**Nick**

* We discussed changing the section headers to match the Heilmier questions, I saw a TA recommend this on Slack as well
* We may need to be more specific about what tools we will use, the project description has this: “how you will do it (what tools? e.g., SQLite, PostgreSQL, Hadoop, Kinect, iPad, etc.)”
  + We could mention D3 for the data viz if we are set on that
  + For the ML do we have a preference between R or Python? I get the vibes the answer may be Python
* The lit survey...after re-reading the project description I think we may need to slightly change the approach, we need to discuss all of the following for each paper, not just drop in a reference. This is also 60% of the proposal grade.
* “For each paper, describe
* (a) the main idea,
* (b) why (or why not) it will be useful for your project, and
* (c) its potential shortcomings, that you will try to improve upon.”
* We need to add what each group member will do to the proposal somehow:
  + “Provide a **plan** of activities and time estimates, per group member. **List what each group member has done, and will do.”**
* Need to explicitly state distribution of team effort:
  + “[-5% if not included] Distribution of team member effort. Can be as simple as "all team members have contributed similar amount of effort". If effort distribution is too uneven, I may assign higher scores to members who have contributed more.”

**Ben**

Approach. I reworded and added sentences into the Approach section and added the two papers I reviewed on charter schools. I consider these in our scope of SES as confounding factors.

“Taking the graduation rate data from https://data.nysed.gov/, we intend to use a multiple models to characterize the relationship between academic performance and SES. We will join any of the available relevant data sets and isolate the most important factors that affect academic performance at the school, neighborhood, district, and state level. We will also compare other factors like school types (e.g. charter schools) to socioeconomic factors .

We plan to use measures for academic performance including, but not limited to, aggregate test scores and aggregate graduation rates. Then using modern [Use "D3" instead of "modern"?] visualization techniques, we will illustrate differences and allow the stakeholder to explore relationships using our interface. We seek to make correlation relationships, not causal relationships, as the data may not be sufficient to include all latent factors.

We will explore machine learning and statistical methods to find appropriate models. For SES, we are using NRC (at the school / district level) and Perhaps County Historical Employment and Wages Data (at the county aggregate level) as features. We also plan to investigate confounding factors that could affect conclusions on SES, such as school types and teacher education.

We will also explore using fixed effect models as used in economic studies of charter school factors [Winters, Jinnai]. Fixed effects models are linear regression models combining confounding factors, such as student/school characteristics, with factors of interest such as charter school exposure in Winters or charter school penetration in Jinnai. However, for our project, charter schools would be a confounding factor for SES effects. We may need to perform feature engineering to define the factors appropriately as with the public-to-charter move term calculated by Winters. We must also consider nuance as some factors like charter schools may only affect grades levels (e.g., 6th grade) that are overlapping [Jinnai].”

Time. This is one requirement: [-5% if not included] Distribution of team member effort. Can be as simple as "all team members have contributed similar amount of effort". If effort distribution is too uneven, I may assign higher scores to members who have contributed more.

I added the following in Time:

“For the current status, Dave has written the first proposal draft in LaTex, and all other members are reviewing the draft. Nick is making the presentation. Ben has written code for processing Access database files for initial exploration. All members have identified literature articles and added summaries.”

On the concept of not using “racist” models, I believe I would agree, but there is no definition here for what is meant.

I suggest we define specifically and clearly what is a “racist” model so it can be reasonably avoided and not risk using a loaded word without more context. The word "stereotyping" or the phrase "making generalizations" are much clear to what we are trying to avoid without leaving the reader/grader guessing.

Note the two charter school related papers I found did not note this as a major risk. They include **student characteristics** (likely included race/gender/etc.) as input data in order to factor out these effects when looking at the coefficients of other effects they are seeking, but their study did not discuss individual student characteristics and outcomes.

The **scientific ethical rule** here to follow here is to have a model that accounts for **all** **important factors** available in our data to be sure we can see an accurate effect that we want to consider and not ignore **multiple latent factors** affecting our output. However, we also have to be careful not to include variables that are **highly correlated** in a ML model, or the feature importance will be diluted. For instance, using football player position and weight are highly correlated variables to predict some health outcomes - WR weigh less than OL.

I believe the **moral ethical rule** we would follow is not to show results that intentionally or accidentally create generalizations/stereotypes, which is why we would be careful about what factors we publish and keep them in line with the purpose of our study.