General Comments

I thought the writing was overall very good. There was good use of topic and stress positions and a logical flow of ideas. There weren't many issues with sentences trying to serve multiple purposes, which definitely helped ease of reading. The main general issue seemed to be that section 3.2 might need some reorganization to clarify ideas.

Specific Comments

Section 3 Intro

The section starts by giving an outline of the method used. I think it might make for a gentler introduction to briefly restate the goal and some important considerations that were made when designing the methodology. I find it easier to understand the how when the why is fresh in my mind.

Paragraph 1, Line 4

The sentence says "For each birth cohort, we compared the fraction of individuals...." It's not clear to me what this was being compared to. Are you comparing usage proportions between cohorts within a specific survey year?

Paragraph 2

This sentence seems long and complicated and probably contains multiple ideas that could use a stress point. I think you could in the first sentence cite that almost all marijuana users start before 25. In the second sentence you could mention the benefit of how the question is structured. I also found that part confusing the way it's written now. When you say "to prevent..." to implies a negative consequence of some alternative question, but I'm not sure what that alternative is meant to be.

Section 3.1 Introduction

I thought this introduction to the section was excellent. The sentences flowed very well from previously discussed topics to new ideas. Sentences also did a nice job of using topic and stress positions.

Section 3.1.1 Paragraphs 3 and 4

These two paragraphs explain where the year of birth variable comes from. It seems like you're saying that even though it's technically confidential, it's a simple function of variables that are in fact publically available (at least for the majority of respondents). That makes me think a more ssuccinct explanation may be possible.

Paragraph 6

I found this paragraph confusing. Paragraph 4 seems to say that the reported "year of first use" is actually imputed from the reported age of first use and true year of birth. Doesn't that mean *the reported year* will sometimes be wrong but subtracting the age will always give the true year of birth?

Section 3.2.1, Line 1

In the first sentence, I think the outcome variable should be described simply as whether the person had used marijuana by age 25. The idea these responses are

coded as 0 and 1 seems like a mathematical abstraction that's not relevant yet. I also think the second sentence could be linked up better, maybe saying something like "This variable comes the NSDUH question 'How old..."'.

Paragraphs 2 and 3

These paragraphs seem to mostly discuss how the range of cohorts was chosen. I'm not sure if this belongs in a section called "Generating our outcome variable". The section also ends with an equation specifying the mathematical coding of the response, but it doesn't seem to be connected to the preceding material.

Section 3.2.2, Line 1

This paragraph uses the phrase "our outcome variable" just like the previous section, but it seems to be describing something different. It might be better to distinguish between an individual outcome and a population outcome.

Equation 2

It's confusing that y and y_i are totally unrelated. Also, "population domain" and "population units" seem like undefined technical terms. The variable x_i is defined but doesn't seem to be used anywhere.

Equation 3

Is this the actual distribution or just an approximation justified by large sample size? I actually don't know if you need to be explicit about this. Also, this equation should probably be switched with equation 4.