Replication and Extension of 'Disrupting Education'

Brendan Schuetze[1]

[1] The University of Texas at Austin

Author Note

The data used for the analyses in this paper is located at http://doi.org/10.3886/E113192V1.

Correspondence concerning this article should be addressed to Brendan Schuetze, 1912 Speedway, Suite 506E, Austin, TX 78712. E-mail: brendan.schuetze@utexas.edu

Abstract

In this paper, I replicate the results of Muralidharan, Singh, and Ganimian (2019), which evaluate the results of a voucher program implemented in India. This program randomly assigned students to receive free technology-enhanced after-school instruction in Hindi or mathematics. Due to missing data, Muralidharan et al. excluded approximately one-sixth of their original sample when calculating their intent-to-treat analyses. To account for the uncertainty associated with this missing data, I extend the initial intent-to-treat analyses performed by Muralidharan et al. using multiple imputation and Manski bounds, ultimately finding similar results independent of the analytic choices employed.

Keywords: Intent-to-Treat, Treatment Effect Heterogeneity, Multiple Imputation, Manski Bounds

Word count: 4226

Replication and Extension of 'Disrupting Education'

In the twentieth and twenty-first centuries, school enrollment has greatly increased throughout India. As Muralidharan et al. (2019) note, 95 percent of school-aged Indian children are enrolled in school, but progress in math and reading has not been as strong as one might expect from such high enrollment numbers. Half of India's fifth graders cannot read at a second grade level.

One hypothesis for the lack of success seen by school systems in India and potentially other developing countries is a mismatch between prior knowledge and the level of instruction offered at each grade level. In essence, students in developing countries come into school with highly inequitable levels of exposure to reading and mathematics at home. This inequity is further compounded by several factors. Spotty attendance rates can result in inconsistent gaps in knowledge across students who share the same classroom and should have ostensibly gained the same knowledge. Furthermore, unlike students in developed countries with a longer history of public education, parents in developing countries are unlikely to be able to tutor or assist their children in learning material that they are falling behind on.

In essence, many of the subtle stabilizing factors built-in to education systems in the Global North are not available to children in developing countries. Thus, when these children do not adequately learn material, they are less likely to be able to get back on grade level. This "falling off the treadmill" effect results in the aforementioned paradoxical effect, where school enrollments can be high, yet learning can be much lower than one might expect. Data from the present paper being replicated supports this notion. Muralidharan et al. (2019) found, yes, the mean level of educational progress in these schools is lower than anticipated. But perhaps more importantly, there is substantial heterogeneity in student outcomes. Students drawn from the present sample *enrolled in the same grade* exhibit performance levels in math and reading "spanning five to six grade levels" (p. 1428).

Earlier field experiments by Banerjee and colleagues throughout India have shown somewhat promising results from what has been called the *teach at the right level* approach. This program starts from the fact that individual student prior knowledge is highly variable. Given research from educational researchers on what is known as the "zone of proximal development" (Kalyuga, 2009; Wass & Golding, 2014), showing that novice and experts in the same domain benefit from different learning materials tailored to their skill level, the variable levels of student knowledge necessitate an individualized approach to learning.

Earlier research on the teach at the right level approach has shown positive results. For example, Banerjee et al. (2016) conducted a pair of experiments implementing extra individualized tutoring both inside and outside of school hours. In one experiment, teachers were given dedicated time during the school day to engage in individualized tutoring. In the other, trained staff from outside of the school system formed the tutoring corps. These two experiments resulted in language acquisition gains of 0.15 and 0.70 σ, respectively, over control schools. Banerjee, Banerji, Duflo, Glennerster, and Khemani (2010) found similar results with an intervention using tutors drawn from a pool of community volunteers.

Although the teach at the right Level program had seen promising results in Indian schools, Muralidharan et al. (2019) found that many of the barriers in the previous implementations of these approaches resulted from the amounts of personnel and funding needed to staff these supplementary programs effectively. Muralidharan et al. sought to rectify these problems and improve upon prior interventions by using a personalized technology platform called *Mindspark*. In essence, Muralidharan et al. replaced the human tutors of previous interventions with intelligent tutoring technology. The Mindspark intelligent tutors used adaptive learning modules to track individual students' performance and ultimately teach to each students' current level of prior knowledge. This intelligent tutoring platform was then implemented in out-of-school tutoring centers (*Mindspark Centers*) around low-income neighborhoods in Delhi. Students from these neighborhoods

chosen for the experimental group were given a voucher to attend these centers for free (they otherwise cost approximately $3 USD per month).

Though the Mindspark technology platform was a key component of the intervention, they were not the only supplemental instruction afforded to the students. Students were scheduled to attend the Mindspark Centers six sessions a week, with each session lasting 90 minutes. Half of this time (45 minutes) was devoted to the intelligent tutor, while the other half was devoted to in-person group-based instruction. These instructional groups were generally composed of around fifteen students per group.

Thus, the research question of Muralidharan et al. was: to what extent can the more cost-effective Mindspark tutoring programs effectively replicate the success of previous human-implemented teach at the right level programs by increasing the math and Hindi abilities of the students in the treatment group? Specifically, since I will be conducting intent-to-treat analyses, I will not be able to speak to the true causal effects of the tutoring centers in and of themselves Rather, I replicate and extend Muralidharan et al.'s evaluation of the causal effect of awarding Indian students a voucher to receive free computer- and human-aided after-school tutoring on these abilities. As one might expect, Muralidharan et al. also used an instrumental variable approach in their paper to estimate the causal effect of actually engaging in tutoring, but I will not be extending these findings in this paper.

Replication

Sample

The study was conducted between September 2015 and February 2016. The sample consisted of 619 middle-schoolers (4^{th} through 9^{th} grades) drawn from five public schools located nearby Delhi Mindspark centers. Most of these students (97.5%) were in grades six through nine. In terms of representativeness of the sample, students in the control and treatment groups scored 0.15 σ higher on pre-program tests than their non-participating

peers. Despite this small overall mean difference, the supports of the control and treatment group test distributions still entirely overlapped, meaning all participants could be relatively well matched. In terms of the differences between the experimental and control groups, the only variable with a standardized mean difference greater than 0.15 was age. The only variable between 0.10 and 0.15 standardized mean difference was grade in school. See *Table 1* for descriptive statistics regarding other sample characteristics.

I note that although Muralidharan et al. tend to cast their results in terms of developing countries as a unified whole, this study and other studies like it have been performed in India, and not other developing countries. This prevents us from generalizing strongly to the rest of the developing world. The nuances of development trajectories and education systems across the globe most likely create unique educational climates, which are likely to moderate the effects of interventions described in this paper. With this in mind, I interpret the results of Muralidharan et al. in the context of India, not developing countries broadly.

Manipulation

In order to assign treatment, Muralidharan et al. randomized students to receiving the teach at the right level program through a program voucher lottery. This was a block-randomized experiment, with students randomized within neighborhoods. These neighborhoods were split up and defined by proximity to each Mindspark center.

Because the Mindspark tutors were used entirely during after-school hours, attendance could not be compulsory. This voucher lottery also served a double purpose of incentivizing the students to use the Mindspark centers, as these vouchers allowed students to use tutoring services that would otherwise cost a small fee (approximately $3 USD per month) to attend. The vouchers enabled six-months of free access to the Mindspark centers. Students on the waitlist were afforded six-months of free access to the centers after the study period ended

contingent on the successful completion of the pre- and post-tests.

Outcome Measures

All participants in both the control and treatment groups took pen-and-paper final tests in Hindi and math. The items in both of these tests ranged from very easy (e.g., math: simple arithmetic; Hindi: matching pictures to words) to difficult (math: interpretation of graphs; Hindi: drawing complex inferences from texts), as to be able to capture the wide range of student ability. Test items were drawn from a variety of pre-validated standardized tests aimed at assessing the math (e.g.PISA, TIMMS; Neidorf, Binkley, Gattis, & Nohara, 2006) and reading (e.g., PISA, PIRLS; Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009) knowledge of fourth through eighth graders.

Theta parameters estimated via a maximum-likelihood using an item-response theory psychometric model. This three-parameter logistic model was trained on data from this final test of Hindi and mathematics knowledge formed the outcome of interest. Theta parameters represent the estimated skill or knowledge level for math and Hindi for each participant in this study. Using theta parameters rather than raw test scores allows for comparison of scores across tests using different items, which is a key requirement of the teach at the right level program. Additionally, the theta parameters used in the following analyses were centered at zero with a standard deviation of one in the baseline group. This simplifies the calculation of the marginal effects of treatment.

Covariates

Two covariates were employed in Muralidharan et al.'s analyses, pre-test scores and strata. Pre-test scores were estimated using the same method described above in the *Outcome Measures* section, for math and Hindi separately, only instead of using the data from the post-test, data from the pre-test was used to train the three-parameter logistic model. Pre-test scores were included in all of the models employed by Muralidharan et al.

(2019) for the purpose of increasing power to detect effects of treatment.

Strata was the other covariate employed in a subset of their analyses. Strata simply refers to the group within each student was randomized and corresponds to the Mindspark center they were assigned to. Because students attended their nearest Mindspark center, this variable also captures effects of neighborhood heterogeneity.

Data Analysis Procedures

Because treatment was randomly assigned, student-to-student, the expected systematic differences between those who received the voucher and those who did not is zero in terms of potential confounding covariates. Although the addition of control variables potentially allows for greater power in detecting causal effects, the randomization of treatment allows for the identification of causal effects without necessarily needing to control for additional covariates through regression models.

As such, the main intent-to-treat analyses employed in this paper took the form of two relatively simple regression models. In the first pair of models, Muralidharan et al. predicted the final test outcomes ($\theta_{post-test}$) using the treatment indicator (0 = not awarded lottery; 1 = awarded lottery) and pre-test scores ($\theta_{pre-test}$). Throughout my analyses, I use the *R* (R Core Team, 2020) regression package *estimatr* (Blair, Cooper, Coppock, Humphreys, & Sonnet, 2020) to obtain heteroskedascity consistent standard errors. When necessary, I use the *wec* package for weighted effects coding (Te Grotenhuis et al., 2016).

ITT Analysis Strata not Included

```
# Math Model
m_simple <- lm_robust(data = df_ITT_math, se_type = "HC1",
            m_theta_mle2_scaled ~ treat + m_theta_mle1_scaled)

# Hindi Model
h_simple <- lm_robust(data = df_ITT_hindi, se_type = "HC1",
            h_theta_mle2_scaled ~ treat + h_theta_mle1_scaled)
```

As reflected in the models above, Muralidharan et al. chose heteroskedasticity consistent standard errors, specifically those suggested by Huber and White known as HC_1 (see MacKinnon & White, 1985). Heteroskedasticity consistent standard errors allow for violations of the assumption of homoskedasticity, without negatively impacting Type 1 error rates. The authors note that one might make the argument that cluster-robust standard errors should be used given the clustered nature of students within strata. The authors make the argument that $HC1$ standard errors are more appropriate here, given that randomization was employed at the invidual – as opposed to group – level.

ITT Analysis, Strata as Fixed Effects

```
# Strata Intercepts
m_strata <- lm_robust(data = df_ITT_math, se_type = "HC1",
            m_theta_mle2_scaled ~ treat + m_theta_mle1_scaled,
            fixed_effects = ~ strata_c)

# Strata Intercepts
h_strata <- lm_robust(data = df_ITT_hindi, se_type = "HC1",
            h_theta_mle2_scaled ~ treat + h_theta_mle1_scaled,
```

```
            fixed_effects = ~ strata_c)
```

Interpretation

The results from these models (see appendix) all reveal significant effects of the lottery intervention. Because the results are nearly identical between the three models, I will summarize all of them briefly. The math intervention tends to result in a 0.37 σ increase in outcome scores, while 4the Hindi intervention results in a smaller 0.24 σ increase in outcome scores. For every σ increase in the math pre-test score, we predict 0.58 σ increase in the post-test score. For the Hindi pre-test, we see that every σ increase predicts a 0.68 σ increase in the post-test. Because these numbers are all on a common scale, we can compare across them and see that the math scores seem a little more malleable than the Hindi scores – the math intervention changes outcomes more than the Hindi intervention and the Math pre-test is less predictive than the test covering Hindi ability.

Assumptions

The assumptions of this analysis are rather minimal, due to the randomization of treatment assignment via the lottery system described above. The two largest assumptions remaining uninterrogated by the authors are that: (A) missing data is not systematically altering the result of the dataset; and (B) there is no treatment effect heterogeneity from strata-to-strata. By failing to account for missing data, the authors potentially leave themselves open to failing to control for confounds that affect whether data is missing and the effects of treatment. By assuming no treatment effect heterogeneity, they potentially fail to account for the variability in program effectiveness from site-to-site and neighborhood-to-neighborhood. Through my extension, the aim of this project is to assess whether the treatment effect estimates replicated above change substantially upon the application of techniques to deal with missingness and treatment effect heterogeneity.

Differences Between Original and Replication

Given the relatively simple form of these models and the exact specification thereof contained in Muralidharan et al. (2019), I was fairly easily able to replicate all four ITT analyses contained in the original article [School Subject (2) x With or Without Strata as Fixed Effects (2)] in terms of the exact betas of both treatment and pre-test. Muralidharan et al. (2019) did not report the specific strata fixed effects, so I could not compare these numbers. I was not, however, able to exactly replicate the intercepts reported in their study. I believe that this discrepancy stems from the standardization (Z-scoring) process the authors applied to the pre- and post- test scores. They do not exactly specify what sample formed the basis of their standardization. I tried scaling the test scores based upon several different reference distributions. None of them resulted in exact replications of the numbers reported in the manuscript. This may also be a difference between Stata and R in that maybe the standardization choices they made may be more obvious to me if I were more familiar with the Stata default options. Overall, I am not particularly troubled by this discrepancy, as the intercept is not particularly meaningful in this situation, and the treatment effect is the clear quantity of interest.

Extension

Inclusion of Treatment Effect Heterogeneity across Strata

In the models below, I account for potential treatment effect by strata heterogeneity by including the interaction term between the *treatment* and *strata* variables. Strata uses weighted effects coding, such that the main effect of treatment captured by the regression model is equal to the overall weighted effect of treatment across strata.

```
# Math Model
m_strata_by_treat <- lm_robust(data = df_ITT_math, se_type = "HC1",
            m_theta_mle2_scaled ~ treat * strata_wec + m_theta_mle1_scaled)


# Hindi Model
h_strata_by_treat <- lm_robust(data = df_ITT_hindi, se_type = "HC1",
            h_theta_mle2_scaled ~ treat * strata_wec + h_theta_mle1_scaled)
```

The results of this analysis are almost exactly the same as the earlier models not including the interaction terms for math ($\beta_{1,\text{ interaction}} = 0.377$, versus $\beta_{1,\text{ main effects only}} = 0.373$) and Hindi ($\beta_{1,\text{ interaction}} = 0.239$, versus $\beta_{1,\text{ main effects only}} = 0.237$)

Manski Bounds

For a preliminary check of the robustness of the models evaluated throughout this paper, I subject the Muralidharan et al. (2019) to the Manski extreme bounds approach suggested by Horowitz and Manski (2000). This approach only assumes the support distribution of the missing variable. In essence, Manski bounds evaluate the treatment effect in the absolute worst and best case scenarios. These scenarios are generated by replacing the missing data dependent on the treatment assignment, such that missing data is either replaced with the minimum or maximum observed value. In the best case scenario, the treatment group receives the maximum value on the outcome when missing data is found and the control group receives the lowest value. The opposite occurs for the worst case scenario. Because there is no absolute theoretical max or minimum for the Z-scores obtained from the psychometric tests used in the analyses of Muralidharan et al. (2019), I will be using the empirical maximums and minimums in each strata for the purpose of this analysis. As one might expect, the extreme maximum and minimum values for the entire unstratified sample are approximately equal to $+/-3\sigma$, respectively.

```r
df_eb <- df


df_eb$extreme_bound_lower <- df_eb$m_theta_mle2

df_eb$extreme_bound_upper <- df_eb$m_theta_mle2


# Calculate min and max by strata

eb_stats <- df_eb %>%

  group_by(strata) %>%

    summarise(min = min(m_theta_mle2, na.rm = TRUE),

              max = max(m_theta_mle2, na.rm = TRUE))


# Perform replacement in accordance with Manski bounds

for(i in 1:nrow(df_eb)) {

  if(is.na(df_eb$extreme_bound_lower[i])) {

    if(df_eb$treat[i] == 0) {

      df_eb$extreme_bound_lower[i] <-

        eb_stats$max[eb_stats$strata == df_eb$strata[i]]

      df_eb$extreme_bound_upper[i] <-

        eb_stats$min[eb_stats$strata == df_eb$strata[i]]

    } else {

      df_eb$extreme_bound_lower[i] <-

        eb_stats$min[eb_stats$strata == df_eb$strata[i]]

      df_eb$extreme_bound_upper[i] <-

        eb_stats$max[eb_stats$strata == df_eb$strata[i]]

    }

  }

}
```

```
# Calculate Regression Estimates
ebl <- lm_robust(data = df_eb, se_type = "HC1",
                 formula = extreme_bound_lower ~ treat + m_theta_mle1)
ebu <- lm_robust(data = df_eb, se_type = "HC1",
                 formula = extreme_bound_upper ~ treat + m_theta_mle1)
```

The extreme bounds calculated for the effect of treatment were [-0.25, 0.85]. This analysis indicates that in the most extreme case, there is a remote-but-not-entirely-impossible chance that missing could impact the treatment effect estimates such that they are no longer positive. With this possibility in mind, I use multiple imputation to perform a more realistic evaluation of the impacts of missing data on our treatment effect estimate.

ITT with Multiple Imputation

Multiple imputation was chosen as the first step of this extension, because approximately one-sixth of the data from this experiment was removed by Muralidharan et al. (2019) during their analysis phase. Using multiple imputation has several benefits over the choice to simply remove the data from students with missing observations entirely. These benefits include unbiased estimates of treatment effects if assumptions are met and higher statistical power, given the preservation of sample size that results (McCleary, 2002). Though there are benefits of multiple imputation, these benefits do come at the cost of assuming that the data is "missing at random" (or the more lenient assumption, "missing completely at random"). That is, we are assuming, after accounting for the covariates present in the dataset, there are no further systematic confounds between missing data and the outcomes (Sterne et al., 2009).

First, I assessed the completeness of the dataset by calculating how many incomplete cases existed therein.

```
# Count number of missing rows
# Remove Student Tuition Columns because NA does not mean
# missing, but rather not applicable for those columns.
incomplete_cases <- df %>% select(!starts_with("st_tui")) %>%
  filter(!complete.cases(.)) %>% nrow()


incomplete_test_data <- df %>%
  select(m_theta_mle1, m_theta_mle2,
         h_theta_mle1, h_theta_mle2) %>%
    filter(!complete.cases(.)) %>% nrow()
```

The number of cases with at least one missing value was 242, which is 39.10 percent of the dataset. Furthermore 13.89 percent of cases were missing pre- or post-test data, which is integral to the ITT analyses calculated by Muralidharan et al. (2019).

Then, I used the *R* package, *mice* for multiple imputation (van Buuren & Groothuis-Oudshoorn, 2011), which I set to impute data using the "predictive mean matching" algorithm. The multi-level structure was accounted for using a fixed-effects approach, by entered strata as a predictor into the imputation algorithm. A rule of thumb suggested by White, Royston, and Wood (2011) for assessing the number of necessary imputations data sets for an analysis is to set the number of imputations equal to the percentage of missing cases. Because the pre- and post-test data are the only variables included in our regression models using the imputed data, I only included cases missing this data in determining the minimum number of imputations I would use. Correspondingly, the minimum number of imputations recommended by White et al. was equal to the percentage of cases missing either pre- or post-test data (14). Due to practical constraints stemming from the mice library and my computer setup (14 is not an even multiple of the number of cores I am using for imputation), I rounded this number up to the nearest multiple of five

and created fifteen imputed copies of the dataset.

```
# Multiple imputation on multiple cores using predictive mean matching.
# Runs quickly given small dataset and not much missing data
imputed <- parlmice(data = df, n.core = 5, n.imp.core = 3, method = "pmm")
df_imputed <- complete(imputed, action = "long")


# Create Math Data
df_imp_scaled_m <- df_imputed %>% scale_scores("m", TRUE)


# This converts a long dataset to a list of individual datasets.
dat_list_m <- long2list(df_imp_scaled_m)


# Convert data back to MIDS format
df_imp_c_m <- miceadds::datalist2mids(dat_list_m)


# Do the same thing for Hindi data
df_imp_scaled_h <- df_imputed %>% scale_scores("h", TRUE)
dat_list_h <- long2list(df_imp_scaled_h)
df_imp_c_h <- miceadds::datalist2mids(dat_list_h)
```

Below is the code for the models using the pooled imputed datasets. Except for the pooling across the multiply imputed datasets, these regression models are of the same form as those used in the earlier intent-to-treat analyses.

```
# Without strata as fixed effects
m_IMP_simple <- with(
    data = df_imp_c_m, exp = lm_robust(se_type = "HC1",
    formula = m_theta_mle2_scaled ~ treat + m_theta_mle1_scaled)
  ) %>% pool()

h_IMP_simple <- with(
    data = df_imp_c_h, exp = lm_robust(se_type = "HC1",
    formula = h_theta_mle2_scaled ~ treat + h_theta_mle1_scaled)
  ) %>% pool()

# With strata as fixed effects
m_IMP_strata <- with(
    data = df_imp_c_m, exp = lm_robust(se_type = "HC1",
    formula = m_theta_mle2_scaled ~ treat + m_theta_mle1_scaled + strata_c)
  ) %>% pool()

h_IMP_strata <- with(
    data = df_imp_c_h, exp = lm_robust(se_type = "HC1",
    formula = h_theta_mle2_scaled ~ treat + h_theta_mle1_scaled + strata_c)
  ) %>% pool()
```

Interpretation of Imputed Analyses

As shown in tables 8 through 14, The results of these analyses are nearly identical to the intent-to-treat analyses conducted on the non-imputed datasets. Although the estimates are largely (and often exactly the same after rounding), we have gained more complete degrees of freedom. Though this is only true before adjustment for non-response by the mice package; after adjustment using formula proposed by Barnard and Rubin (1999) the degrees of freedom are in fact smaller. Though the estimates are the same as the non-imputed data, arguably I am now making better justified inferences, due to the fact that the uncertainty induced by missingness in the data is accounted for by the analysis.

```
# Math, With Strata-by-Treatment Interaction
reg_list_m <- vector(mode = "list", length = length(dat_list_m))
for(i in 1:length(dat_list_m)) {
  reg_list_m[[i]] <- lm_robust(data = dat_list_m[[i]], se_type = "HC1",
                         formula = m_theta_mle2_scaled ~ treat * strata_wec +
                           m_theta_mle1_scaled)
}


m_IMP_strata_by_treat <- pool(reg_list_m)


# Math, With Strata-by-Treatment Interaction
reg_list_h <- vector(mode = "list", length = length(dat_list_h))
for(i in 1:length(dat_list_m)) {
  reg_list_h[[i]] <- lm_robust(data = dat_list_h[[i]], se_type = "HC1",
                         formula = h_theta_mle2_scaled ~ treat * strata_wec +
                           h_theta_mle1_scaled)
}
```

```
h_IMP_strata_by_treat <- pool(reg_list_h)
```

Discussion

Throughout this paper, the treatment effect estimates of the Mindspark intervention have been remarkably stable. This is most likely due to the fact that I am analyzing a well-randomized experiment, which we would expect to result in stable treatment effect estimates, due to the expected differences in confounding factors between treatment and control groups to be zero. The only approach that showed any meaningful differences from the estimates reported by Muralidharan et al. (2019) was the extreme bounds approach. But, as the name suggests, this approach is not realistic. The overall effects of assignment to the Mindspark treatment group are positive and equal to approximately 0.37 standard deviations for math outcomes and 0.23 standard deviations for Hindi outcomes. As noted by Kraft (2018), many effect sizes in education research are quite small, and considering this ITT analysis most likely underestimates the effect of attending the tutoring sessions, this result seems promising for the prospects of implementing after-school tutoring programs throughout other areas of India. More research should be done in other developing countries to assess the generalizability of this type of intervention to other educational contexts, where the problems affecting education may be substantially different.

This study had many strengths, particularly the use of randomization in assignment to treatment and the relatively sophisticated psychometric models used to evaluate the students' test scores. Nevertheless, the most important limitation of the analyses presented above are that they are intent-to-treat analyses, meaning I am not identifying the causal impact of attending the program in-and-of itself, but rather the causal effect of assigning someone a lottery to attend these centers. Not all children who received assignment to the treatment condition attended the Mindspark centers. Furthermore, even within those who did attend the Mindspark centers, there was variability in terms of how many days of

tutoring they attended. Ultimately, the effects of actually *attending* Mindspark are also of interest to policy makers. Muralidharan et al. (2019) did perform an instrumental variable (IV) analysis to get at the answer of this question, and I if I were to extend this paper further, I would want to perform IV analysis on my multiply imputed dataset. There is also the question of to what extent the human-aided versus computer-aided portions of this tutoring program impacted the children's outcomes. Future research might evaluate whether computer tutoring programs could be implemented without need for human tutor oversight, as such a program could be more cost effective and potentially reach more students.

References

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., ... Walton, M. (2016). *Mainstreaming an effective intervention: Evidence from randomized evaluations of "teaching at the right level" in india*. National Bureau of Economic Research.

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in india. *American Economic Journal: Economic Policy*, *2*(1), 1–30.

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*(4), 948–955. Retrieved from http://www.jstor.org/stable/2673599

Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2020). *Estimatr: Fast estimators for design-based inference*. Retrieved from https://CRAN.R-project.org/package=estimatr

Horowitz, J. L., & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, *95*(449), 77–84.

Kalyuga, S. (2009). The expertise reversal effect. In *Managing cognitive load in adaptive multimedia learning* (pp. 58–80). IGI Global.

Kraft, M. A. (2018). Interpreting effect sizes of education interventions. *Brown University Working Paper*.

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*,

29(3), 305–325.

McCleary, L. (2002). Using multiple imputation for analysis of incomplete data in clinical research. *Nursing Research*, *51*(5), 339–343.

Mullis, I. V., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework.* ERIC.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in india. *American Economic Review*, *109*(4), 1426–1460.

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). Comparing mathematics content in the national assessment of educational progress (naep), trends in international mathematics and science study (timss), and program for international student assessment (pisa) 2003 assessments. Technical report. NCES 2006-029. *National Center for Education Statistics*.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *Bmj*, *338*, b2393.

Te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2016). When size matters: Advantages of weighted effect coding in observational studies. *International Journal of Public Health*, 1–5. Retrieved from http://doi.org/10.1007/s00038-016-0901-1

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from https://www.jstatsoft.org/v45/i03/

Wass, R., & Golding, C. (2014). Sharpening a tool for teaching: The zone of proximal development. *Teaching in Higher Education*, *19*(6), 671–684. https://doi.org/10.1080/13562517.2014.901958

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399.

Table 1

	0	1	SMD
n	305	314	
Math Pre-Test (mean (SD))	0.01 (1.02)	-0.01 (0.98)	0.016
Hindi Pre-Test (mean (SD))	-0.05 (1.04)	0.05 (0.96)	0.096
SES (mean (SD))	0.04 (1.69)	-0.03 (1.72)	0.041
Age (mean (SD))	12.41 (1.50)	12.67 (1.56)	0.174
Grade (%)			0.118
4	3 (1.0)	2 (0.7)	
5	6 (2.0)	4 (1.3)	
6	90 (30.1)	81 (26.6)	
7	77 (25.8)	80 (26.2)	
8	85 (28.4)	92 (30.2)	
9	38 (12.7)	46 (15.1)	
Gender (Female) = 1 (%)	231 (75.7)	239 (76.1)	0.009

Table 2

Intent-to-treat analysis without strata on math outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.37	0.06	5.80	<.001	0.24	0.49	532.00
Math Pre Test	0.58	0.04	13.93	<.001	0.50	0.67	532.00

Table 3

Intent-to-treat analysis with strata as fixed-effects on math outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.37	0.06	6.03	<.001	0.25	0.50	514.00
Math Pre Test	0.57	0.04	14.51	<.001	0.49	0.64	514.00

Table 4

Intent-to-treat analysis with treatment effect heterogeneity on math outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.38	0.06	6.14	<.001	0.26	0.50	496.00
Math Pre Test	0.58	0.04	14.94	<.001	0.50	0.66	496.00

Table 5

Intent-to-treat analysis without strata on Hindi outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.23	0.06	3.66	<.001	0.11	0.35	534.00
Hindi Pre Test	0.71	0.04	17.88	<.001	0.63	0.79	534.00

Table 6

Intent-to-treat analysis with strata as fixed-effects on Hindi outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.24	0.06	3.91	<.001	0.12	0.36	516.00
Hindi Pre Test	0.68	0.04	18.41	<.001	0.61	0.76	516.00

Table 7

Intent-to-treat analysis with treatment effect heterogeneity on Hindi outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.24	0.06	3.94	<.001	0.12	0.36	498.00
Hindi Pre Test	0.68	0.04	18.46	<.001	0.61	0.75	498.00

Table 8

Imputed ITT analysis without strata on math outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.31	0.06	4.90	<.001	0.19	0.44	309.12
Math Pre Test	0.53	0.04	13.19	<.001	0.45	0.61	320.02

Table 9

Imputed ITT analysis with strata as fixed-effects on math outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.31	0.06	4.96	<.001	0.19	0.43	293.65
Math Pre Test	0.52	0.04	13.84	<.001	0.45	0.60	367.49

Table 10

Imputed ITT analysis with treatment effect heterogeneity on math outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.31	0.06	4.98	<.001	0.19	0.43	286.43
Math Pre Test	0.53	0.04	14.08	<.001	0.46	0.60	331.35

Table 11

Imputed ITT analysis without strata on Hindi outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.22	0.06	3.47	<.001	0.10	0.35	395.57
Hindi Pre Test	0.62	0.04	15.32	<.001	0.54	0.70	453.24

Table 12

Imputed ITT analysis with strata as fixed-effects on Hindi outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.22	0.06	3.54	<.001	0.10	0.34	374.94
Hindi Pre Test	0.59	0.04	15.21	<.001	0.52	0.67	447.97

Table 13

Imputed ITT analysis with treatment effect heterogeneity on Hindi outcomes. HC1 SE.

Term	Estimate	Std Error	Statistic	p Value	Conf Low	Conf High	Df
Treat	0.22	0.06	3.56	<.001	0.10	0.34	365.99
Hindi Pre Test	0.59	0.04	15.18	<.001	0.51	0.67	434.83