# Take Another Look: The Importance of Accounting for Unobserved Characteristics presented by Imperfect Randomization

Hunter Busick

2024-12-10

## Introduction

Any researcher understands how difficult it can be to perfectly align the pieces of a study. Participants must have the correct demographics and be available for specific periods of time to meet the demands of the study. Once you have the right people in the right place, you must then rely on your study design, as well as the participants willingness to cooperate. If a participant is in a particularly bad mood, maybe they will exhibit extreme response bias. If a participant is apathetic, they may exhibit neutral response bias, making it unlikely to yield interesting results. Once all these ducks are in a row, you must finally embark on a data cleaning process that will oftentimes take more effort than the actual statistics or analysis you are trying to run. I say all of this to highlight that conducting research and the statistical processes that come with it are often both difficult and time consuming. This poses the question: What do we do when we find an error in our methods in the post-study world? What if we or someone else finds a discrepancy or ethical issue in the way we conducted our work, but re-running the entire experiment does not make logistical sense?

The answer is often to use retrospective analysis. Once we have identified where a form of bias may have entered our experiment, we can then use bias correction methods to retroactively account for it's presence. Examples of bias corrections are when we place more weight on a population that was underrepresented in our experiment. We may use classification methods to assign a class label to a observation that didn't receive one. While these methods are great for going back and correcting errors that appeared in our work or the data, it is also important to consider that they do not come with zero risk. Increasing the weight for an underrepresented population will also magnify the role of outliers, if any exist within that minority. Additionally, we may overfit classifier models such that we generalize it to only be useful on our training set. A tangible example of bias being introduced to a study can be found in the Highscope Perry Preschool Program.

### Setting

The Highscope Perry Preschool Program was a longitudinal study that followed 123 African-American children beginning in 1962. These children were identified as being "high risk" for failing school as a result of low IQ scores and low socioeconomic status. The preschool aged children were initially assigned to to two groups in a series of 5 cohorts, 58 in the treatment group and 65 in a control group. Those in the treatment group attended a high quality preschool for 150 minutes each day, focusing on developing both the child's social and academic skills. Being in the program also came with a weekly 90 minute visit from a faculty member in the home of the child, thanks to a stellar student to teacher ratio. As we will see later, this binary variable of whether or not a students parent is available for this meeting is a gateway for bias into the study[1].

[1] https://heckmanequation.org/resource/perry-preschool-midlife-toolkit/

The outcomes of the program are hard to refute, with many impressive outcomes for those in the treatment group. Those who experienced treatment demonstrated less social misconduct resulting in incarceration, more healthy marriages, increased academic success, and greater socioeconomic outcomes than those who were in the control group. Additionally, cost benefit analysis shows that each child enrolled in the treatment group saved taxpayers $88,433 (Parks, 2000).

Despite these conclusions, we may draw concern to the fashion in which students from the larger group were separated into treatment and control. The process was as follows:

1) Younger siblings of those in the program received the same treatment program as their elder siblings.
2) Remaining children were ranked on Stanford-Binet IQ score at program beginning. Same scores were randomly prioritiezed within the sequence.
3) Keeping IQ scores roughly the same as before, participants were swapped between groups to balance gender and socio-economic proportions.
4) Two groups labeled control and treatment at random.
5) Participants in the treatment group whose mother was not available for the weekly meeting because of work were switched to the control group.

While this appears to be a normal process for separating subjects into groups, we open the door to biased results when we move children from the treatment to control group because they have a busy mother. In an entry to the Journal of Econometrics by James Heckman, Rodrigo Pinto, and Azeem Shaikh, this is defined as **imperfect randomization**: "When treatment status is reassigned after an initial randomization on the basis of characteristics that may be observed or unobserved by the analyst."(Heckman et al., 2023) The main goal of their paper is to propose a method in which results from studies with imperfectly randomized groups can still be inferred.

# Analysis of Methods and Results

One of the most important factors in making a statistical model is the ability to explain how it works to others so that they may critique it, or eventually incorporate it into their own work and studies. This paper will seek to both analyze the methods used in Heckman et al, and offer a simplified version of the methods through novel analysis. In this way, those with a less sturdy background in statistics and mathematics may still be able to apply a methodology with the goal of accounting for imperfect randomization.

### Analysis of Methods

The authors begin by defining variables for *potential outcomes*, *treatment group assignment*, and *observed characteristics* such as *gender*, *IQ score*, and *socioeconomic status*. They introduce useful shorthand to keep the equation digestible. After defining, they lay out the goal of their method to test a family of null hypotheses that controls the **familywise error rate**. Because we are testing on a series of $K$ potential outcomes it is imperative that we use the familywise error rate. Testing on multiple outcomes increases the chances that we falsely reject a null hypothesis simply because of our inflated amount of random fluctuations. $U_j$ is defined as a variable representing whether the mother was working.

We then move onto a series of "weak" assumptions establishing the conditions for running the test. It is important to note that these will not be generalizable and are study specific. We will revisit this lack of generalizability in our moral analysis. We establish conditional independence noting that outcome variables are independent of a given treatment characteristic. This ensures that our variables for observed and unobserved characteristics accurately account for outcome changes, and allows us to derive an effect of treatment status on these outcomes. We then note assumptions for invariance in balancing groups and symmetry of

balancing groups. These assumptions further allow us to construct fully random sampling in group assignment. Assumptions 3.4 and 3.5 are the most unique to our model and its workings. These assumptions work together to note that it must obviously be the case that the family has a working mother if a subject is moved from the treatment to control group after initial randomization. This is treated as a binary variable which affected reassignment, a feature rather unlikely to generalize to other studies.

Section 4 develops a method for testing a single null hypothesis. The first statements outline definitions for transformations of groups and symmetry of groups. Accounting for asymmetry is essential as this is what we introduced when the groups were imperfectly randomized based on whether or not the mother was working. Building transformations will allow us to later evaluate these hypotheses under multiple different variations of the treatment and control groups. We are able to construct the null distribution using these symmetries by evaluating all possible transformations. Lemma 4.1 makes it such that the assignment process follows certain symmetries, allowing us to construct this null distribution reliably. Theorem 4.1 then defines our probability for incorrectly rejecting a single null hypothesis.

We then move onto testing for a family of null hypotheses. As mentioned earlier we test for a family because we are testing for multiple outcomes at the same time. These multiple outcomes include cognitive and educational abilities, socioeconomic level, and behavioral tendencies. There are also different hypotheses constructed for males vs. females for each outcome. There are many advantages to testing a family of null hypotheses in this specific example, as it allows us to control our false positive rate while having a multitude of outcomes. Specifically we are controlling the familywise error rate by defining a maximal test statistic then use iterative stepwise testing. The method used for multiple stepwise testing was devised by Romano and Wolf (2005). Our authors combine previously defined terms as inputs into this process as follows:

1) Construct a null distribution for the test statistics and generate test statistics for all transformations of the data.

2) Define a maximal test statistic consisting of a subset of hypothesis in the current set and the test statistic for the k-th null hypothesis.

3) Stepwise test the hypothesis by:

   3.1) Beginning with all hypotheses.

   3.2) Compute critical values for each subset of $L_j$ (constructed of observed ($Z$) and unobserved ($U$) characteristics).

   3.3) Compare these critical values to the test statistic and see if we satisfy conditions for rejecting the null.

   3.4) Remove any rejected hypotheses from the larger set $L_j$ and iterate to the next.

   3.5) Halt the process when we no longer have any test statistic larger than our critical value.

Using this iterative process ensures that we maintain our type 1 error control by refining our hypothesis along the way. This procedure is effective for this specific example as its main goal is to exploit the fact that we still produce symmetrical test and control groups, despite imperfections in the randomization process.

**Result Analysis**

The authors compare three models. One assuming perfect randomization $L(Z, U = 0)$, one factoring in only observed factors $L(Z = 1)$, and one incorporating both observed and unobserved factors $L(Z, U)$. The main way of accounting for imperfect randomization in this paper is using a stepwise adjustment method while controlling for familywise error rate. The model assuming that randomization was perfect tended to overshoot how important the program was, as many results that are significant in this model are later not in models that utilize observed and unobserved characteristics. The $Z = 1$ model for only observed characteristics proved more reliable than the $U = 0$ model as it accounts for the same observed characteristics observed in the original study. This model incorporates factors such as gender, family dynamics, and socioeconomic

status. The $L(Z, U)$ model incorporates all of the observed characteristics from the $Z = 1$ model, and also introduces our unobserved characteristic ($U = 1$) of whether or not the mother was available for a meeting because of her work schedule. This is really the model that we have the highest expectations for as it incorporates the factor that caused our imperfect randomization in the first place.

We are most interested in the differences in significant outcomes between models. That is, what findings were different between the $L(Z, U)$ model and the first two? For cognitive outcomes in males we see improvements across all ages for those who underwent the treatment condition in all models, but less in magnitude for the $L(Z, U)$ model. The effect is similar for females. Educational attainment for males is significant across models for males, but yet again looses magnitude once we get to the $L(Z, U)$ model. Females see more of their educational attainment values loose significance once accounting for unobserved characteristics. A potential explanation for this ties to the role that young females play in the home when the mother is busy. Oftentimes females more so than males will take over roles associated with motherhood. As they age these roles may become more prominent and hinder a young woman's desire or ability to attain further education. Thus, it makes sense why there is no longer a significant difference between groups when we introduce our unobserved characteristic. For behavioral outcomes, males and females alike maintain strong significance throughout all models, suggesting that the differences in treatment vs. control can be strongly linked to program involvement, and that the mothers work hours didn't dictate behavioral changes later in life. Economic outcomes provide a very intriguing finding. Employment and earnings outcomes are moderately improved for those in the program in the first two models. However, when we use our third model incorporating our unobserved characteristic, we see the significances completely erased. This may suggest that a mothers involvement in a son's life is extremely important to their future economic success. Additionally economic success is often a culmination of academic and behavioral characteristics, which we saw loose strength of significance when switching to the $L(Z, U)$ model. Females saw no changes between models.

# Normative Considerations

### Ethically assigning treatment groups

One of the clearest normative considerations that this paper and the Highscope study as a whole gives rise to are the ethics of assigning treatment groups. In line with some of our findings from the methods analysis, it is not far fetched to think that those who could've benefited from the study the most were those whose mother may have been less involved in their childhood as a result of needing to work to keep the family afloat. Keep in mind that our children were defined as "high risk" for educational failure, likely implying that a rigorous or untimely work schedule for the mother was not something she is choosing for the sake of convenience. In this imperfect randomization process the researchers accidentally introduced a classifier variable that essentially represented whether the mother was spending a lot of time at home with the child, or whether or not they were working to keep food on the table. However, rather than being able to study the difference in program effectiveness between those with a more present mother, it was simply decided that those whose mother wasn't available would not get treatment. Even worse, this was decided after the initial randomization process, highlighting the fact that there were children set to undergo a life-changing program but were stripped of that opportunity because of circumstances outside of their control.

Appealing to the framework set forth by deontology, it is fair to say that we have violated the categorical imperative by treating the children switched from the group as means to an end, rather than ends themselves. Because we maintain so many of the significances after accounting for the unobserved characteristic of mothers working, it is fair to say that regardless of a child's mother's employment situation, they would have benefited from enrollment in the treatment group. We then treat the children as means to an end when we initially randomly assign them to receive treatment from the program, but then switch them after the fact. In this, we are ignoring the fact that this is a life altering experience, and rather than adjusting the program to fit the needs of the students, the students are shifted to meet the needs of the program, which at the end of the day the study was about. The consequentialist may also have a bone to pick with this imperfect randomization, as it widened the disparity between advantaged and disadvantaged children within the subset

of our experiment. While the consequentialist would likely argue that our subset of children having to be switched out of the treatment group so that the experiment could run was likely for the betterment of society, they would also argue that we widen the disparity by advocating for a program that some may only be able to afford with government assistance.

**Risks of Normalizing Imperfect Randomization**

What then do we think about about accounting for imperfect randomization generally? While I believe that this is a method that is great for looking back at experiments where we may not have been able to get the randomization right, we should be cautious in the middle of experiments when we suddenly realize that we will have to realign groups because of an unobserved variable. That is to say, we should not begin an experiment and be satisfied with non-perfect randomization practices simply because we know we can just perform retrospective analysis on the said unobserved variable like we have done above. It should remain the standard that randomization be perfect or as close to perfect as possible. In our specific example, maybe there were subjects whose mother could've met for a slightly lesser period of time, or maybe over multiple intervals. Once we have committed to randomizing our groups, it should then be our responsibility to alter the experiment in specific ways such that we do not have to re-balance groups, and subsequently perform analyses that we hope account for the unobserved variables we introduce as a result of imperfect randomization practices.

In this case the deontologist is most likely to appeal to the fact that in a random sample of a subset, all participants should have an equal chance of being selected. This is the backbone of simple random sampling which we utilize so often in research and discovery. Normalizing imperfect randomization techniques with the idea that we can account for them later perpetuates the idea that we do not have to treat all participants equal, and if we find a variable that is not convenient for us to work with, we can split groups based on it and label it an unobserved characteristic. While this proposed framework is a great step towards ensuring we pull the correct observations from our work, the consequentialist may argue that we risk becoming too reliant upon retrospective analysis for our own good, and damage outcome opportunities for our participants in the process.

**Explainability Concerns**

As discussed in the final days of our class, we have an obligation to produce methods and results that are properly interpretable. While the methods proposed by Heckman and colleagues are thorough, concerns may arise surrounding their explainability reproducibility. Since this is an introduction to a newly devised method, it only succeeds under fairly generous assumptions that will not generalize well to other studies. Additionally, the methods that they do perform are on a dataset that is not generally available to the public. One would have to reach out to Highscope and see if they have the qualifying credentials to work with the data spanning over generations and containing partially sensitive information. The combination of generous assumptions and a private dataset make it difficult to pull much from the paper unless if you are a well trained statistician. This is of no fault of the the producers of the paper, as this is how any key finding begins. Despite this, I still believe that accounting for imperfections in the randomization process is a method that many fields, from biology to economics, could benefit from in their studies. Researchers at this moment do not have a quick and easy way to run this retrospective analysis. Could we devise a method using more explainable statistics, simulated data, and R?

## Novel Analysis

I begin by simulating 200 observations of 5 simple variables below. It is important to note that all numbers are simulated for the sake of demonstration, and do not reflect any actual observation. The treatment group experiences more optimal outcomes for the sake of demonstration, and are imperfectly split by a combination of predictors, forming the unobserved variable.

Table 1: Initial Simulation

| Group | SES | Education | Income | Crime |
|---|---|---|---|---|
| Control | 43.73546 | 16.41856 | 31.56656 | 1 |
| Treatment | 56.51745 | 18.93662 | 45.82702 | 0 |
| Control | 41.64371 | 17.02959 | 29.51331 | 0 |
| Control | 65.95281 | 19.66972 | 36.48523 | 1 |
| Control | 53.29508 | 19.22516 | 43.17114 | 4 |

where *Group* represents control or treatment assignment, *SES* represents socioeconomic status with mean *50* and standard deviation *10*, *Education* is the number of years of education completed, calculated as mean *12* for control and *14* for treatment varying $1 * SES$ for each individual with standard deviation *1*, *Income* in thousands with mean *25* for control and *30* for treatment varying $.2 * SES$ for each individual with standard deviation *5*, and *Crime* inveresly related to *SES* using the Poisson distribution.

I then define predictor coefficients as the 4 variables other than treatment status and create a dataframe with their scaled values. I take this dataframe of scaled predictors and use the *kmeans* function to cluster observations into $k = 3$ groups. This value is then added back to our simulated data as *Unobserved_Characteristic*.

Table 2: Simulation + Unobserved Characteristic

| Group | SES | Education | Income | Crime | Unobserved_Characteristic |
|---|---|---|---|---|---|
| Control | 43.73546 | 16.41856 | 31.56656 | 1 | 1 |
| Treatment | 56.51745 | 18.93662 | 45.82702 | 0 | 2 |
| Control | 41.64371 | 17.02959 | 29.51331 | 0 | 1 |
| Control | 65.95281 | 19.66972 | 36.48523 | 1 | 2 |
| Control | 53.29508 | 19.22516 | 43.17114 | 4 | 2 |
| Control | 41.79532 | 15.57653 | 36.00060 | 1 | 1 |
| Treatment | 53.73240 | 18.98237 | 43.45744 | 0 | 2 |
| Control | 57.38325 | 17.32210 | 35.79328 | 0 | 1 |
| Treatment | 63.67183 | 19.99153 | 37.05070 | 0 | 2 |
| Control | 46.94612 | 16.32798 | 26.90609 | 0 | 1 |
| Control | 65.11781 | 18.21610 | 36.90663 | 4 | 3 |
| Treatment | 54.74754 | 20.91657 | 50.95810 | 1 | 2 |
| Control | 43.78759 | 15.68122 | 34.86604 | 1 | 1 |
| Control | 27.85300 | 14.39713 | 31.39246 | 3 | 3 |
| Control | 61.24931 | 18.77747 | 38.91298 | 1 | 2 |

I then run regressions using *SES* and the initial variables to compare to the same models with the addition of our *Unobserved_Characteristic*. We see that the model performance improves under the *Observed_Characteristic* and that the treatment effect is smaller once it is introduced, suggesting that there may be a grouping variable that was not initially present in our data that helps better explain relationships. This is similar to the effect seen for the variable representing mothers working in the Highscope experiment.

Table 3: Simple Regression Results

| Model | term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|---|
| Education | (Intercept) | 12.2954654 | 0.4101317 | 29.979310 | 0.00e+00 |
| Education | GroupTreatment | 2.0946811 | 0.1523326 | 13.750707 | 0.00e+00 |
| Education | SES | 0.0925921 | 0.0078499 | 11.795386 | 0.00e+00 |
| Income | (Intercept) | 22.7515143 | 2.0597354 | 11.045843 | 0.00e+00 |
| Income | GroupTreatment | 4.8004343 | 0.7650344 | 6.274795 | 0.00e+00 |
| Income | SES | 0.2370799 | 0.0394230 | 6.013741 | 0.00e+00 |
| Crime | (Intercept) | 2.1719383 | 0.3031293 | 7.165055 | 0.00e+00 |
| Crime | GroupTreatment | -1.5744714 | 0.1609813 | -9.780462 | 0.00e+00 |
| Crime | SES | -0.0262507 | 0.0061668 | -4.256796 | 2.07e-05 |

Table 4: Regression Results Including Unobserved Characteristic

| Model | term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|---|
| Education | (Intercept) | 13.5178510 | 0.4843700 | 27.9081100 | 0.0000000 |
| Education | GroupTreatment | 1.7337114 | 0.1704996 | 10.1684190 | 0.0000000 |
| Education | SES | 0.0628679 | 0.0098185 | 6.4030097 | 0.0000000 |
| Education | Unobserved_Characteristic2 | 0.9903882 | 0.2205646 | 4.4902413 | 0.0000122 |
| Education | Unobserved_Characteristic3 | 0.0232746 | 0.2097881 | 0.1109435 | 0.9117753 |
| Income | (Intercept) | 31.3393167 | 2.3048102 | 13.5973524 | 0.0000000 |
| Income | GroupTreatment | 2.2546136 | 0.8112997 | 2.7790145 | 0.0059858 |
| Income | SES | 0.0305237 | 0.0467200 | 0.6533330 | 0.5143113 |
| Income | Unobserved_Characteristic2 | 6.8138237 | 1.0495273 | 6.4922788 | 0.0000000 |
| Income | Unobserved_Characteristic3 | -0.0306991 | 0.9982489 | -0.0307530 | 0.9754980 |
| Crime | (Intercept) | 0.7080931 | 0.3761135 | 1.8826579 | 0.0597467 |
| Crime | GroupTreatment | -1.0643529 | 0.1847874 | -5.7598796 | 0.0000000 |
| Crime | SES | -0.0092725 | 0.0075611 | -1.2263328 | 0.2200735 |
| Crime | Unobserved_Characteristic2 | -0.0123275 | 0.2178771 | -0.0565800 | 0.9548797 |
| Crime | Unobserved_Characteristic3 | 1.1689431 | 0.1543722 | 7.5722404 | 0.0000000 |

# Conclusion

Running experiments is difficult, and sometimes there simply aren't ways to avoid the fact that groups have to be reassigned after we do our initial randomization process. This was the case in the Highscope Perry Preschool Program in which certain subjects were reassigned their group label after it was discovered their parents schedule conflicted with that of the program. Heckman and colleagues devise a method to account for this imperfect randomization, which is when participants are reassigned a condition based on an unobserved variable. While there are ethical concerns with normalizing reassigning groups and using retroactive analysis instead, this paper still provides a necessary first step towards what will hopefully one day be as simple as following a function guide in a statistical software.

**Works Cited**

Heckman, James, Rodrigo Pinto, and Azeem Shaikh. "Dealing with Imperfect Randomization: Inference for the Highscope Perry Preschool Program." Journal of Econometrics 243, no. 1–2 (December 2023). https://doi.org/10.3386/w31982.

Office of Justice Programs, and Greg Parks, Juvenile Justice Bulletin § (2000).

Romano, Joseph P., and Michael Wolf. "Stepwise Multiple Testing as Formalized Data Snooping." Econometrica 73, no. 4 (July 2005): 1237–82. https://doi.org/10.1111/j.1468-0262.2005.00615.x.