# Pre-Analysis Plan: 3. Spillovers from Sibling Choices*

Adam Altmejd

2017-04-19

## Contents



# 1 Introduction

The purpose of this project is to evaluate how peer experience influences education choices. We will look at how the education of the older sibling affects the younger's behavior. Using the university application data described in the introduction, we aim to answer two questions:

1. **Imitation**: How is a younger sibling's choice of higher education affected by the education of the older sibling?
2. **Inspiration**: How is a younger sibling's grades affected by the older's admission success?

*Stockholm School of Economics, adam@altmejd.se

The question of behavioral spill-overs to siblings is interesting for many reasons. For one thing, not many empirical studies of peer influence on decision making exist, and this study will be able to provide some well needed evidence on how we rely on our siblings. Furthermore, information availability about higher education is highly variable over socio-economic groups. In a low SES family, an older sibling that is admitted to college could inspire the other children to apply. Evidence of such behavior would be useful in order to understand the mechanisms of intergenerational mobility within families. Last, because of the nature of this data, We will be able to get quantitative and heterogenous measures of these effects. For example, one could imagine that attractive fields influence younger siblings to study the same topic, while less attractive fields have effects in the opposite direction. We will measure not only the direction but also the sizes of these different responses, quantities that could be of importance for education policy.

This document is a pre-analysis plan (PAP), registered in a public repository before the author has been given access to the dazta set needed for analysis. Without the possibility to explore the data set it is likely that we will run into many unexpected obstacles. If we for any reason need to deviate from this plan because of such circumstances it will be clearly stated.

## 1.1 Previous Literature

This project is closely related to Schrøter Joensen and Skyt Nielsen (2017). They study how younger siblings are affected by their older's choice of high school education when the older sibling experiences an increase in availability of high school maths. They find that younger siblings are 2-3 percentage points more likely to to choose a math/science track in high school if the older sibling was given a quasi-randomly introduced expanded choice margin for science related fields. Another related paper is Dustan (2018). He studies sibling spillovers in high school choice in Mexico in a regression discontinuity setup, also exploring the mechanisms potentially driving these effects.

There is a small literature about the effect of role models and social transmission mechanisms. Kosse et al. (2016) randomly expose both low and high SES children to a pro social mentor, and show that the observed gap in prosociality between groups of different SES is closed, even still 2 years after the treatment. Exposure to the mentor also increases probability to apply for gymnasium, the German academic track high school. Moreover, Dahl et al. (2014) find strong peer effects in parental leave uptake between coworkers and brothers, giving credence to the existence of a social transmission mechanism between siblings. Of course, there is also a large body of research on intergenerational mobility (see Black and Devereux (2011) for an overview and e.g. Fagereng et al. (2015) for some causal evidence), where correlations in both education attainment and earnings between generation is prevalent.

The existence of an information channel with an impact on student choices has been found in a number of studies. In a similar institutional environment to the Swedish one, Pekkala Kerr et al. (2015) use a large scale field experiment to show that informing students about labor market outcomes of different fields has a large impact on choices of the least informed students. Furthermore, Fricke et al. (2016) use random assignment of the subject matter of students' research paper to show that exposure to Economics increases the likelihood of the field being chosen for major. Hastings et al. (2016) perform a large representative survey of Chilean applicants and find that they systematically overestimate earnings of past graduates. Their respondents list prestige and accreditation as the primary reason for their degree choice. Between 35% and 47% of applicants do not even know what they will earn after graduation. There are many other papers that study the determinants of degree choice, all stressing the importance of non-pecuniary factors. Scott-Clayton (2012) provides a review of this literature, with many examples of how lack of information affects choice.

There is also evidence that the lack of information about higher education is unequal, and affects low SES applicants more. Hastings et al. (2015) use a randomly administered earnings disclosure policy that induces low SES individuals to apply for fields with higher returns. Their results fit well with the hypothesis that I present below; that earnings disclosure mostly affects low SES individuals could be because they are worse informed about the actual outcomes. Bowen et al. (2009) find that a large proportion of highly qualified but poor applicants did not attend the most selective institution that they were qualified for, even though such institutions often offer superior financial aid.

## 1.2 Hypotheses

For reasons outlined below we believe that the interesting behavioral responses will be found in analysis that allows for heterogeneity in behavior. Nonetheless, we will also study the aggregated response to use as a baseline model.

When looking at aggregate imitation response of younger siblings, we believe that the measured imitation behavior of the younger sibling will be indistinguishable from zero or slightly negative. This is because siblings will imitate when they are inspired, but not when they are dissappointed, and also occaisionally want to go their own way and thus avoid the choice of the older sibling.

Non-zero effects however will be identified when we look at how behavior varies with how positive the information transmitted from the older to the younger sibling actually is. Having an older sibling that is in medical school gives the individual a lot of information about what such studies entail. More precise estimates of career prospects as well as data on the difficulty of the program and what is actually tought in class, are all factors that the younger sibling initially probably only had a vague idea about. The behavioral response of the younger sibling to this information will thus depend on his or her prior knowledge about

the subject. The further the information is from these prior beliefs, the stronger the reaction. The same is true when the sibling is more uncertain and has flatter beliefs. We thus hypothesize that when looking at the distribution of sibling responses in research question (1) imitation effects will be stronger when the information transmitted is more positive, as these are more likely to cause a stronger positive news chock. When sorting by field quality, the worst fields will cause younger siblings to apply less frequently, and the best will cause the opposite reaction. Further, when comparing siblings accross socio-economic status, those with lower SES (flatter priors) will have stronger responses.

With regards to the second research question, we hypothesize that younger siblings of those students that are successfully admitted to their preferred choice will study harder and get better grades. The size of the grade-improvement will depend on the difference in grade requirements between the field that the older sibling was successfully admitted to and their next-best choice. When the difference is large, the younger sibling will be more encouraged to study hard.

These hypothesis will be tested using statistical tests described below. P-values below 0.05 will be deemed significant.

# 2 Analytical Framework

For an individual $i$, let $D_i$ be their choice of education among a finite set of choices $\mathbb{J} = \{0, 1, 2, \ldots\}$. Think of these as different fields or courses at different institutions. The variable $E_i$ is a measure of progress in that choice. Further, characterize $i$'s younger sibling (possibly more than one) by $s(i)$.

We are interested in a model where the younger sibling's behavior is causally affected by an information shock about the quality of choice $D_i$. We can describe the process using a function, unique for each younger sibling, $I_{s(i)}(E_i, D_i, \xi_{s(i)})$, below written $I_{s(i)}$. What information the younger sibling receives will depend on what field the older sibling is studying, $D_i$, how much they know about that field $E_i$, as well as factors, $\xi_{s(i)}$. The laet variable characterizes things like the relationship between siblings and how attentive the younger sibling is, but also, and perhaps most importantly, it captures the younger sibling's prior beliefs about the choice. For example, if the younger sibling already believes that a field is completely amazing, even fairly positive information could be be dissappointing.

For the first research question, we can describe this causal relationship with

$$\text{Imitation}_{s(i)}(D_i) = \alpha + \beta I_{s(i)} + \varepsilon_{s(i)}.$$

The $\beta$ coefficient captures the degree of imitation as the information content changes. While such an average imitation rate is interesting in itself, it is likely

that both the direction and strength of sibling imitation varies with different dimensions, which is why it will be important to study treatment heterogeneity.

For the second question, we have

$$\text{Inspiration}_{s(i)} = \alpha + \beta I_{s(i)} + \varepsilon_{s(i)}.$$

But here, the information transmission function might be very different. On one hand, inspiration might be correlated with imitation since if the younger sibling has lower grades than the older, an increase in preference for the older sibling's choice should also make the younger sibling work harder to secure an offer. On the other hans, an effect on the younger sibling's score could come from the older's success in itself. That the older sibling is admitted to the program of their dreams might inspire the younger to work harder even if they focus that effort on a completely different field of study.

We will now explain how we plan to identify and test these causal effects, first presenting the identification strategy and then defining all variables that will be used.

## 2.1 Causal Identification

It is well known that the correlation between education choices and different outcomes is highly endogenous (going back to Mincer (1958)). Students sort by ability, choosing different levels of education depending on how skilled they are to start with. Only with random assignment can we get any proper causal estimates. As explained in the introduction, this study will exploit two sources of exogenous variation; an admission lottery used to break ties and the discontinuities in admissions that can be found around each grade cutoff. Both these sources of variation affect $I(\cdot)$ through the $D_i$ variable. Students are randomly admitted to either a preferred choice (that we refer to as field $j$) or deferred to a lower prioritized option ($k$), producing information shocks about those specific choices relative to others.

We start with a reduced form model. Let $Y_{s(i)}^{\{1,2\}}$ be the outcome variable for either research question. The causal models above can then be described by

$$Y_{s(i)}^{\{1,2\}} = \alpha + \beta z_i + X_{s(i)}\gamma + \varepsilon_{s(i)}.$$

It yields the aggregate intention to treat estimates effect from the information shock. With $Z_i \in \mathbb{J}$ being the choice that the applicant is randomly admitted to, our instrument, $z_i$, is 1 whenever the successful randomization occurs and the applicant is admitted to their preferred choice,$z_i = \mathbb{I}(Z_i = j)$. Here, $X_{s(i)}$ is a set of control variables that we include to improve precision, and the $\beta$ coefficient gives the effect on ambition and imitation from the random admission.

### 2.1.1 Using heterogeneity to explore mechanisms

#### 2.1.1.1 News Quality Interaction

To test the second part of our hypotheses, that responses will vary heterogeneously with the positivity of the news shock we use an interaction effect between choice and shock quality, $d_i \times Q_j$. According to our hypothesis, the interaction effect coefficient should be positive and significant. We would then have the following second stage

$$Y_{s(i)}^{\{1,2\}} = \beta_1 d_i + \beta_2 d_i \times Q_j + \delta_1 Q_j + X_i\gamma + \varepsilon_{s(i)},$$

and two first stage equations,

$$d_i = \pi_1 z_i + \phi_1 Q_j + X_i\psi_1 + u_i$$

and

$$d_i \times Q_j = \pi_2 z_i \times Q_j + \phi_1 Q_j + X_i\psi_2 + v_i.$$

where $Q_j$ is a measure of how positive the news schock from starting preferred choice $j$ is. A higher value means a more positive shock. The interaction effect the captures the influence from this higher quality choice when the older sibling is actually admitted. We expect to see similar effects for both imitation and inspiration using this specification, but we will use slightly different definitions of $Q_j$ for the two models.

#### 2.1.1.2 Imitation by field, institution or city

Between a preferred and less preferred choice many things can vary. The applicant could be randomized between different schools, that sometimes lie in different cities, it could be a randomization between different programs at the same school, etc. To better understand what drives siblings to imitate we will look at these samples separately and evaluate the different magnitudes. Is it the case that siblings mainly follow to the same school, but are not as interested in studying the same field at different schools? Could this be because siblings prefer to live in the same city, creating easier access to e.g. housing. What if said city is the home town of the family?

#### 2.1.1.3 Other Subgroups

Moreover, we will test the aggregate models in different subgroups. We will divide the sample into three socio-economic status groups by the education level of the siblings' parents. Our hypothesis is that any effect will be attenuated by

higher socio-economic status, as kids with highly educated parents have much better access to information about university education. We will also look at the effects separately by gender.

Schrøter Joensen and Skyt Nielsen (2017) find an effect of sibling imitation but only for sibling pairs where the age difference is small. We will study how the age difference interacts with our treatment using an interaction model similar to the one presented above, and also just by dividing the sample into two groups, $\leq 4$ years and $\leq 10$ years, like they do.

They also study heterogeneity through birth order- and gender interaction effects. We will test their claims by by limiting our sample to the interaction between first- and second-born siblings, and also look at genders separately. Schrøter Joensen and Skyt Nielsen (2017) argue that competition is an important factor driving imitation, and find that a large part of the imitation comes from brother pairs. The younger brother is 70% more likely to choose STEM fields if their older brothers did so. However competition could also have the opposite effect, where the younger sibling does not want to risk loosing.

There is a different mechanism potentially at play when younger siblings imitate the choices of the older. Having an older student at a specific school or in a certain city could decrease the transaction costs of moving there. The younger sibling could perhaps move in with the older sibling. There are a number ways to study if this material transmission mechanism is driving the results. We will check if any results remain after removing those students who only apply to the exact same field-institution combination as their older sibling. We will also estimate institution-specific imitation effects and compare their size to the main results.

Last, to increase the sample size we will also test the effect including the siblings of those that lotteries and who are thus assigned to their less preferred option. How much does loosing the lottery increase the likelihood that the sibling applies there instead?

### 2.1.2 Instrumentation

The causal model of information transmission above makes it clear that it is not only the admission in itself that has an impact. Information is transmitted between siblings all throughout the older's studies ($E_i$) and is affected by many other factors ($\xi_{s(i)}$). Our instrument directly affects what choice the subject is admitted to ($Z_i$), but in the model, information transmission is affected by the student actually starting their studies ($D_i$). To capture the actual treatment effect we will use the random admission to instrument for the applicant starting their studies. As was argued in the introductory document, the assumptions needed to get accurate estimates of the LATE are likely to hold, but we will of course also test and report the strength of the instruments. In our main analysis, we will also interact this instrument with an estimate of how surprising

the information is to try to get at some of the variation caused by $\xi_{s(i)}$. In supplementary specifications we will explore different approaches and also try to capture the effect from $E_i$.

Let $d_i = \mathbb{I}(D_i = j)$, be an indicator variable that is 1 if the older sibling starts studying their preferred field $j$. Starting from the simple aggregate versions of the models, using 2SLS we can write the second stage as

$$Y_{s(i)}^{\{1,2\}} = \beta d_i + X_i\gamma + \varepsilon_{s(i)},$$

and instrument for information transmission with the first stage

$$d_i = \pi z_i + X_i\psi + u_i.$$

This specification produces LATE estimates for the complier group. However, as we argued above, the effect is likely heterogenous and the estimates from this model will actually be a weighted average of many different treatment dimensions. Our hypothesis for this aggregate effect is that $\beta$ will not be significantly different from zero when studying sibling imitation, although somehwat positive when estimating the effects on younger sibling ambition.

## 2.2 Robustness

Supplementary to the main specifications above, we will estimate a number of alternative models. The purpose is to (1) distinguish the information transmission mechanisms from other possible causes of imitation and inspiration, (2) analyze how the effect varies across different interesting sub groups, and (3) test the robustness of the findings.

A problem with the above tests is that the data actually consist of multiple different experiments, one for each admission margin, and it is not completely clear what the $\beta$ coefficient above would capture. Moreover, the control group is not well-defined. Failure in the lottery leads to deferral to a less preferred choice, but these next-best choices vary between individuals and also admission into them is not certain. Some lottery losers are offered a spot in their next-best choice, others are deferred again.

To get around these issues we will estimate the treatment effect separately for different choices, and test our hypotheses also on these disaggregated models,

$$Y_{s(i)}^{\{1,2\}} = \sum_j [\beta_j d_{ij}] + X_i\gamma + \varepsilon_{s(i)},$$

with one first stage for each preferred choice $j$,

$$d_{ij} = \sum_{j'} [\pi_{j'} z_{ij'}] + X_i \psi_j + u_{ij}.$$

Each instrument $z_{ij} = \mathbb{I}(Z_i = j)$ is 1 if an applicant $i$ is randomly admitted to $j$. We use all instruments in every first stage. The $\pi_{jk}$ coefficients can then be thought of as the fraction of students randomly admitted to choice $j'$ who still transmit information about choice $j$. When $j = j'$ this is the rate of compliers, while all other $\pi_{j \neq j'}$ are fractions of always takers from other fields.

To test our hypotheses on this disaggregated model is somewhat complicated. An F-test should be significant, as we believe the full model does have explanatory power. We can also rank all choices by quality and should then see a more positive effect for better alternatives. However since the supplementary analysis should be seen mostly as exploratory, we will refine this analysis after we have received the data.

One problem with this disaggregated model is that without collapsing choices the number of treatment margins is very large (one for each combination of choices). To get around this we will follow Kirkeboen et al. (2016) and pool choices by field of study and institution separately. But we will also use machine learning techniques to identify the most important heterogenous causal effects of the full model, and then hopefully be able to test these on the larger data set that will become available at a later stage.

If selection varies systematically not only by what choice the applicant is admitted to but also by what their less preferred choice is (as Kirkeboen et al. (2016) argues it does when looking at financial returns) we should estimate the second stage and the set of first stage equations separately for each such next best choice $k$, adding fixed effects also for next-best fields (see their paper for more details on this specification). While it does not seem likely that this preference margin is as important when explaining sibling responses we will include the specification as well.

To check robustness further we will also test different definitions of the outcome variables and endogeneous variables, and vary the bandwith around the admission cutoff for the regression discontinuity approach. But since we have yet to explore the data, it is not clear how to exactly specificy these supplementary tests. There will probably be many aspects of the data set that could create confounders, and we will want to study if these affect our results. Any such supplementary analysis will be important, but should be seen as exploratory rather than as evidence for or against our hypotheses.

# 3 Variable Definitions

We have yet to exactly define many of the main variables of the models presented above. How to exactly measure outcome is far from obvious. We want variables that do not induce unecessary noise, but on the other hand do answer to changes in the instrument.

Since we did not yet explore the data, many of the exact definitions of variables below have been made based on very little information. To avoid data mining, the current definitions will be used in the main specification, but we will also want to study how changing them affects results. These variations should be seen as robustness checks and considered exploratory. If it happens that a different definition seems to perform better in the first data set, we will hopefully be able to test this on the supplementary data that will be shared with us at a later stage. But then we will first register a new version of this pre analysis plan where it has been clearly specified.

## 3.1 Main Variables

The main identifying variable (instrument in 2SLS) is the result of the admission lottery, but we will also study the effects using the regression discontinuity approach, with the variable being predicted admission based on what side of the grade cutoff the applicant's score was.

When estimating the IV equations also the endogeneous variable needs to be clearly defined. A major part of the information about a field of study is transmitted from older to younger sibling in the early stages of the education and is captured by $D_i$. At this time one is introduced to a lot of data on what it means to be a student in the field and what the degree could actually be used for. This definition also allows to include as many observations as possible, which is why it will be used in the main model. More specifically we use a dummy variable that takes the value 1 when the older sibling is randomly admitted and actually starts (finishes at least some credits within the program), and 0 otherwise.

As the older sibling continues through their studies, they do learn more and can potentially supply new valuable information ($E_i$ in the causal model). Supplementary to the main specification, we plan to explore many versions of the endogeneous variable to try to capture this mechanism. We will make use of the (non-random) variation in time between when the two siblings starts, both due to age difference and preferences, and include one dummy for each extra year that the older sibling is within the program before the younger applies. On a smaller subsample we can use actual wage after graduation. Also the disaggregated analysis over choices will be informative, as we can look at the distribution of effects over different field and institutions. Given that the extra data set can be used, we will also try an automatic approach and data mine for

strong instruments for information transmission that we can then test out of sample after having registered the strategy in an updated version of this plan.

Optimally, we want to somehow measure not only the information transmission, but rather a *news* effect that takes the priors of the younger sibling into account, to distinguishign between information that is new and surprising, and that which is not. The interaction with $Q_j$ is included with the purpose to capture some of this effect. As an alternative to the two main definitions of the variable (below) we will also try to directly measure the "surprise" of a specific choice with the residuals from a regression model that predicts school popularity on observable characteristics, and try to estimate the value added between different choices if possible.

### 3.1.1 Imitation

For the first research question there are two potential routes for how to specify the dependent variable. Either using a binary measure of whether or not the younger sibling includes the older siblings choice in their application, or a discrete measure of the actual rank of the older sibling's choice. The benefit of the second is that it captures an interesting intensive margin of preferences, but its problem is that the length of the total application list is not fixed.

- The primary, binary, measure is $Y^1_{s(i)} = \mathbb{I}(R_{s(i)}(j) = 1)$, an indicator function that is equal to 1 if the younger sibling ranks the older sibling's preferred choice $j$ as their most preferred choice, and 0 otherwise.
- The discrete, supplementary, measure is $Y^1_{s(i)} = \frac{R_{s(i)}(j)}{\max(R_{s(i)})}$, the ranking of $j$ in the younger sibling's application, divided by their total number of applications (to get around the problem of variable ranking length).

We define $Q_j$ as the popularity of choice $j$ to proxy for news quality. More specifically, we will rank all choices by average selectivity (in terms of application score) and break ties by ranking on the number of eligible applicants to that choice. This way we get a ranking of all choices by popularity that we can use to approximate how positive the information is. One problem with this measure is that popularity is probably also correlated with knowledge about a specific field, which would decrease the potential impact of the interaction. If younger siblings have more accurate priors about popular choices, the information transmitted for these options will have lower news value even though it is very positive. As described above, we will include additional measurements to alleviate this potential problem.

### 3.1.2 Inspiration

For the second research question, the dependent variable should be a measure of performance or grit that captures how the younger sibling is motivated to study

harder by the fact that older is admitted. For our main test we will use a measure of the change in the younger sibling's GPA. The score will be standardized over the whole applicant pool, separately for each type of admission group score.

- The main specification will use the difference in standardized GPA between elementary school and high school for those siblings that did not yet have a high school degree when the older sibling applied, $Y^2_{s(i)} = \Delta\text{GPA}_{s(i)}$. However, this requires younger siblings to be in the early years of high school at the time of applications since their elementary school grades need to have been set beforehand. This limits the sample to some extent.
- To increase sample size we will also use an exploratory version which measures the change across individuals, and where the outcome variable is $Y^2_{s(i)} = \text{GPA}_{s(i)}$.

In the interaction specification for this research question we define $Q_j$ somewhat differently. With the thought that the student will be more inspired, the larger the "impact" of the random assignment is, we let the variable be defined as the difference in required standardized admission score between the preferred $j$ and next-best choice $k$. We take the difference between the required score in the admission group for $j$ where $i$ is randomized and the cutoff score in the same admission group but for the less preferred choice $k$. We use standardized required scores over choices for each admission group. The interaction effect coefficient will then give the effect on the change in younger sibling grades (in s.d.) from admission, when the older sibling is admitted to a choice that is 1 s.d. harder to get in to than their next-best alternative.

## 3.2 Control Variables

We will use the same control variables for both research questions and include them in the main specification to increase precision. As a robustness check we will estimate the model without controls as well. The variables are:

- Gender of both siblings and a gender interaction effect
- Age at application and age difference between siblings
- Cohort (year fixed effects)
- Number of siblings
- Foreign background (binary, according to SCB's definition)
- Application score of the older sibling
- Preferred choice (or field/institution) fixed effects, to control for differences in preferences across sibling pairs
- When studying imitation: the grades of the younger sibling that were set at the time when the older sibling was randomly admitted
- When not specifically studying subgorups by parental education:
  - fixed effects for parents completed level of education, separately for each parent and each level of education (primary, secondary, tertiary)

- parental income: mean of household disposable income (`DispInk`, by consumption unit) when applicant is between ages 13 and 16
- dummy variables for if any of the parents studied the same *field* as the older sibling is applying to

### 3.3 Sample Selection and Construction

When constructing the data set we will have to make a number of decisions on what data to keep and exactly how to measure each feature. Before matching the SCB data to the application data we will construct an application data set that contains one observation per individual, focusing on the relevant admission margins where randomization has occurred (and thus only include a preferred choice $j$ and a less preferred choice $k$). To produce this data set we will:

1. Keep only applications to degree programs and drop applications to free-standing courses.
2. Remove invalid applications. This could be when the student ends up not being eligible or when application data is missing for some reason.
3. Keep only the first application period for each individual to a degree program where they either (a) pariticipate in a lottery, or (b) have an application score close enough to a cutoff. Set choice over which the randomization was performed to choice $j$ (the preferred choice).
4. Identify the correct treatment margin, i.e. what would the applicant be admitted to if the lottery failed, and set this to the next-best choice $k$.
5. When there are multiple randomizations, keep the margin that includes a successful admission.

As we discussed above we will have to pool applications into aggregated fields or institutions when looking at the effect heterogeneously, and collapse choices into these pooled variables. For example, if a student applied to medical school in two different cities as their preferred choice, then to three engineering schools, and last to a business school; we collapse their choice into ($j$) medicine, ($k$) engineering and ($l$) business. Or if collapsing by institution, if the applicant has medical school at Karolinska Institutet at their first choice, psychology at the same school as their second and medicine at Lund University as their third, collapse into Karolinska ($j$) and Lund ($k$). We then also need to find those applications where the treatment margin "bites", and keep only those observations where the preferred choice and next-best choice are in different fields/institutions.

Sometimes there will be multiple relevant randomization margins for one individual. When using the admission lottery we will then use the first lottery that the applicant wins. I.e. if the applicant is in a lottery for field $j$ and looses only to participate in a lottery for $k$ that he wins, keep the $j/k$ margin. If he looses the second lottery, and instead is admitted to $l$, use the $k/l$ margin. Since we do not have access to the lottery numbers, we have no way of knowing what

would have happened to an applicant that wins the first lottery but would have faced a second lottery had they lost. In this case we keep the first $j/k$ margin. Similarly for the RD approach. If the applicant is predicted to be admitted to field $k$, but with slightly lower grades would be admitted to $l$ and with slightly higher to $j$, we keep the $j/k$ choice margin.

This will yield a data set of applicants that have been randomly admitted to field $j$ rather than field $k$, with one observation per individual. I will then join this data to the individual characteristics data set from SCB, matching siblings and parents. The final data set will contain one observation per sibling pair.

After joining the data we will also drop all those cases where the younger sibling already has some university education when the older is treated. We will also drop any observations where the dependnt variable is missing. For example, if a younger sibling is not in high school when the older is treated, we cannot calculate $\Delta\text{GPA}_{s(i)}$ for them and would drop the observation.

# References


Black, Sandra E., and Paul J. Devereux. 2011. "Recent Developments in Intergenerational Mobility." In *Handbook of Labor Economics*, edited by David Card and Orley Ashenfelter, 4:1487–1541. http://www.sciencedirect.com/science/article/pii/S0169721811024142.

Bowen, William G., Matthew M. Chingos, and Michael S. McPherson. 2009. *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton University Press. http://press.princeton.edu/titles/8971.html.

Dahl, Gordon B., Katrine V. Løken, and Magne Mogstad. 2014. "Peer Effects in Program Participation." *American Economic Review* 104 (7): 2049–74. doi:10.1257/aer.104.7.2049.

Dustan, Andrew. 2018. "Family Networks and School Choice."

Fagereng, Andreas, Magne Mogstad, and Mats Rønning. 2015. "Why Do Wealthy Parents Have Wealthy Children?"

Fricke, Hans, Jeffrey Grogger, and Andreas Steinmayr. 2016. "Exposure to Academic Fields and College Major Choice."

Hastings, Justine, Christopher A. Neilson, Anely Ramirez, and Seth D. Zimmerman. 2016. "(Un)Informed College and Major Choice: Evidence from Linked Survey and Administrative Data." *Economics of Education Review*, Access to higher education, 51 (April): 136–51. doi:10.1016/j.econedurev.2015.06.005.

Hastings, Justine, Christopher A. Neilson, and Seth D. Zimmerman. 2015. *The Effects of Earnings Disclosure on College Enrollment Decisions*. Working Paper 21300. National Bureau of Economic Research. doi:10.3386/w21300.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad. 2016. "Field of Study, Earnings, and Self-Selection." *The Quarterly Journal of Economics* 131 (3): 1057–1111. doi:10.1093/qje/qjw019.

Kosse, Fabian, Thomas Deckers, Hannah Schildberg-Hörisch, and Armin Falk. 2016. *The Formation of Prosociality: Causal Evidence on the Role of Social Environment*. IZA Discussion Paper Series 9861. IZA. https://papers.ssrn.com/abstract=2761338.

Mincer, Jacob. 1958. "Investment in Human Capital and Personal Income Distribution." *Journal of Political Economy* 66 (4): 281–302. doi:10.1086/258055.

Pekkala Kerr, Sari, Tuomas Pekkarinen, Matti Sarvimäki, and Roope Uusitalo. 2015. *Post-Secondary Education and Information on Labor Market Prospects: A Randomized Field Experiment*. 9372. IZA Discussion Paper Series. IZA. http://legacy.iza.org/en/webcontent/publications/papers/viewAbstract?dp_id=9372.

Schrøter Joensen, Juanna, and Helena Skyt Nielsen. 2017. *Spillovers in Educational Choice*. SSRN Scholarly Paper ID 2548702. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2548702.

Scott-Clayton, Judith. 2012. *Information Constraints and Financial Aid Policy*. Working Paper 17811. National Bureau of Economic Research. doi:10.3386/w17811.