Beyond Average Impact Estimates: A Case for Examining Subgroup Differences in Locally-Conducted Group Randomized Trials

Authors: Andrew Jaciw, Denis Newman, Boya Ma, Empirical Education Inc.

Empirical Education

1. Purpose.

We often find that the primary concern of a local education agency conducting a group randomized trial is to measure differential impacts of an intervention on specific student populations in their local settings. If differential impacts can be detected readily with relatively small experiments, there is support for the general program described by Newman (2008) for mounting such trials. Our goals are:

1) To address power considerations in detecting differential impacts for student subgroups in cluster randomized trials.

2) To empirically examine whether the average difference in performance between subgroups of interest does not vary across clusters. This condition is important for obtaining added power to detect differential impacts between the subgroups.

3) If this condition holds in the experiments examined, to consider the implications for small-scale cluster randomized trials that address differential impacts for student subgroups of local concern.

2. A Motivation for Investigating this Problem.

One motivation for studying this problem is our finding, across several locally-conducted group randomized trials, that the estimate of the interaction between student-level covariates and treatment reaches statistical significance even when the average impact estimate does not.

One example is illustrated in the following tables and graph from a randomized trial on a technology intervention for Algebra. Although we detected an interaction with English proficiency, we were concerned that our estimate was not conservative enough—that it reflected uncertainty due to the re-sampling of students only, and not the re-sampling of the units of randomization (i.e, teachers.)

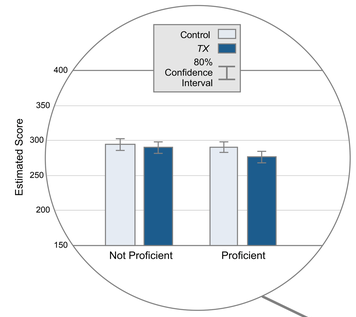


Table 1. Overview of Sample and Impact of *Treatment X* on CST Algebra

	Condition	Means	Standard deviations	No. of students	No. of classes	No. of teachers	Effect size	*p* value	Percentile standing
Unadjusted effect size	Control	294.26	43.11	453	28	13	-0.09	.93	-3.59%
	Treatment X	290.23	45.90	279	20	8			
Adjusted effect size	Control	294.26	43.11	The same sample is used in both calculations.			-0.26	.35	-8.32%
	Treatment X	282.81	45.90						

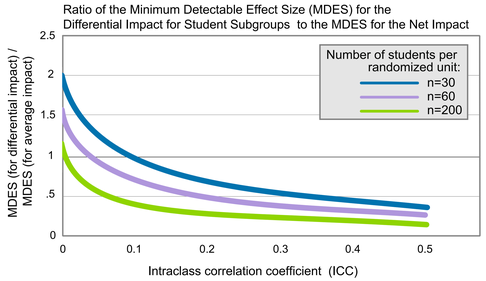
Table 2. Moderating Effect of English Proficiency on the Impact of *Treatment X* on CST Algebra

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Outcome for the non-English proficient control with an average pretest	293.93	24.68	6	11.91	<.01
Change in outcome for each unit-increase on the pretest	27.21	1.53	694	17.78	<.01
Control group difference (English proficient minus not proficient) in the outcome	-3.50	3.79	694	-0.92	.36
Effect of *Treatment X* for non-English proficient student	-3.94	12.48	6	-0.32	.76
Average difference (English proficient minus not proficient) in the effect of *Treatment X*	-10.13	5.87	694	-1.73	.08
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	432.26	271.57		1.59	.06
Within-teacher variation	972.11	52.17		18.63	<.01

Note. For .05<*p* value<.15, we conclude that we have some confidence that the effect observed is not due to chance only.

3. The MDES for Differential Impacts Versus Average Impacts.

Bloom (2005) notes that we have greater power to detect differential impacts than average impacts of the same size, when the subgroups of interest are below the level of randomization. This is because the differential impact estimator "differences away" the cluster error component and thereby eliminates the uncertainty due to between-cluster differences in average performance. The following figure shows the ratio of the MDES (minimum detectable effect size) for a differential effect, to the MDES for an average effect, holding other parameters constant. There is a power advantage for detecting differential effects when the value of the ratio is less than 1.



We observe that the ratio depends on the sample size of students in the randomization units and the ICC, but not on the number of randomization units. The advantage for detecting the differential impact (ratios below 1) is observed for values of the ICC that are frequently found. For example, with n=30, this advantage happens for ICC>0.1. The advantage for detecting the differential impact that is illustrated holds only under certain conditions. ***It will not hold if the average difference in performance between subgroups is not constant across randomized clusters.*** Therefore, we are interested in the more general expression for the standard error for the differential impact estimate, where we don't assume a constant average difference between subgroups.

(In the hypothetical example above, we assume two subgroups of students and an equal proportion of students in each subgroup; the formula used to generate the graph is given in Bloom, 2005.)

4. Modeling Result.

We obtain a generalized expression for the standard error (expanding on Bloom's (2005) derivation):

Assume that there are n students per cluster and J clusters. Assume that n/2 students are in each of two subgroups (e.g., boys and girls) in each cluster. Assume that J/2 clusters are randomized to each of two conditions.

Consider a model of performance of student i in cluster j (for schools in one of the two conditions.)

$$y_{ij} = \beta_0 + \beta_1 isM_{ij} + u_{1j} isM_{ij} + u_{0j} + e_{ij}$$

isM_{ij} is an indicator of subgroup membership.
β_0 is the grand average of performance for subgroup $isM_{ij}=0$.
β_1 is the cross-school average difference in performance between subgroups $isM_{ij}=0$ and $isM_{ij}=1$.
u_{0j} is the school-specific deviation in average performance for subgroup $isM_{ij}=0$.
u_{1j} is the school-specific deviation in the difference in subgroup performance from the cross-school average of this difference.
e_{ij} is the student-specific error term ($Var(e_{ij})=\sigma^2$).

If we assume that the variance in the estimate of the average subgroup difference is the same in both conditions, then under this model, we obtain the following expression for the variance of the differential impact estimate:

$$Var(\bar{\Delta}(istx=1)-\bar{\Delta}(istx=0)) = 4\left[\frac{Var(u_{1j})}{J}+\frac{4\sigma^2}{nJ}\right]$$

We see that:
1) u_{0j} is differenced away
2) if the subgroup effect is not constant across upper level units then the power calculation for detecting differential impacts needs to figure in this additional source of variation

5. Empirical Results.

We examined whether subgroup differences vary across clusters. In the following table we show the significance levels of the estimates of the variance, among units randomized, in the average difference in performance between subgroups (the u_{1j}'s discussed above). This is taken from a sample of eight experiments that we have conducted.

We establish empirically that, at least in some cases, the variance across schools in the average difference in performance between student subgroups is not statistically significant. This result supports Bloom's model which assumes that there is no variation among schools in the average difference in performance between subgroups. Under this condition we have more power to detect differential effects among subgroups of students than average effects of the same size.

Table 3. Significance Level of the Random Effect

			Subgroups (gray box indicates that this effect is not estimated for the experiment)			
Experiment	Intervention	Randomization	Gender	English Proficiency	Pretest	SES
1	Math technology (Grades 7-9)	24 classes (17 students per class)	.32	*	*	
2	Reading program (Grades 7-8)	28 classes (15 students per class)	*		*	.43
3	Reading program (Grades 3-5)	30 teachers/classes (5 students per randomization unit)	.32		*	
4	Science program (Grades 3-5)	92 teachers (22 students per teacher)	.27	*	*	*
5	Science program (Grades 3-5)	16 teachers (23 students per teacher)	.09	.15	*	
6	Math technology (Grades 9-12)	10 teachers (58 students per teacher)	.16		*	
7	Math program (Grades 3-6)	30 teachers (14 students per teacher)	.37		.02	.37
8	Reading program (Grades K-3)	30 teachers (4 students per teacher)	.38	.01	.14	

*the maximum likelihood procedure does not yield an estimate (this indicates that the model is too complex [Singer and Willett, 2003] which often implies that the random effect is too small to warrant estimation.)

6. Statistical Issues and Implications.

1) Effect sizes for differential effects – how big or how small?
2) Student moderators – student-level or upper-level effects?

Implications for doing Small Scale Trials:
This work provides support for a strategy of conducting relatively small experiments to answer questions of local interest to a school district (Newman, 2008). Small and less expensive experimental program evaluations focused on moderating effects can provide more valuable information to decision makers than large-scale experiments intended for broad generalization, which cannot provide confirmatory evidence for all interactions of interest to schools. These results suggest a strategy for investments in effectiveness research that builds up broader generalizations from smaller scale studies focused on local needs.

References:

Bloom, H. S., (2005). *Randomizing groups to evaluate place-based programs.* In H. S. Bloom (Ed)., Learning More From Social Experiments. New York, NY: Sage.
Newman, D. (2008). *Toward School Districts Conducting Their Own Rigorous Program Evaluations: Final Report on the "Low Cost Experiments to Support Local School District Decisions" Project.* Empirical Education Research Reports, Palo Alto, CA: Empirical Education Inc.
Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press.

Empirical Education Inc. • www.empiricaleducation.com
425 Sherman Avenue, Suite 210 • Palo Alto, CA 94306-1850
T: (650)328-1734 • F: (650) 328-1794

THE LEADER IN SCIENTIFIC EVIDENCE FOR SCHOOLS

Beyond average impact estimates: A case for examining subgroup differences in locally-conducted group randomized trials

Authors: Andrew Jaciw, Denis Newman, Boya Ma
Empirical Education Inc.

Purpose

We often find that the primary concern of a local education agency conducting a group randomized trial is to measure differential impacts of an intervention on specific student populations in their local settings. If differential impacts can be detected readily with relatively small experiments, there is support for the general program described by Newman (2008) for mounting such trials.

Our goals are:

1. To address power considerations in detecting differential impacts for student subgroups in cluster randomized trials.
2. To empirically examine whether the average difference in performance between subgroups of interest does not vary across clusters. This condition is important for obtaining added power to detect differential impacts between the subgroups.
3. If this condition holds in the experiments examined, to consider the implications for small-scale cluster randomized trials that address differential impacts for student subgroups of local concern.

A motivation for investigating this problem

One motivation for studying this problem is our finding, across several locally-conducted group randomized trials, that estimates of the interactions between student level covariates and treatment often reach statistical significance even when the average impact estimates do not. One example is illustrated in the following table and graphs from a randomized trial on a technology intervention for Algebra. Although we detected an interaction with English proficiency, we were concerned that our estimate was not conservative enough—that it reflected uncertainty due to the re-sampling of students only, and not the re-sampling of the units of randomization (i.e., teachers.)

Table 1. Overview of Sample and Impact of *Treatment X* on CST Algebra

	Condition	Means	Standard deviations[a]	No. of students	No. of classes	No. of teachers	Effect size	*p* value[b]	Percentile standing
Un-adjusted effect size	**Control**	294.26	43.11	453	28	13	-0.09	.93	-3.59%
	Treatment X	290.23	45.90	279	20	8			
Adjusted effect size	**Control**	294.26	43.11	The same sample is used in both calculations			-0.26	.35	-8.32%
	Treatment X	282.81[c]	45.90						

[a] The standard deviations used to calculate the adjusted and unadjusted effect sizes are calculated from the scores of the students in the sample for that row.

[b] The *p* value for the unadjusted effect size is computed using a model that figures in clustering of students in teacher level but does not adjust for any other covariates. The *p* value for the adjusted effect size is computed using a model that controls for clustering and that includes both the pretest and indicators for upper-level units within which the units (pairs) of randomization are nested.

[c] Modeling separate intercepts for upper-level units leads to estimates of performance, in the absence of treatment, which are specific to those units. For purposes of display, to set the performance estimate for the control group, we compute the average performance for the sample of control cases used to calculate the adjusted effect size. The estimated treatment effect, which is constrained to be constant across upper-level units, is added to this estimate to show the relative advantage or disadvantage to being in the treatment group.

Table 2. Moderating Effect of English Proficiency on the Impact of *Treatment X* on CST Algebra

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Outcome for the Non-English proficient control with an average pretest	293.93	24.68	6	11.91	<.01
Change in outcome for each unit-increase on the pretest	27.21	1.53	694	17.78	<.01
Control group difference (English proficient minus not proficient) in the outcome	-3.50	3.79	694	-0.92	.36
Effect of *Treatment X* for Non-English proficient student	-3.94	12.48	6	-0.32	.76
Average difference (English proficient minus not proficient) in the effect of *Treatment X*	-10.13	5.87	694	-1.73	.08
Random effects	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	432.26	271.57		1.59	.06
Within-teacher variation	972.11	52.17		18.63	<.01

Note: For .05<*p* value<.15, we conclude that we have some confidence that the effect observed is not due to chance only.

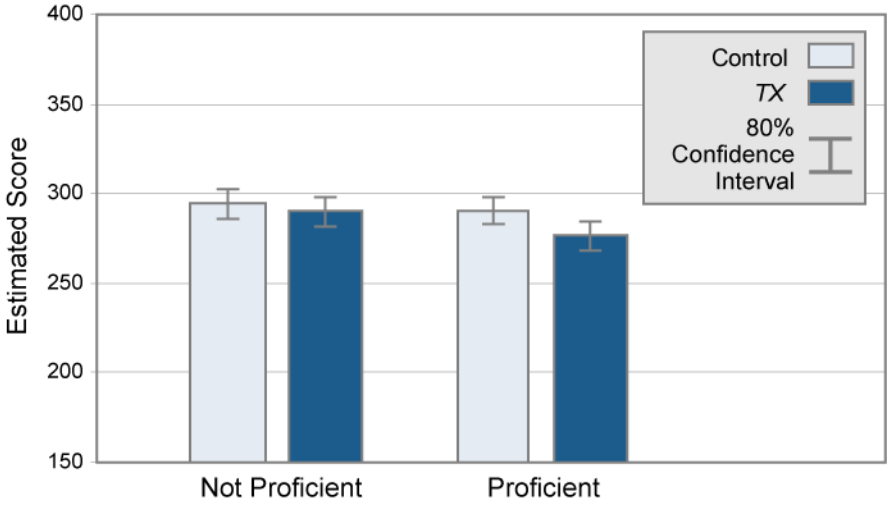


Figure 1. Moderating Effect of English Proficiency on Impact of *Treatment X* on CST Algebra

A model for computing power to detect differential impacts for student subgroups

Bloom (2005) notes that we have greater power to detect differential impacts than average impacts of the same size, when the subgroups of interest are below the level of randomization. This is because the differential impact estimator "differences away" the cluster error component and thereby eliminates the uncertainty due to between-cluster differences in average performance.

HOWEVER: We are interested in whether a particular assumption of Bloom's model holds; namely, whether the subgroup difference is constant across clusters. If this assumption does not hold, then the benefits due to the differencing away of the random intercepts may be counteracted.

In the following graph, the Y-axis represents the ratio of the Minimum Detectable Effect Size (MDES) for differential impacts (for student subgroups) to the MDES for the average impact. Thus values of the ratio that are smaller than 1 indicate we have greater precision to detect a differential impact of a given size than an average impact of the same size. (The formula used to generate this graph is given in Bloom, 2005.)

The X-axis represents the Intraclass Correlation Coefficient (ICC). The lines represent different sizes of the randomized units (e.g., with teacher-level randomization the blue, red and green lines represent cases where each teacher has 30, 60 and 200 students, respectively).

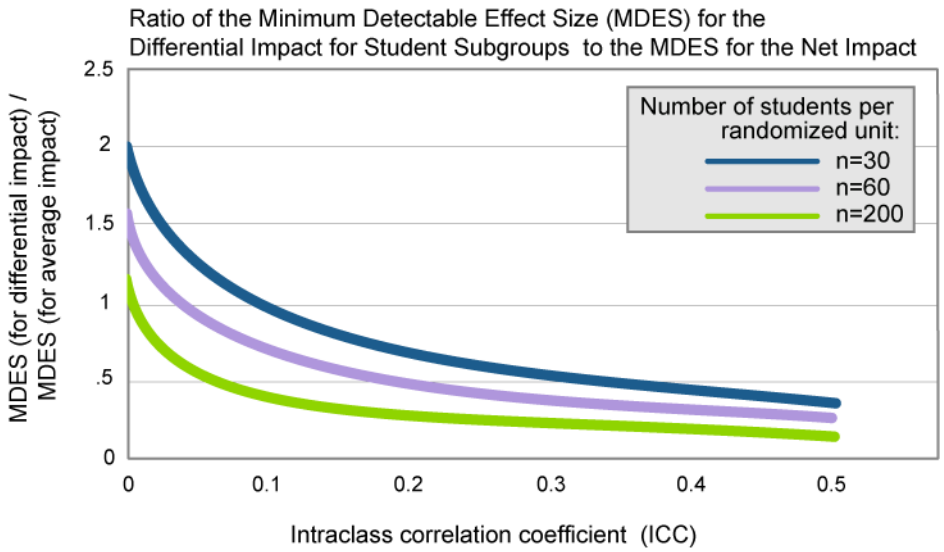


Figure 2. Ratio of the Minimum Detectable Effect Size (MDES) for the Differential Impact for Student Subgroups to the MDES for the Net Impact

We observe that the ratio depends on the sample size of students in the randomization units and the ICC, but not on the number of randomization units. The advantage for detecting the differential impact (ratios below 1) is observed for values of the ICC that are frequently found. For example with n=30, this advantage happens for ICC>0.1.

The advantage for detecting the differential impact that is illustrated holds only under certain conditions. ***It will not hold if the average difference in performance between subgroups is not constant across randomized clusters***. Therefore, we are interested in the more general expression for the standard error for the differential impact estimate, where we don't assume a constant average difference between subgroups.

(In this example, we assume that there are two subgroups of students, and that there is an equal proportion of students in each subgroup.)

Generalization of the Expression for the Standard Error

We first considered a generalization of the expression for the standard error of the differential impact estimate for the case where the subgroup difference is assumed to vary across clusters. We expand on Bloom's (2005) approach.

Assume that there are n students per cluster and J clusters. Assume that n/2 students are in each of two subgroups (e.g., boys and girls) in each cluster. Assume that J/2 clusters are randomized to each of two conditions.

Consider a model of performance of student i in cluster j (a treatment school).

$$y_{ij} = \beta_0 + \beta_1 isM_{ij} + u_{1j} isM_{ij} + u_{0j} + e_{ij}$$

isM_{ij} is an indicator of subgroup membership.

β_0 is grand average of performance for subgroup $isM_{ij} = 0$.

β_1 is the cross-school average difference in performance between subgroups $isM_{ij} = 0$ and $isM_{ij} = 1$.

u_{0j} is the school-specific deviation in average performance for subgroup $isM_{ij} = 0$.

u_{1j} is the school-specific deviation in the difference in subgroup performance from the cross-school average of this difference.

e_{ij} is the student-specific error term ($Var(e_{ij}) = \sigma^2$).

The average performance of students in school j who are in subgroup $isM_{ij} = 0$ is:

$$\overline{y_j}(isM_{ij} = 0) = \beta_0 + u_{0j} + \frac{\sum e_{ij}}{n/2}$$

The average performance of students in school j who are in subgroup $isM_{ij} = 1$ is:

$$\overline{y_j}(isM_{ij} = 1) = \beta_0 + \beta_1 + u_{0j} + u_{1j} + \frac{\sum e_{ij}}{n/2}$$

Consider the average of the difference between these means across all the treatment schools:

$$\overline{\Delta}(istx = 1) = \frac{\sum[\overline{y_j}(isM_{ij} = 1) - \overline{y_j}(isM_{ij} = 0)]}{J/2}$$

Assuming independence of the two subgroups, as well as the independence of the average subgroup difference across clusters:

$$Var(\overline{\Delta}(istx = 1)) = \frac{4}{J^2}\sum Var[\overline{y_j}(isM_{ij} = 1) - \overline{y_j}(isM_{ij} = 0)]$$

$$= \frac{4}{J^2}\sum[Var(u_{1j}) + \frac{4\sigma^2}{n}] = \frac{4}{J^2}[\frac{J}{2}Var(u_{1j}) + \frac{J}{2}\frac{4\sigma^2}{n}] = 2[\frac{Var(u_{1j})}{J} + \frac{4\sigma^2}{nJ}]$$

We assume an identical expression for the variance of the average difference between subgroup means across all control schools.

Assuming independence of observations across conditions, we have the following expression for the variance of the differential impact estimate:

$$Var(\overline{\Delta}(istx = 1) - \overline{\Delta}(istx = 0)) = 4[\frac{Var(u_{1j})}{J} + \frac{4\sigma^2}{nJ}]$$

We see that:

1) u_{0j} is differenced away

2) If the subgroup effect is not constant across upper level units, then the power calculation needs to figure in this additional source of variation. The power advantage obtained by differencing away the random intercept term, to some extent will be reduced by this additional source of random variation.

Empirical Results

We examined whether subgroup differences vary across clusters.
In the following table we show the significance levels of the estimates of the variance, among units randomized, in the average difference in performance between subgroups (the u_{1j}'s discussed above). This is taken from a sample of six experiments that we have conducted.

Table 3. Significance level of the random effect

			Subgroups (gray box indicates that this effect is not estimated for the experiment)			
Experiment	**Intervention:**	**Randomization**	**Gender**	**English proficiency**	**Pretest**	**SES**
1	Math technology (Grades 7-9)	24 classes (17 students per class)	.32	*	*	
2	Reading program (Grades 7,8)	28 classes (15 students per class)	*		*	.43
3	Reading program (Grades 3-5)	30 teachers / classes (5 students per randomization unit)	.32		*	
4	Science Program (Grades 3-5)	92 teachers (22 students per teacher)	.27	*	*	*
5	Science Program (Grades 3-5)	16 teachers (23 students per teacher)	.09	.15	*	
6	Math Technology (Grades 9-12)	10 teachers (58 students per teacher)	.16		*	
7	Math program (Grades 3-6)	30 teachers (14 students per teacher)	.37		.02	.37
8	Reading program (Grades K-3)	30 teachers (4 students per teacher)	.38	.01	.14	

We establish empirically that, at least in some cases, the variance across schools in the average difference in performance between student subgroups is not statistically significant. This result supports Bloom's model which assumes that there is no variation among schools in the average difference in performance between subgroups. Under this condition we have more power to detect differential effects among subgroups of students than main effects of the same size.

Statistical Issues

1. **Effect sizes for differential effects – how big or how small?**
   While it is useful to know that under certain conditions we have more power to detect a differential impact than an average impact of the same size, the question remains whether differential impacts are normally larger or smaller than average impacts. This is an empirical question and one that we will examine in a future study.
2. **Student moderators – student-level or upper-level effects?**
   A student-level covariate, such as the pretest, reflects both student and school effects. The power advantage for detecting differential effects across levels of this covariate discussed in this work applies only to a component of this covariate. In future work we will explore power for detecting differential effects for covariates that contain information at and below the level of randomization.

Implications for doing Small Scale Trials

This work provides support for a strategy of conducting relatively small experiments to answer questions of local interest to a school district (Newman, 2008). In this case, small-scale experiments that may be underpowered for detecting average effects may still yield valuable information about whether a treatment is differentially effective. Small, and less expensive, experimental program evaluations focused on moderating effects can provide more valuable information to decision-makers than large-scale experiments intended for broad generalization, which cannot provide confirmatory evidence for all interactions of interest to schools. These results suggest a strategy for investments in effectiveness research that build up broader generalizations from smaller scale studies focused on local needs.

References

Bloom, H. S., (2005). *Randomizing groups to evaluate place-based programs.* In H. S. Bloom (Ed). Learning More From Social Experiments. New York, NY: Sage.

Newman, D. (2008) *Toward School Districts Conducting Their Own Rigorous Program Evaluations: Final Report on the "Low Cost Experiments to Support Local School District Decisions" Project.* Empirical Education Research Reports, Palo Alto, CA: Empirical Education Inc.

Singer, J. D.,& Willett, J. B., (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press.

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 210
Palo Alto, CA 94306
(650) 328-1734