Preliminary impacts of the Digging Deeper Program

A short report as of 8/2017

Catrina Adams, Joe Taylor, Anne Westbrook

BACKGROUND

PlantingScience: Digging Deeper Together – A Model for Collaborative Teacher/Scientist Professional Development (Digging Deeper) is a partnership among the Botanical Society of America, Biological Science Curriculum Study (BSCS) and the American Society of Plant Biologists (ASPB). The project aims to develop and research the efficacy of a blended professional development (PD) model that engages high school science teachers and scientists in learning experiences that provide opportunities for them to examine their own ideas, understandings, and practices. The overarching goal of the PD model is to increase student engagement and prepare students to proficiently integrate science content with practices of science.

Digging Deeper builds on the previously NSF-funded PlantingScience program, which links high school students in classrooms with scientist mentors via an online community. Digging Deeper adds a PD component for teachers and scientists that includes online community building sessions, week-long summer institutes, an ongoing community of practice during the school year and continued engagement in the community after the school year.

The research component of Digging Deeper examines the effectiveness of the model in transforming teacher practice and increasing student learning. The overarching research questions include:

1) **To what extent does participation in the Digging Deeper community of teachers and scientists affect teacher knowledge and practice?**

2) **To what extent does student use of the online program and participation in the learning community with mentors affect student learning.**

This report shares preliminary results of efficacy research pertaining to the effects of the Digging Deeper/PlantingScience experience on student science achievement and students' attitudes about scientists.

METHOD

Design

This study employed a cluster randomized trial design with biology teachers randomly assigned to treatment and comparison conditions. Random assignment of teachers occurred in 5 waves. However, the effect of wave was fixed in the analysis.

Measures

Student Outcome Measure: Science achievement. Two forms of an achievement test were administered to students pre- and post-intervention, with each form randomly assigned to be taken by approximately 50 percent of the students. Each form (identical pre and post) had 26 multiple choice items covering photosynthesis and cellular respiration. Correct and incorrect responses were coded dichotomously and then converted using the Rasch measurement model measures of individual student ability that lie on an equal interval, 0-100 logarithmic scale (cite – Linacre, 2012). The two forms shared 15 common items and these common items were used to equate the two forms into one equated outcome measure with 37 total items (15 common and 22 unique). The process used is called *common item equating* in a Rasch measurement framework (Linacre, 2012). In a Rasch measurement analysis, indices called *person reliabilities* ranging from 0.00 to 1.00 are generated. These indices can be interpreted in a similar fashion to more traditional reliability measures from classical test theory such as Cronbach's alpha and KR-20 where larger values suggest better instrument performance. For the student science achievement outcome, the model based person reliability for the equated test was 0.70. In the analyses, the pre and post-intervention equated science achievement test is referred to symbolically as PREACH and POSTACH, respectively.

Outcome Measure: Student attitudes toward scientists. Two forms of an attitude scale were administered to students pre- and post-intervention, with each randomly assigned to be taken by approximately 50 percent of the students. Each form (identical pre and post) had 10 Likert scale items covering students' attitudes toward scientists. Rating scale responses were converted using the Rasch model to measures of individual student attitude that also lie on an equal interval, 0-100 logarithmic scale. The two forms shared 5 common items and these common items were used to equate the two forms into one equated attitude measure with 15 total items (5 common and 10 unique). For the student attitudes toward scientists outcome, the model based person reliability for the equated scale was 0.52. In the analyses, the pre and post-intervention equated science achievement test is referred to symbolically as PREATT and POSTATT, respectively.

Demographic and developmental indicators: Students self-reported their inclusion in a set of demographic and developmental groups. These included students' sex (SEX), English language learner status (ELL), free or reduced-price lunch status (FRL) as well as a set of dichotomous race indicators for Asian (ASIAN), African American (AFAM), American Indian or Alaska Native (AMIND), Native Hawaiian or Pacific Islander (HPI), Hispanic or Latina/Latino (HISP), White (WHITE), and Other (OTHER). In these indicators, a code of "1" denoted inclusion in the specified group (for SEX, 1 indicates female). Finally, students also reported their grade level (GRADE) on an ordinal scale (1=9th grade, 2=10th grade, 3=11th grade, and 4=12th grade).

Sample

Baseline sample. The baseline sample for this study included 85 qualifying and consenting teachers (36 treatment, 49 comparison). Nested within those teachers were 2794 consenting students (882 treatment, 1912 comparison).

Analytic sample. The analytic sample of teachers included 64 teachers (27 treatment, 37 comparison). The final analytic sample of students were those with complete data on outcome measures (pre- and post-intervention) as well as all demographic indicators (514 treatment, 1021 comparison).

Attrition rates. Using the baseline and analytic samples described above, we provide in Table X overall and differential attrition rates for the student outcome measures. Because the analytic sample had complete outcome data for both achievement and attitudes, the attrition rates are identical for both outcomes. Note that the number of students assigned is the number assigned to teachers who were *retained* in the study. In this study, 85 teachers were randomly assigned (36 treatment, 49 comparison) and 64 teachers were retained (27 treatment and 37 comparison) The teacher attrition rate was 25.0% in the treatment group and 24.5% in the comparison group. This corresponds to an overall teacher attrition rate of 24.7% and a differential teacher attrition rate of 0.5%.

Table 1. Student Attrition Rates

Attrition Details		
Overall for Students	Number of Students Randomly Assigned	2794
	Number of Students in the Analytic Sample	1535
	Overall Student Attrition	45.1%
Treatment Students	Number of Treatment Students Randomly Assigned	882
	Number of Treatment Students in the Analytic Sample	514
	Treatment Student Attrition	41.7%
Comparison Students	Number of Comparison Students Randomly Assigned	1912
	Number of Comparison Students in the Analytic Sample	1021
	Comparison Student Attrition	46.6%
	Differential Student Attrition	4.9%

Overall sample characteristics. To assist the reader in assessing relevance of the overall analytic sample, we provide in Table X student frequencies in key demographic and developmental categories. Note that students could choose to designate multiple race categories so race percentages do not add to 100.

Table 2. Overall Sample Characteristics (n=1535)

Student Characteristic	Frequency	Proportion of Total Sample
Female	883	58%
English not the native language	290	19%
Receives free or reduced-price lunch	510	33%
Asian	137	9%
African American	116	8%

American Indian or Alaska Native	92	6%
Native Hawaiian or Pacific Islander	28	2%
Hispanic or Latino/Latina	447	29%
White	1018	66%
Other Race	28	2%
Grade 9	653	42%
Grade 10	369	24%
Grade 11	303	20%
Grade 12	210	14%

Baseline and group equivalence. For the analytic sample of students (n=1535) baseline equivalence was tested using the pre-intervention data on student outcomes and group equivalence was tested using frequencies of student characteristics, disaggregated by treatment condition. The metric for equivalence in each case is a standardized mean difference (Cohen's d) effect size and these are reported in Table X. Frequencies for demographic variables were first converted to log odds ratios and, in turn, to Cohen's d using equation 1 from Borenstein et al., (2009):

$$d = \text{log odds ratio} * (\sqrt{3}/\pi) \quad [1]$$

Table 3. Baseline and Group Equivalence (n=1535).

Baseline Outcome OR Student Characteristic	Treatment		Comparison		Effect Size (d)
	Mean or Frequency	SD	Mean or Frequency	SD	
Baseline Achievement	44.46	8.19	45.60	7.89	-0.14
Baseline Attitudes	50.76	6.19	51.75	6.07	-0.16
Female	277	-	606	-	-0.12
English not the native language	61	-	229	-	-0.42
Receives free or reduced-price lunch	153	-	357	-	-0.13
Asian	48	-	89	-	0.04
African American	52	-	64	-	0.29
American Indian or Alaska Native	49	-	43	-	0.48
Native Hawaiian or Pacific Islander	9	-	19	-	-0.03
Hispanic or Latino/Latina	80	-	367	-	-0.61
White	395	-	623	-	0.41
Other Race	12	-	16	-	0.22
Grade 9	234	-	419	-	0.10
Grade 10	151	-	218	-	0.24
Grade 11	87	-	216	-	-0.15
Grade 12	42	-	168	-	-0.44

Analysis Approach

Confirmatory impact analysis. As random assignment occurred at the teacher level and outcomes were measured at the student level, a multilevel analytic approach was selected to get an accurate standard error for the treatment effect. This decision supported by the magnitude of the unconditional intraclass correlation (ICC) estimates (0.15 for attitudes, and 0.29 for achievement). Each multilevel model, one for achievement and one for attitudes, regressed the post-intervention outcome measure (POSTACH or POSTATT) on a vector of grand mean-centered level 1 covariates including baseline outcome values (PREACH or PREATT) and the demographic and developmental indicators. Level 2 independent variables were also grand mean-centered and included the treatment indicator (TREAT) and student level baseline, developmental, and demographic indicators aggregated to the teacher level. For example, the level 2 model for achievement outcomes included a mean for baseline student achievement (PREACH..) and for grade level (GRADE..) as well as a variable indicating the percentage of that teacher's students who had indicated membership in a demographic or developmental category (e.g., PCNTAFAM, PCNTELL).

Covariate selection. The initial approach to choosing covariates for the impact analysis was to use the baseline equivalence information to inform specification of a parsimonious multilevel model. Specifically, the intent was to only use covariates for which imbalance was observed in the baseline variables. Imbalance was defined as being indicated by an effect size whose absolute value was greater than 0.05 standard deviations per *What Works Clearinghouse* guidelines (IES, 2017). Table X indicates that two race variables, HPI and ASIAN, were under this threshold. However, because these variables were part of a larger scheme of race dummy codes, removing them would complicate the interpretation of the race coefficients. As such, we decided to use the full set of available covariates including these two race variables.

Sensitivity analyses for impact estimates. Noting that impact estimates can sometimes be volatile to use of different combinations of covariates, we adopted a reasoned approach to testing whether the magnitude of the impact estimates would be robust to use of different vectors of covariates. Beyond the confirmatory impact model discussed above, we also estimated the treatment effect using three additional models. The first was an impact model identical to the confirmatory model except that the AMIND, HPI, and OTHER race categories were collapsed into a larger race category called OTHER*. This choice was due to our observations of small student frequencies in these three individual race categories and the potential impact on the parameter estimates in the confirmatory model. The second model was identical to the model with the collapsed race categories except that it only included covariates that had relatively small type II error probabilities ($p \leq 0.20$). This approach to improving the precision of impact estimates is suggested in Maldonado and Greenland (1993), Budz-Jorgensen et al (2001), and Price et. al (2007). The final model applied this same criterion ($p \leq 0.20$) to the confirmatory model with all disaggregated race categories.

Effect sizes for impacts. Effect sizes corresponding to the treatment effects estimated using the multilevel models were computed using guidance from the What Works Clearinghouse (WWC) Procedures and Standards Handbook 3.0 (see IES, 2017). The WWC recommendation is to use in the numerator the treatment effect estimate from the multilevel model (i.e, the covariate-adjusted mean difference) and in the denominator the pooled student-level standard deviation.

RESULTS

Confirmatory Analyses

Outcome Descriptive Statistics

Table 4. Descriptive Statistics Post-Intervention Outcomes for the Analytic Sample of Students

	Treatment				Comparison				Effect Sizes	
Outcome	Sample Size	Standard Deviation	Unadjusted Mean	Adjusted Mean[1]	Sample Size	Standard Deviation	Unadjusted Mean	Adjusted Mean[2]	Unadjusted Effect Size	Adjusted Effect Size
Achievement	514	10.030	48.441	48.359	1021	11.448	47.002	45.237	0.131	0.284
Attitudes	514	6.881	50.903	51.312	1021	9.055	50.268	48.960	0.076	0.280

1= Estimating by adding ½ of the covariate-adjusted treatment effect to the grand mean (intercept) from Table X.

2= Estimating by subtracting ½ of the covariate-adjusted treatment effect from the grand mean (intercept) from Table X.

Treatment Effects on Student Achievement

Table 5. Impact on Student Achievement

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	46.797651	0.587753	79.621	50	<0.001
PREACHPM, γ_{01}	0.122069	0.174392	0.700	50	0.487
TREATMEN, γ_{02}	**3.120545**	**1.259707**	**2.477**	**50**	**0.017**
PCNTFEM, γ_{03}	0.673021	4.026502	0.167	50	0.868
MEANGRAD, γ_{04}	1.507069	0.830066	1.816	50	0.075
PCNTASIA, γ_{05}	3.307397	7.357740	0.450	50	0.655
PCNTAFAM, γ_{06}	-7.540127	6.822333	-1.105	50	0.274
PCNTAMIN, γ_{07}	-24.904564	8.815991	-2.825	50	0.007
PCNTHPI, γ_{08}	57.074011	19.989669	2.855	50	0.006
PCNTHISP, γ_{09}	-4.342246	6.028967	-0.720	50	0.475
PCNTWHIT, γ_{010}	-2.960047	7.089081	-0.418	50	0.678
PCNTOTHE, γ_{011}	-25.606459	17.278901	-1.482	50	0.145
PCNTELL, γ_{012}	-3.496976	7.589485	-0.461	50	0.647
PCNTFRL, γ_{013}	5.676595	3.438334	1.651	50	0.105
For GENDER slope, β_1					
INTRCPT2, γ_{10}	-1.010108	0.504282	-2.003	1460	0.045
For GRADE slope, β_2					
INTRCPT2, γ_{20}	0.517054	0.445367	1.161	1460	0.246
For ASIAN slope, β_3					
INTRCPT2, γ_{30}	1.294603	0.945366	1.369	1460	0.171
For AFAM slope, β_4					
INTRCPT2, γ_{40}	-2.157397	1.015022	-2.125	1460	0.034
For AMIND slope, β_5					
INTRCPT2, γ_{50}	-2.424741	1.106724	-2.191	1460	0.029
For HPI slope, β_6					
INTRCPT2, γ_{60}	-2.935400	2.039248	-1.439	1460	0.150
For HISP slope, β_7					
INTRCPT2, γ_{70}	-1.365672	0.773454	-1.766	1460	0.078
For OTHER slope, β_8					
INTRCPT2, γ_{80}	-3.970837	1.881429	-2.111	1460	0.035
For ELLBINAR slope, β_9					
INTRCPT2, γ_{90}	-0.606709	0.702921	-0.863	1460	0.388
For FRL slope, β_{10}					
INTRCPT2, γ_{100}	-1.605417	0.648640	-2.475	1460	0.013
For PREEQACH slope, β_{11}					
INTRCPT2, γ_{110}	0.247034	0.034494	7.162	1460	<0.001

Treatment Effects on Attitudes

Table 6. Impact on Student Attitudes

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. $d.f.$	p-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	50.136360	0.348225	143.977	50	<0.001
TREATMEN, γ_{01}	**2.350514**	**0.738541**	**3.183**	**50**	**0.003**
PCNTFEM, γ_{02}	-2.527903	2.322230	-1.089	50	0.282
MEANGRAD, γ_{03}	-0.264844	0.534658	-0.495	50	0.623
PCNTASIA, γ_{04}	2.690891	4.414424	0.610	50	0.545
PCNTAFAM, γ_{05}	2.608799	4.084732	0.639	50	0.526
PCNTAMIN, γ_{06}	-26.750778	5.262957	-5.083	50	<0.001
PCNTHPI, γ_{07}	24.612670	11.874721	2.073	50	0.043
PCNTHISP, γ_{08}	5.808063	3.633047	1.599	50	0.116
PCNTWHIT, γ_{09}	10.510628	4.191675	2.508	50	0.015
PCNTOTHE, γ_{010}	-2.848900	10.570042	-0.270	50	0.789
PCNTELL, γ_{011}	3.925348	4.724177	0.831	50	0.410
PCNTFRL, γ_{012}	3.711631	2.043331	1.816	50	0.075
PREATTEQ, γ_{013}	0.342679	0.207962	1.648	50	0.106
For GENDER slope, β_1					
INTRCPT2, γ_{10}	0.851984	0.413686	2.059	1460	0.040
For GRADE slope, β_2					
INTRCPT2, γ_{20}	0.319545	0.365091	0.875	1460	0.382
For ASIAN slope, β_3					
INTRCPT2, γ_{30}	0.967845	0.775026	1.249	1460	0.212
For AFAM slope, β_4					
INTRCPT2, γ_{40}	1.772850	0.832251	2.130	1460	0.033
For AMIND slope, β_5					
INTRCPT2, γ_{50}	-1.750680	0.907888	-1.928	1460	0.054
For HPI slope, β_6					
INTRCPT2, γ_{60}	-0.295712	1.673305	-0.177	1460	0.860
For HISP slope, β_7					
INTRCPT2, γ_{70}	-0.154965	0.634036	-0.244	1460	0.807
For OTHER slope, β_8					
INTRCPT2, γ_{80}	-0.513175	1.542946	-0.333	1460	0.739
For ELLBINAR slope, β_9					
INTRCPT2, γ_{90}	0.527391	0.575914	0.916	1460	0.360
For FRL slope, β_{10}					
INTRCPT2, γ_{100}	-0.912619	0.531813	-1.716	1460	0.086
For PREATTME slope, β_{11}					
INTRCPT2, γ_{110}	0.422864	0.032552	12.990	1460	<0.001

Interpretation. Controlling for the effects of student and teacher-level covariates, the Digging Deeper program demonstrates a statistically significant impact on both student achievement and attitudes about scientists.

Exploratory Analyses

Sensitivity Analysis for Achievement and Attitude Impacts. Impact estimates from sensitivity analyses are reported in Tables X and X. Effect sizes are Hedges' g with associated 95% confidence intervals. The full regression output from these analyses is available by request.

Table 7. Sensitivity Analyses for Impacts on Achievement

Model	Impact Estimate	SE	t	p	Effect Size (g)
Confirmatory Model (as above)	3.121	1.260	2.477	**0.017**	0.284 [0.177, 0.390]
Confirmatory model with only predictors where p≤0.20.	3.663	0.573	6.391	**<0.001**	0.333 [0.226, 0.440}
Confirmatory model with collapsed race categories.	3.092	0.596	5.192	**<0.001**	0.281 [0.175, 0.288]
Confirmatory model with collapsed race categories and only predictors where p≤0.20.	3.199	0.586	5.458	**<0.001**	0.291 [0.184, 0.397]

Table 8. Sensitivity Analyses for Impacts on Attitudes

Model	Impact Estimate	SE	t	p	Effect Size (g)
Confirmatory Model (as above)	2.351	0.739	3.183	**0.003**	0.280 [0.174, 0.387]
Confirmatory model with only predictors where p≤0.20.	2.402	0.711	3.379	**0.001**	0.286 [0.180, 0.393]
Confirmatory model with collapsed race categories.	1.853	0.838	2.210	**0.031**	0.221 [0.114, 0.327]
Confirmatory model with collapsed race categories and only predictors where p≤0.20.	1.714	0.816	2.100	**0.040**	0.204 [0.098, 0.310]

Interpretation: The statistically significant impacts of Digging Deeper on both achievement and attitudes are robust to variation in how the impact model is specified