Test of Foundations Learning System in Austin Texas: Summary of Results

Bob McMurray

6/8/2020

DRAFT

Sample

The initial sample consisted of 461 students across five middle schools. The intended design was to assign some children to a control group who received no intervention (business as usual), and others to the intervention group using the Foundations Learning System. Both groups would be tested, using the Foundations Diagnostic, at three timepoints: pre-test, mid-way through the intervention, and post-test. At each test point the would be tested on both *decoding* and *automatic word recognition*. Due to Covid19, the study was halted early. As a result, 404 completed the pre-Diagnostic, 99 completed the mid-Diagnostic and only three students completed the post-Diagnostic.

Results

The analysis of these results addressed two questions. First, we examined the pretest scores descriptively. While this was not a random sample of children, it was intended to assess face-validity of the Foundations Diagnostic by asking if scores varied across grade, IEP and dyslexic status and so forth. Second, we addressed change from pre- to mid-test in the experimental group relative to the control group to evaluate change in performance over the course of the study.

Pretest

Table 1 shows demographic characteristics of the 386 children in the pretest. We conducted a series of paired t-tests and ANOVAs to investigate the influence of these factors on pre-test scores for both decoding and automatic word recognition. Results are shown in Table 2.

For ***grade***, we found no significant differences in pre-test decoding, though there was a downward trend for older grades. For automatic word recognition this trend was even greater and was significant (p=.018) – older children tended to perform worse. The fact that scores did not increase with grade was expected – word level skills as measured here are not commonly taught at these grades, and these kids were deemed by the schools to be "stuck". Thus, any differences across grades likely reflect differences in who was chosen to receive the intervention. Similarly, there was no significant differences between ***genders.***

However, we found effects in the predicted direction for several other variables. In general children classified as ***English Language Learners (ELL)*** scored significantly poorer on both measures (p<.0001). Children classified as having ***dyslexia*** also performed significantly worse on both measures (decoding: p=.042; automatic word recognition: p=.001). However, children receiving an IEP or 504 (pooled) did not score significantly worse than children who did not have an IEP.

Table 1: Results of statistical tests on pre-test scores. * = p<.05

			Decoding			**Automaticity**		
Factor	**Level**	**N**	**Mean (SD)**	**Test**	**Sig**	**Mean (SD)**	**Test**	**Sig**
Grade	6	93	233.3 (155)	F(2,387)= 2.69	0.069	231.3 (121)	F(2,387)= 4.047	0.018
	7	154	215.4 (138)			218.5 (92)		
	8	143	191.2 (132)			195.2 (93)		
Gender	F	184	215.8 (139)	T(384)= 0.75	0.45	213.2 (106)	T(384)= 0.14	0.89
	M	202	205.0 (142)			211.8 (96)		
ELL	No	244	249.9 (135)	T(388)= 7.59	<.0001	243.8 (95)	T(388)= 8.45	<.0001
	Yes	146	145.5 (125)			161.6 (90)		
Dyslexia	No	300	218.7 (143)	T(388)= 2.04	.042	221.3 (105)	T(388)= 2.99	0.001
	Yes	90	184.4 (131)			185.4 (82)		
IEP or 504	No	172	205.9 (137)	T(388)=.64	.52	219.0 (103)	T(388)=1.04	0.298
	Yes	218	206.8 (143)			208.3 (99)		

Table 2: Composition of the control and treatment groups. Shown are counts of students and the proportion of their respective groups.

Factor	Level	Control	Treatment
Gender	F	11 (58%)	43 (54%)
	M	8 (42%)	37 (46%)
Grade	6	9 (47%)	19 (24%)
	7	7 (37%)	37 (46%)
	8	3 (15%)	24 (30%)
ELL	N	16 (84%)	59 (74%)
	Y	3 (16%)	21 (26%)
Dyslexic	N	13 (68.4%)	62 (78%)
	Y	6 (31.5%)	18 (23%)
IEP/504	N	0 (0%)	40 (50%)
	Y	19 (100%)	40 (50%)

Thus, at pre-test, scores in both decoding and automatic word recognition largely reflected expectations in terms of which factors led to higher or lower scores.

Changes in Performance from Pre- to Mid-test

The analyses of change in performance focuses only on change between the pre- and mid-test. There were 99 students in these analyses. Of these 99, control students were only available at one school (Gorzycki, N=19). The remaining 80 students in the treatment group spanned four schools (Gorzycki, Lamar, Lively, Covington). All 19 control students came from a single classroom. The experimental students came from 8 classrooms.

The study was stopped early due to school closures associated with Covid19. Thus, only 52 of the 80 treatment students completed the 12 weeks before the mid-test. Students for whom treatment was stopped early took the mid-test then rather than at the predefined unit in the Foundations Learning System. The mean was 9.58 (SD=3.87, range = 1 to 12). Given this we adopted an 'intent to treat' analytic approach treating all students in the treatment group as if they had completed the treatment. We later report a dosage analysis to determine if differences in exposure were correlated with outcomes.

Table 2 shows a summary of the 99 students that were analyzed here. As can be seen despite the unbalanced sample size, groups were well balanced by grade, gender, ELL status and dyslexia status. Both groups had roughly equal composition with respect to these factors. However, all of the control children had a 504 or IEP, where the experimental students were evenly split.

Figure 1 shows performance on both measures at both points of test as a function of group. As can be seen, at pre-test (red bars), the treatment group outperformed the control group, particularly at decoding. This lack of balance between groups was not unexpected due to the inability to complete control testing. However, the difference between pre- and mid-testing was substantially larger for the treatment groups: for decoding, the control group gained 62 points, while the treatment group gained 101; for automatic word recognition, the control group gained 27 points while the treatment group gained 66. This was

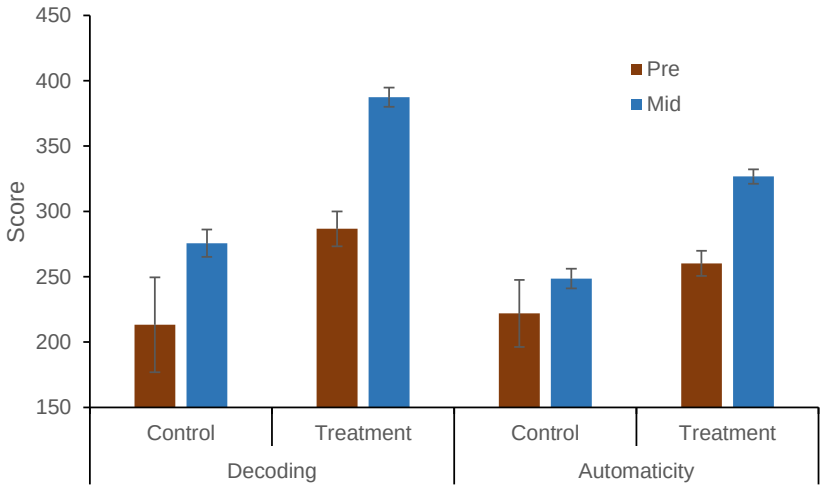

Figure 2: Performance on the two reading measures (decoding and automaticity) as a function of group and time of test (pre- vs. mid-). Error bars represent SEM.

also reflected in the number of children showing improved scores between pre- and mid-testing. In the control group 12 of 19 (63%) made gains in decoding and the same number made gains in automaticity. However, in the treatment group 72 of the 80 children (90%) made gains in decoding and 70 (87.5%) made gains in automaticity.

Following Dugard and Todman (1995) and Van Breukelen (2006) results were analyzed in an ANCOVA which examines differences between groups at the mid-test, controlling for pre-test scores. This has been shown to have more power than a mixed ANOVA, and it can better account for the observed differences between the groups. Pretest scores were centered prior to analysis.

The first ANCOVA analyzed decoding. We found a significant effect of group membership ($F(1,96)=6.18$, $p=.015$) with a small effect size ($\eta_p^2=.06$): the treatment group performed better than the control group, after accounting for pre-test score. We next turned to automaticity. Similar results were found with a significant effect of group ($F(1,96)=10.4$, $p=.002$) with a small effect size ($\eta_p^2=.098$).

One concern is that all of the students in the control group had IEPs or 504s, but only half of the treatment group did. To determine if this could have been driving the treatment effect, we repeated the analyses but excluding all of the children that did not have an IEP. For decoding, we found a significant treatment benefit ($F(1,56)=5.14$, $p=.027$). Similarly, for automatic word recognition we found a significant effect of treatment ($F(1,56)=8.59$, $p=.005$). Thus, the treatment benefit was not driven by the children without IEPs/504s.

Moderators

We next asked the overall group difference was moderated by any other factors. Here, we focused on gender, grade, ELL status, and dyslexia. IEP status was too unbalanced between groups to be considered as a moderator. We investigated these effects by repeating the analyses, adding a single between subjects factor.

We started by considering gender. For decoding, the main effect of group was significant, even accounting for gender ($F(1,94)=4.98$, $p=.028$). While there was no main effect of gender ($F(1,94)=2.12$, $p=.149$), there was significant interaction of gender and group ($F(1,94)=4.91$, $p=.029$). This was driven by the fact that the treatment effect was restricted to girls. This interaction was not observed for automatic word recognition ($F<1$) with effects of treatment condition in both genders.

Next, we examined grade. There was no moderation of the group effect by grade for decoding ($F<1$) or automaticity ($F<1$) suggesting equal benefits of the intervention for each group.

Our third analysis examined ELL status. There was no significant interaction of ELL status and treatment group for decoding ($F(1,94)=2.53$, $p=.115$), though we note that differences were particularly pronounced in the ELL group: while the ELL students in the control group lost 24 points, those in the treatment gained 93 (compared to 79 and 103 respectively among the non ELL students). Automatic word recognition showed a similar pattern. While there was no interaction $F(1,94)=1.26$, $p=.265$), the ELL group lost 9 points in the control group but gained 66 in the treatment group (compared to a gain of 33 and 66 respectively for the non-ELL children). Thus, the fact that there was no significant group x ELL interaction in either analysis, means that there was no statistical evidence to support the claim that ELL children responded differently to the treatment (and numerically they may have shown even more of a benefit).

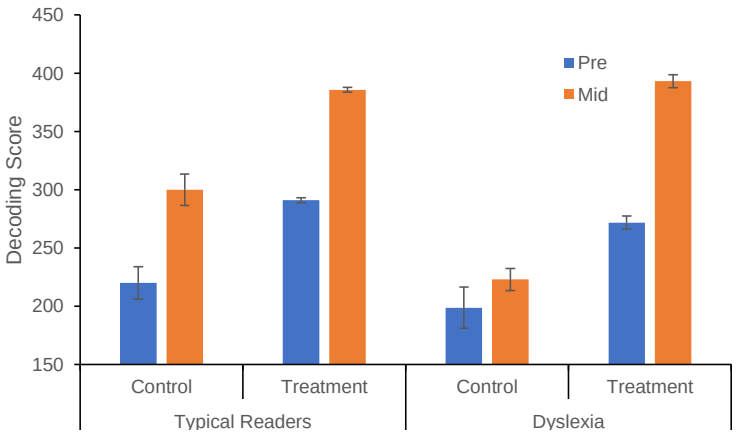

Figure 2: Decoding performance as a function of dyslexia status, treatment group, and time of test (pre- vs. mid-). Error bars represent SEM.

Finally, we examined dyslexia. We found a marginally significant interaction of dyslexia

status and treatment group for decoding (F(1,94)=3.21, p=.076). This is shown in Figure 3 where the change between pre- and mid-test were numerically larger for children with dyslexia than for typical readers. This was not observed for automatic word recognition (F(1,94)=.83, p=.36). Thus, while neither interaction was significant, the numerical results (as well as the marginal interaction for decoding) offer no statistical evidence that children with dyslexia responded differently to the intervention.

Dosage

Given that the intervention was stopped early due to Covid19, subjects varied in the number of weeks of treatment they received. To investigate dosage effects, we considered only the 80 subjects in the treatment group and computed each subjects change from pre- to mid-test. These were then correlated with the number of units received. Both decoding (r=.418, p<.001) and automaticity (r=.443, p<.001) showed moderate and significant correlations.

Conclusions

In the larger sample available for pre-test, scores were low but significantly related to the expected factors, with lower scores in children with dyslexia, ELL status or IEPs, no difference between genders, and mixed results for grade. These results should be interpreted with caution as they may reflect legitimate group differences or differences in the criteria by which children were selected for the intervention (e.g., a child had to be even lower performing than usual to be selected if they were ELL). It is also worth investigating the unexpectedly high relationship between decoding and automaticity measures.

Our analysis of the treatment effect showed statistical evidence for larger gains in both decoding and automatic word recognition in the treatment group. This did not appear to be moderated by ELL status or grade, though decoding may have been moderated by gender and there was marginal moderation for dyslexia status. However, with a small sample (particularly in the control group), this may be statistical noise. Perhaps the more important take home message from the moderation analyses was that no group of children seemed to respond differentially to the treatment. Finally, we also found significant dosage effects for both measures. Given that all of the children got at most half the treatment, this suggests that effects may be larger in future interventions if all students complete the treatment.

These results should be interpreted with caution for several reasons. First, there were far fewer control children that necessary. Second, the control children all came from a single teacher in a single school. As a result, we could not include school and teacher-level factors in the model, which is necessary given that random assignment was done at the classroom level. Third, while the treatment and control groups were reasonably balanced with respect to gender, grade, ELL status and dyslexia, they were not balanced with respect to IEP/504 status. Student can have an IEP or 504s for a variety of reasons (speech, behavioral) that may have little bearing on reading. However, we found that even when excluding the children without IEPs the effects of treatment remained. While we should be cautious in drawing strong conclusions about IEPs, this does not appear to be a factor driving the treatment effect here. Fourth, we note that the measures of performance were all test-internal methods that were developed in conjunction with the treatment – an independent measure is to be preferred. Finally, and perhaps most importantly, none of the children actually finished the intervention, and the presence of a dosage effect suggests that we could be underestimating the treatment effect that would have been observed if they had.

Nevertheless, with these caveats in mind, results show evidence for gains due to the treatment.

References

Dugard, P., & Todman, J. (1995). Analysis of pre-test-post-test control group designs in educational research. *Educational Psychology, 15*(2), 181-198. doi:10.1080/0144341950150207

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*(9), 920-925. doi:10.1016/j.jclinepi.2006.02.007