# Analysis of Non-response in Student Surveys

Ronaldo Iachan[1], William H. Robb[2], James G. Ross[1], Katherine H. Flint[1], and Alan Bloch[3]
ORC Macro, Calverton[1]
ORC Macro, Burlington[2]
Office of Smoking and Health, CDC[3]

## Abstract

This paper investigates non-response in school surveys at the levels of student and item non-response with a focus on whether hard-to-reach students differ significantly from other students along key dimensions. The probability of student non-response will be modeled as a function of school characteristics (merged from the school sampling frame), student demographics, number of absences and whether the student was administered a makeup survey (absent initially). Propensity (logistic) models will help identify the best predictors of non-response. As a by-product of multivariate models, we investigate weighting approaches that incorporate these predictors in weighting adjustments. We will compare alternate weighting procedures in their effectiveness in reducing the potential bias of non-response and variance impacts. At the item level, additional variables will include position in the survey and type of item. Finally, we will examine multilevel models for item and student non-response using these same predictors at the student and school levels, models that take into account between-item and between-student correlations.

## 1. Introduction

This article assesses the non-response biases associated with the 2004 National Youth Tobacco Study (NYTS 2004). Non-response bias refers to the potential for systematic under-representation and consequent bias in survey results due to non-response. Non-response bias analyses can offer a multitude of implications including statistical adjustments to the response data, changes in how responding institutions or individuals are recruited, and accommodations in how the survey itself is fielded.

In the NYTS, non-response may occur at three levels, namely at the school level, student level, and at the level of missing data for questionnaire items. These will be designated school non-response, student non-response and item non- response. (In general, item non-response should be distinguished from unit non-response, or total non-response.) These different levels of non-response are examined in this article together with an analysis of their potential bias.

In general, unit non-response bias may arise as a function of two factors: 1) the magnitude of non-response, and 2) differences between respondents and non-respondents along the key outcomes of interest in the survey. The first factor is typically quantified by response rates attained at different levels; in this case, at the school and student levels. The impact of the second factor can only be assessed when one has information about non-responding units, either for those key survey outcomes or for variables believed (or known) to be highly correlated with these outcomes.

The study of non-response bias includes first an assessment of whether it may constitute a serious problem in a particular survey, overall or for certain population subgroups. If a problem is revealed in a given study, then this assessment may also suggest that more intensive efforts should be undertaken to attain participation overall or for certain subgroups. Even in surveys where non-response is not considered a serious problem, as in the NYTS, efforts may be suggested to reduce or to adjust for the residual bias that may be induced by non-response.

Then, the analyses may suggest ways for compensating for potential non-response. As noted above, statistical adjustment methods typically involve the use of weighting class adjustments and post-stratification adjustments. This article provides an initial assessment of non-response bias in the 2004 NYTS, and suggests additional ways of reducing the bias potential even further.

Section 2 describes the response rates achieved in the 2004 NYTS at the school and student levels. This section also includes a discussion of school-level non-response. Because so few schools refuse to participate, and any potential school non-response bias is negligible, this issue is regarded as trivial and not examined in depth here.

Section 3 investigates differential student non-response within schools as a function of a variety of school characteristics. These analyses of student non-response use aggregate school-level data as a way of circumventing the lack of student-level data for non-respondents. These analyses suggest some useful potential correlates of student non-response.

The analysis in Section 4 examines auxiliary data available to assess non-response bias, and potentially to adjust for this bias. These auxiliary data collected in the 2004 NYTS survey include the number of absences reported by a participating student during a 30-day period. These data also include whether the survey was completed as part of the original administration or in a make-up session. We find that these variables are associated with key survey outcomes (e.g., tobacco use); therefore, they may be usefully incorporated in weighting class adjustments.

Finally, Section 5 presents analyses of item non-response. This analysis focuses on the effects of item placement and other potential factors that may explain missing data for survey items. In particular, Section 5 asks whether middle school students tend to have insufficient time to complete the NYTS and, if so, should a special, shortened version of the NYTS be created for middle school students.

## 2. Response Rates and School Non-response

Participation rates all cycles of NYTS have been exceptionally high. On NYTS, 267 of 288 or 93% of sampled schools participated. Of 31,774 selected students in the 267 participating schools, 27,933 or 88% participated. This results in a combined participation rate of 82%.

A detailed analysis of school non-response seems academic because the school participation rates are very high. Moreover, school subgroup sample sizes would be too small for school non-response analysis to be practicable. As described in the next section, however, it is possible to develop instructive analyses of student non-response data aggregated at the school level.

## 3. Student Non-response: A Within-school Analysis

Student non-response analyses are limited by the sparseness of individual student data that are available for non-respondents. However, the analysis of student non-response aggregated at the school level can capitalize on the gamut of school-level data available in the school sampling frame (QED files). For this analysis, we merged the following variables as potential explanatory for each sample school: percentages of minority students (Percent Black and Percent Hispanic), poverty levels, urban status (3 categories) and school level (middle or high school).

For public schools, we modeled student response rates within schools as a function of these auxiliary school-level variables. Specifically, log-linear (logistic) regression models were developed for the participation rate, i.e., for the proportion P(i) of students responding in school-i. [1]

Before developing these log-linear models, we also examined the distribution of the student participation rates within schools. The median student non-response rate across schools is approximately 8%, and that about half the schools have student non-response rates between 5% and 18%.

Table 1 presents the results of this modeling. It shows that schools with larger percentages of Blacks or Hispanics have significantly higher rates of student non-response. In addition, high schools have significantly lower student response rates than middle schools. By contrast, poverty levels and urban status did not exhibit significant effects on student non-response.

**Table 1 - Log-linear Regression Model for Participation Rates within Schools**

| School-level Variable | Estimated Coefficient | Standard Error |
|---|---|---|
| Percent Black | -0.017 | 0.0034 ($p<0.0001$) |
| Percent Hispanic | -0.015 | 0.0034 ($p<0.0001$) |
| Poverty | 0.0033 | 0.0038 (NS) |
| Urban status (3 categories) | -0.332 | 0.135 (NS) |
| Level (HS vs. MS) | -0.1544 | 0.135 ($p=0.015$) |

Again, we stress the value of this approach for finding correlates of student non-response albeit in a somewhat indirect fashion. The findings confirm that student response rates are lower for minority students (Blacks and Hispanics) and high school students. These lower response rates might lead to non-response bias if these groups have different survey characteristics; e.g., different tobacco use prevalence. The usual NYTS weighting procedures, however, compensate for this likely bias by including race/ethnicity and school level

[1] For non-public schools, however, almost all of these variables were unavailable. Therefore, the analysis by school type was limited to a comparison of response rates for public and non-public schools. A t-test suggests a significant difference in student participation rates within schools between these two school types.

(grade) among the key factors for post-stratification and other weight adjustments. Nevertheless, these findings do imply that the sampling design could result in lower absolute numbers of Black students, Hispanic students, and high school students than targeted in the sampling design. Two potential courses of action may then be considered: (1) modification of the sampling design to increase the targeted yield of Black, Hispanic and high school students; and (2) strengthening our recruitment methods to increase cooperation and participation among Black, Hispanic, and high school students.

## 4. Student Non-response

Student non-response is generally regarded as a potential source of bias in estimating health risk behavior. It is broadly believed that higher-risk students are less likely to attend school and more likely not to complete a survey because of absenteeism. To assess the effects of absenteeism on non-response, we included an item on the questionnaire to measure the amount of absenteeism each participant experienced over the prior 30 days. We also documented for participating students whether they completed the questionnaire in the original group administration or as part of a make-up session. This second piece of information was intended to assess the extent to which the make-ups might reduce the effects of absenteeism on response rates and on the biasing effects of student non-response. This section assesses the potential non-response impact of student absences and the role of make-ups in mitigating the effect of absences. The rationale behind this line of inquiry is the feasibility of performing weight adjustments based on student absenteeism data.

### 4.1 Rationale for Analyses of Student Absenteeism Data

The rationale for analyses of student absenteeism data is the potential for weight adjustments based on the number of self-reported absences during the prior 30 days, taken in conjunction with whether the student completed the survey as part of a make-up session rather than during the original administration. In principle, such weight adjustments are designed to reduce the potential for non-response bias that would result if students who tend to have frequent absences tend to differ from other students along key outcomes measured in the survey (e.g., smoking and tobacco use in general).

Several approaches may be considered that take into account the number of absences. Conceptually, if a student is absent “m” times during the previous 30 days, as reported in Question 7, the student weight could be adjusted upward to reflect his/her reduced probability of selection. In other words, students reporting with frequent absences can be allowed to represent those students actually absent during the day(s) of survey administration.

Conceptually, the adjustment factor is based on the fraction “F” of days present out of the total T=22 school days in the period, F= (T-m)/T; for example, the adjustment equals 2 if the student is absent half of the time, and it equals 1 for students with no absences in the period. As the question (Q7) collects categorical data on absences, any such adjustments would involve at best an approximation. However, this information can be used in the effective development of weighting classes as suggested below.

However, these weighting adjustments to reflect absence patterns may overstate the level to which higher absence students should be allowed to represent students who are actually absent for the survey and do not complete a make-up session. Therefore, it is also important to take into account make-up survey administration. The actual benefit of weighting up data reported by students with relatively high absence rates to reduce non-response bias is reduced by these make-ups. This occurs because intensive efforts are made to schedule make-ups for those students most likely to have frequent absences, and most likely to induce non-response biases in survey estimates.

We considered alternative approaches that also take into account the degree to which the make-up process compensated for absenteeism in each participating school. These approaches refine the same concept to inflate the weights of students who either a) have frequent absences in the previous 30 days, or b) were absent during the day of the survey, or both (a and b). With this approach, weighing classes are constructed that are based on both factors, absences (Q7), and make-up information. This approach seems preferable because the number of absences collected in Q7 is a categorical variable with only 5 categories.[2]

[2] I.e., it’s not continuous as required by an adjustment of the type explained above using a fraction “F”.

### 4.2 Application of Absenteeism and Make-up Data in Bias Assessment

In the NYTS, weighting class adjustments typically adopt weighting classes based on design strata. Instead of, or in addition to these adjustments, we may recommend the use of weighting classes based on these two factors, absences and make-ups.

Table 2 illustrates the potential benefits of these approaches for bias reduction. It shows that smoking prevalence rates increase steadily with the number of absences (over the past 30 days). As shown below, these two measures--makeup indicator and number of absences—are associated. However, the measure of absenteeism is a better predictor of risk behaviors (smoking) because it is measured on a finer scale with more categories than the dichotomous indicator of make-up administration.

**Table 2 - Smoking Prevalence (Ever smoked) by Number of Absences**

| Number of Absences (Q7) | Ever Smoking Prevalence |
|---|---|
| 0 days | 30.2% |
| 1 day | 39.6% |
| 2-5 days | 49.2% |
| 6-10 days | 58.1% |
| More than 10 days | 68.4% |

As mentioned earlier, non-response bias is a function of two factors: a) the amount of non-response, and b) differences between respondents and non-respondents in the key survey characteristics. Table 2, together with estimates computed for other prevalence estimates, all suggest that frequent absentees tend to differ from survey respondents in the key survey outcomes, and therefore indicate potential benefits in adjusting for absenteeism. The potential benefits of these adjustments are conditional on the impact that absenteeism may have on survey response rates (the first factor above), an impact that was presumably minimized through intensive follow-up efforts including make-up administration of the survey.

To help assess how much the make-up surveys mitigate the potential impact of absences, we investigated differences in the distribution of absences between the subgroups of students taking make-up surveys and those with regular survey administration. As expected, absences are much more frequent among the make-up students. Statistical (chi-square) tests showed significant differences between the two subgroup distributions ($p < 0.0001$); these differences go well beyond those expected simply because many students became eligible for a make-up administration because they were absent at the original administration.

These findings confirm the potential utility of exploring methods for the calculation of non-response adjustments based primarily on self-reported absenteeism during the 30 days prior to the survey, with some possible adjustment for the magnitude of make-ups in each participating school.

## 5. Item Non-response

Analyses of item non-response were based on two working hypotheses. The first was that some items might be experiencing high non-response rates because of a variety of factors other than placement, including lack of clarity, perceived sensitivity or invasiveness, and lack of importance. The second was that item non-response will become substantially greater near the end of the questionnaire and that this surge in item non-response approaching the end of the questionnaire will correlate inversely with grade in school, with middle school students experiencing higher item non-response toward the end of the questionnaire than high school students.

This section examines the potential effects of item position (early vs. late) and respondents' school, identifies items with particularly high rates of missing data (item non-response rates), and then plots rates of item non-response across all items.

### 5.1 Effects of Questionnaire Position

It is difficult to tease out the causes of item non-response because these effects can be both confounded and compensating (i.e., work in opposite directions). One reasons for placing items considered more important earlier in the survey questionnaire is the expectation that item response rates may be lower for later items; and survey designers don't want to risk losing too much data on key items or sections.

For the topic-related analysis, we divided the questionnaire into 4 sections, while noting that several alternative partitions could be considered:

Section 1, Questions 1 to 7: Demographics
Section 2, Questions 8 to 37: Cigarette Use
Section 3, Questions 38 to 49: Other tobacco products
Section 4, Questions 50 to 81: Other areas (ETS, attitudes, media etc.)

We then computed the average item non-response over items in each of these sections with results shown in Table 3.

**Table 3 - Average Item Non-Response Rate by Section**

| Section | Questions | Average Item Non-Response Rate |
|---|---|---|
| 1 | Q1-Q7 | 3.69% |
| 2 | Q8-Q37 | 3.04% |
| 3 | Q38-Q49 | 2.59% |
| 4 | Q50-Q81 | 3.89% |

These results illustrate that other, confounding effects compensate for the potential effects of item position in the questionnaire. Item non-response rates are highest for the Section 4 questions which come at the end of the questionnaire. They are also relatively high for questions related to tobacco products considered more obscure by respondents as well as for demographic questions that may be considered more invasive. In fact, item position may be considered less influential than other factors including item saliency, sensitivity, and clarity of the questions. Non-response rates for individual items are examined in more detail further below.

### 5.2 Potential Effects of Grade Level

The analysis of item non-response also investigated whether the decline in item response rates may be explained mostly or partly by the age or grade of the student and the consequent slower reading speed among younger students. In other words, a hypothesis of interest is whether item non-response in items positioned late in the questionnaire is significantly greater for middle school students (grades 6-8) than for high school students (grades 9-12). More generally, we investigated whether the amount of missing data for later items increased for the lower grades.

For this comparison, we considered the average item non-response proportion for items 1-30 (“Early Items”) compared to the average item non-response proportion for items 31-81 (“Late Items”) separately for middle school and high school students. Table 4 demonstrates that there is no evidence that middle school students have higher mean item non-response rates than high school students for late items (or, for that matter, for early items).

To the contrary, item non-response rates for later items are greater for high-school students than for middle school students. Section 5.4 also highlights the differential rates of decline in item response rates between middle school and high school students as respondents move further into the questionnaire. That section provides model-based analyses of the decline in item response rates with the separate models fit for middle school and high school groupings being virtually indistinguishable.

**Table 4 - Comparison of mean missing data for items placed early and late in the questionnaire for High School and Middle School students**

| | Early Items | Late Items |
|---|---|---|
| Middle School (n=14,034) | 2.14% | 3.09% |
| High School (n=13,738) | 2.68% | 4.01% |

The lack of evidence for higher mean item non-response rates among middle school students suggests that a more finely grained analysis by grade could be revealing. Table 6 presents mean item non-response rates (or proportion missing data) for the same two blocks of survey items categorized as early and late. This table shows that average item non-response for the late items does not increase with age; to the contrary, mean item non-response for the late items is highest for students in grades 9, 10 and 12. The lowest item non-response rates were among students in grades 7 and 8.

### 5.3 Analysis of Item Non-response

This section looks at non-response rates for individual questionnaire items. The first part of the analysis examined specific item non-response rates separately for middle school and high school students, as well as for the overall student sample, in an effort to identify items that may be particularly problematic. (These results are not included in this paper.) In the second part of the analysis, models were developed to reflect the premise that item response rates decay as students progress further into the questionnaire.

The model-based analysis focused on the position of the item (i) in the questionnaire. This analysis examines the proportion of item non-response (or missing data) strictly as a function of the item’s position expressed as “i” so that i=1 for Q1, i=2 for Q2, and so on. We developed models for the proportion of item missing data, P(i) for item-i, first over all students, and then separately for middle school and high school students (stratified models).

Because the dependent variable is a proportion, namely the non-response rate P(i) for item-i, a transformation is needed to go along with a simple linear regression model. The analysis considered a regression model for the logit of P(i), that is, for the log (P/(1-P)).

Figure 1 shows the result of this linear regression fit for logit(P). The model had a good fit (R-square=0.21) with a slope of 0.0136 suggesting a significant increase in the amount of missing data (item non-response) as we move further into the questionnaire. An additional finding of interest is that the first-order auto-correlation is 0.19 suggesting that items near each other tend to have similar item non-response rates.

Figure 2 shows the fits of similar models stratified by grade level; i.e., separate models were developed for middle school and high school students. Both models had good fit (R-square=0.19 and 0.21) and demonstrate the same pattern of declining response—or increasing item missing data-- as the overall model. The rate of decline seems slightly greater for middle school students, suggesting that a small effect of reading level may occur. In other words, slower reading or other age effects for middle school students may tend to lower item response rates for the items positioned late in the survey; however, these effects are far from significant.

## 6. Conclusions

Some expected results were obtained from this analysis of school, student, and item non-response. First, high school students are less likely to participate in the survey than middle school students. Second, as the percent Black or Hispanic increases in a school, student participation rates decrease. The exploratory analysis of absenteeism data suggests that use of self-reported absence data might provide real utility in weighting up data of high absentees who completed the questionnaire to account for high absentees who did not complete a questionnaire.

One surprising finding was that grade in school does not have a definitive effect on mean item non-completion rates. The highest mean item non-response on later items on the questionnaire was not found among middle school students but among high school students; this can be explained more plausibly as a function of sagging motivation rather than slower reading speeds.

**FIGURE 1**
**Proportion of Item Missing Data as a Function of Item Position in Questionnaire**

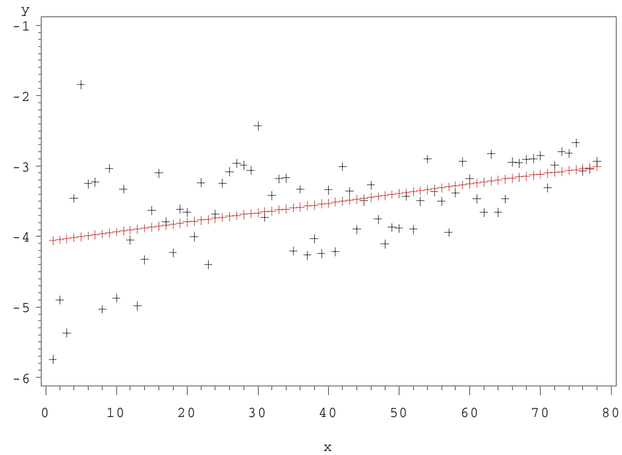


**FIGURE 2**
**Proportion of Item Missing Data as a Function of Item Position: Separate Models for Middle School (in Green) and High School students (in Blue)**