5, lead to a positive treatment effect (though not significantly different from zero). This sensitivity to the choice of lagged control variable is similar to the sensitivity to the choice of base year observed in the differences-in-differences estimates for religious schools. Finally, column 6 shows that adding both pretreatment scores to the regression for religious schools leads to a very small negative estimate that is not significantly different from zero. Since religious treatment and control schools clearly differed in both pretreatment years, this last estimate seems most credible. The absence of an effect in religious schools may be due to the fact that the training program began later than in the secular schools and because the scale of the intervention in religious schools was smaller.
C. Additional Matching Estimates

As a final check on the estimates for secular schools, we matched individual pupils on the basis of their 1994 test scores, instead of school averages as in panel B of table 6. An advantage of pupil matching over school matching is that for both reading and math it is possible to find pupils with similar 1994 scores in all treatment and control schools even though their school averages may differ. The pupil-matching strategy was implemented by dividing the distribution of 1994 test scores into quartiles, comparing treatment and control scores in each quartile, and then summing the quartile-by-quartile treatment effects into a single weighted average with weights given by the distribution of treated observations across quartiles. Assuming the only difference between treatment and control pupils besides the treatment is their 1994 scores, this procedure produces a nonparametric estimate of the effect of treatment on the treated. We chose to match pupils based on quartiles of the 1994 score distribution instead of a finer breakdown because it turns out that division into four groups is enough to eliminate almost all treatment-control differences in 1994 test scores.

The only significant within-quartile treatment-control difference in 1994 scores is for math in the first quartile. This can be seen in column 3 of table 8, which reports the within-quartile difference in 1994 scores for math and reading. The overall average effect of treatment status on 1994 scores (i.e., the average across quartiles weighted by the number treated in each quartile) is .031 with a standard error of .031 for math and -.052 with a standard error of .04 for reading. This suggests that the quartile-matching strategy does indeed balance the treatment and control groups.11

The matching estimates of treatment effects on 1996 scores are reported in column 6 of table 8. Except for the upper quartile of the reading score distribution, the within-quartile contrasts for 1996 are all positive and, in some cases, individually significant. The overall matching estimate is .25 with a standard error of .16 for math scores and .4 with a standard error of .16 for reading scores. The matching estimates for math are virtually identical to the corresponding regression estimates (reported in col. 10 of table 6). The matching estimates for reading are somewhat larger than the corresponding regression estimates (reported in col. 5 of table 6). Thus, these results reinforce the finding that pupils with same 1994 test scores did better in 1996 if they were in schools where teachers received additional training.

11 The standard errors reported in this table were corrected for within-school correlation in test scores.
Table 8
Matching on 1994 Scores—Estimates for Secular Schools

<table>
<thead>
<tr>
<th>Subject</th>
<th>1994 Scores</th>
<th>1996 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Math, 1994 quartile:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>-1.483</td>
<td>-1.611</td>
</tr>
<tr>
<td></td>
<td>(.062)</td>
<td>(.055)</td>
</tr>
<tr>
<td>II</td>
<td>-.350</td>
<td>-.364</td>
</tr>
<tr>
<td></td>
<td>(.055)</td>
<td>(.055)</td>
</tr>
<tr>
<td>III</td>
<td>.339</td>
<td>.370</td>
</tr>
<tr>
<td></td>
<td>(.055)</td>
<td>(.055)</td>
</tr>
<tr>
<td>IV</td>
<td>1.045</td>
<td>1.070</td>
</tr>
<tr>
<td></td>
<td>(.057)</td>
<td>(.057)</td>
</tr>
<tr>
<td>Average effect</td>
<td>.031</td>
<td>.250</td>
</tr>
<tr>
<td></td>
<td>(.031)</td>
<td>(.158)</td>
</tr>
</tbody>
</table>

Reading, 1994 quartile:

<table>
<thead>
<tr>
<th></th>
<th>1994 Scores</th>
<th>1996 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>I</td>
<td>-1.433</td>
<td>-1.355</td>
</tr>
<tr>
<td></td>
<td>(.069)</td>
<td>(.069)</td>
</tr>
<tr>
<td>II</td>
<td>-.589</td>
<td>-.538</td>
</tr>
<tr>
<td></td>
<td>(.058)</td>
<td>(.058)</td>
</tr>
<tr>
<td>III</td>
<td>.164</td>
<td>.200</td>
</tr>
<tr>
<td></td>
<td>(.055)</td>
<td>(.055)</td>
</tr>
<tr>
<td>IV</td>
<td>.969</td>
<td>.957</td>
</tr>
<tr>
<td></td>
<td>(.063)</td>
<td>(.063)</td>
</tr>
<tr>
<td>Average effect</td>
<td>-.052</td>
<td>.399</td>
</tr>
<tr>
<td></td>
<td>(.040)</td>
<td>(.157)</td>
</tr>
</tbody>
</table>

Note: Standard errors are reported in parentheses. Standardized test scores were used in this analysis. The standard errors are corrected for within-school correlation in test scores.

V. The Cost-Effectiveness of Training Inputs

The estimates in tables 6 and 8 suggest that a conservative estimate of the effect of the training program in nonreligious schools is .25σ, where σ denotes the standard deviation of pupils’ test scores. The overall cost of the program in Jerusalem was about $12,000 per class. To determine whether this expenditure was worthwhile, we would need to estimate the value of an increase in children’s test scores, something that is hard to quantify. However, we can compare the cost of attaining a given increase in test scores obtained by increasing teacher training with estimates of the cost of achieving a similar increase in test scores obtained by reducing class size or lengthening the school day. This tells us which of these three important inputs is most likely to be worth increasing.

Angrist and Lavy (1999) estimated that reducing the maximum class size in Israel from 40 to 30 would raise test scores by .15σ and require a 28% increase in the number of classes. An unpublished memorandum

---

12 This is broadly consistent with estimates of effect size for other inputs reported in the education research literature (see, e.g., Hedges et al. 1994).
from the Ministry of Education suggests that the annual operating cost of each new class would be about $75,000, so the proposed reduction in class size would cost about $21,500 per existing class. Assuming the relationship between class size and performance is linear, we estimate that it would cost $35,000 per existing class to achieve a .25σ test-score gain. Using the same data, Lavy (1998) estimated that lengthening the school week by 3.8 hours would raise test scores by .15σ. Since the annual cost of an extra weekly hour of instruction is $2,000 for each class, this would cost $7,600 per class. Again, assuming a linear relationship between inputs and performance, the cost of a .25σ increase in scores is estimated to be about $12,600. These calculations suggest that if the objective is improving pupil achievement, teacher training may be at least as cost-effective a strategy as lengthening the school day and considerably cheaper than reducing class size.

VI. Summary and Conclusions

The relationship between teachers' characteristics and their pupils' achievement has been the subject of many studies, but few have looked at the impact of in-service training. Our estimates suggest that an in-service training program run in Jerusalem's secular elementary schools raised children's achievement in reading and mathematics. These findings appear using a variety of statistical methods, including differences-in-differences, regression, and matching. The estimates for religious schools are not clear cut, but this is possibly because the training program in religious schools started later and was implemented on a smaller scale. The estimates for secular schools suggest that teacher training may provide a less costly means of increasing test scores than reducing class size or adding school hours.

Of course, the question remains whether the particular training program studied here is similar to training programs that might be used in other settings. Discussions with school officials lead us to believe that the approach taken in Jerusalem is not unusual in the Israeli context. Moreover, the type of training given to reading and math teachers using 30 Towns money was based on widely used pedagogical strategies ("humanistic mathematics" and "individualized instruction") originally developed in U.S. schools. However, one possibly unique feature of the 30 Towns intervention is the fact that it was closely linked to a weekly curriculum. Another important and possibly unique feature is the immediate feedback that participating teachers received from instructors. Since the basic philosophies behind the training strategies studied here are not especially foreign, however, it seems reasonable to imagine that the Israeli variant on training delivery could also be adapted for non-Israeli schools. At a
minimum, the results here suggests that highly focused but relatively inexpensive teacher training of this type warrants further study.

Appendix

The 30 Towns Project

Teachers' In-Service Training

From the program's inception in September 1994, math and Hebrew teachers in program schools received in-house training on teaching methods. The training was carried out by instructors who met weekly with the math and Hebrew teachers for each grade, separately, and reviewed teaching methods for the material to be covered the following week. The math teachers received training based on the "humanistic mathematics" philosophy of teaching, developed by White (1987, 1993) for U.S. schools.

Hebrew training was carried out by two organizations: the Ministry of Education (six schools) and the Center for Educational Technology (four schools), which develops educational software marketed under well-known brand names worldwide. Curriculum for Hebrew teachers was based on the "individualized instruction" approach to education, which uses a systematic processes of diagnostics, individualized instruction, follow-up, and evaluation. This approach is based on a school of thought developed in the United States (see, e.g., Bishop 1971; Gibbons 1971; Glazer 1977). More recent applications of this approach are described in Corbo and Snow (1986), Fullan and Miles (1992), and Fullan (1993). Adaptation and development for Israeli schools was carried out by a group of researchers at the Center for Educational Technology.

Beginning in January 1995, math and Hebrew teachers in program schools also received a course aimed at refreshing and deepening knowledge of subject matter. This training also covered a number of other teaching issues, for example, identifying and working with nonreaders and failing students.

Other Elements of the Intervention

In addition to in-house training, the project established a learning center to assist failing students in the treatment area. Targeted students from all treated schools were sent to the center for additional instruction after school. The project also funded a center for students with learning disabilities. A learning disability diagnostic expert and two teachers staffed the center. Only about 50 elementary school students received additional instruction in the centers, which also served middle schools and high schools in the treatment area. Finally, the project funded special activities with immigrant pupils intended to enhance learning and social integration and to get parents more involved with school and community activities.
Teacher Training and Pupil Learning

References


Carroll, J. G. “Effects of Training Programs for University Teaching


Dildy, P. “Improving Student Achievement by Appropriate Teacher In-Service Training: Utilizing Program for Effective Teaching (PET).” *Education* 103, no. 2 (Winter 1982): 132–38.


