



2025 Minnesota Student Survey Data Removal and Cleaning Procedure

If questions arise, contact the Minnesota Student Survey Interagency Team at MDE.StudentSurvey@state.mn.us. View our website, mss.mn.gov, for additional resources.

Note: variable names used in this document refer to the variable names in the data file, not in the Minnesota Student Survey (MSS).

Overall, these checks include:

- Check for mischievous responding:
 - Excessive marking
 - Straight-line marking
 - Highly improbable combinations (e.g., age 12, grade 12)
- Check for logical inconsistencies within items (both yes and no)
- Check for logical inconsistencies across items
- Check for appropriate denominators (the question was presented to the right population or subset)

Criteria for Removal

1. Remove cases in which the student did not answer any questions beyond the background section. For 2025, this would be questions A1-A12, grade through sexual orientation.
2. Remove cases in which age is sharply incompatible with reported grade. Also, remove anyone who said they were 21 or older, unless they are in a special population group (ALC or JCF).
3. Remove cases that show pattern of exaggeration around tobacco use. Exaggerated tobacco use is the reported use of all **five** kinds of tobacco (cigarettes, cigars, smokeless tobacco, e-cigarettes, and hookah) on 20+ days of the past thirty days. (Note, in previous administrations we also looked at exaggerated drug use, since the 2025 questions on drug use did not ask about the number of occasions used in the past year, but simply have they used, it was decided in consultation with DHS subject matter experts to not exclude these cases.)
4. Remove cases for patterns of excessive marking or straight-line marking (so named because of students drawing a straight line through responses on the paper version of the survey). These occur when questions are organized into lists or boxes, and students mark the extreme answer for all questions in the list. Twelve criteria were used in 2025. Cases were removed if the pattern of excessive marking was

found on **three or more** of the twelve items listed below. (Note, in previous years we used the criteria of 2 or more cases of excessive marking, since some of the items were changed to mark-all that apply questions and we added a new check (marijuana source), we reduced the criteria for removal to 3 or more.

Excessive marking criteria

- a) All race/ethnicity groups were checked.
 - b) All gender identities were checked.
 - c) All racial/ethnic subcategories were marked among students who marked “Asian/Asian American,” “Black, African or African American,” “Hispanic or Latino/Latina” and/or “Middle Eastern or North African.”
 - d) Students marked all 18 reasons for missing days or part days of school (for regular public school students) or scheduled classes or assigned activities (for online students).
 - e) Students reported being bullied or harassed for all eight reasons listed.
 - f) Students reported going to all six locations listed after school for a typical week.
 - g) Students reported participating in all eight kinds of organized activities in a typical week.
 - h) Students reported drinking all eight kinds of beverages listed four or more times per day during the last seven days.
 - i) Students reported engaging in all five kinds of gambling “every day”.
 - j) All nine sources of e-cigarettes were marked among students who reported using e-cigarettes.
 - k) All nine sources of alcohol were marked among students who reported using alcohol in the last 30 days.
 - l) All nine sources of marijuana were marked among students who reported using marijuana.
5. Clean cases showing pattern of logical inconsistency. There are several situations in which survey answers can be clearly inconsistent. Though the survey software utilized was programmed to prevent such occurrences, cases were still examined and checked in 2025. For example, here is a sampling of some of our checks:
- a. Answering “No” to dental problems yet marking one or more dental problems.

- b. Answering “Yes” to activities participated in, and answering questions about why they do not participate in activities.
- c. Answering “none of these” and one or more adults student can talk to.
- d. Answering “no” and “yes” to suicidal thoughts.
- e. Answering “no” and “yes” to suicide attempts.
- f. Answering “no” to gambling, and answering gambling questions.
- g. Answering “did not drink alcohol in the last 30 days” and answering where they got alcohol in the last 30 days.
- h. Answering “have never used nicotine/tobacco” and answering questions about their use of tobacco products.

Additional Case Removal

Additional cases were removed for the following reasons:

1. The survey was determined to be a test survey (i.e., a single or small number of surveys taken on a day *before* a day that a large number of surveys was taken). These were determined on a case-by-case basis.
2. The survey was the only survey in a building. Since some such cases are legitimate (especially in charter schools) and would still contribute toward state and county results, these were determined on a case-by-case basis.
3. Students were asked to enter their school code and school name when they started the survey. There were a number of cases where the code and name were entered incorrectly. We manually checked each record when we could not determine the school or district the student belonged to. We removed cases where school and district could not be ascertained programmatically or by manually checking each case.

Other Data Checks

1. Run frequencies and crosstabs of questions that involve skip patterns to be sure that denominators of follow-up questions match the number who responded yes (or whatever response triggered the skip) in the base question (denominators match or are actually *smaller* due to missing data in the follow-up questions).
2. Check the coding of **all** variables in **all** survey levels **separately** to be sure that the same coding scheme is used for questions that are common across levels.

3. Run frequencies of all variables by **online vs. in-person** to be sure that all types are getting the appropriate questions.
4. Level 1, Level 2, and Level 3 data need to be together in one final data file. Need to make sure items are lined up correctly in the data file.
5. When asking the students for their height and weight, there are quite a few respondents who added a comma, extra zeros, or a space between the values. Those cases were manually fixed.
6. When grade (A1) was missing, we imputed grade based on level of survey and age, it did not correct all instances of missing grade, but did correct most cases (e.g., if grade is missing, but Level = 1 and age = 10 or 11, imputed grade = 5).
7. Items that were separated by online vs. in-person schools (e.g., reasons for missing school) were combined into one variable when appropriate. Variables include: A23, A25.
8. Items that were worded for the Level 1 survey (5th grade), but have the same meaning in the Level 2 and 3 surveys, were combined into one variable, these items include: A136C, A58A, A58N.
9. Check that the Level taken is consistent with the student's grade. If Level = 3 and grade = 7 or 8, it is assumed the student was given the incorrect level and the record is removed.
10. As noted previously, if there is excessive marking on 3 or more items, the student's responses are assumed to be mischievous, and the record is removed. However, if fewer than 3 cases are noted, the record remains in the data file. For these records, we clean the excessive marking by changing the response to every option of the question to missing. Items cleaned in this process include:
 - a) Checking all options to the race/ethnicity subgroups (e.g., checking all 13 Asian subgroup categories)
 - b) Checking all Race categories (American Indian or Alaskan Native; Asian, South Asian, or Asian American; Black, African, or African American; Hispanic or Latino/Latina; Middle Eastern or North African; Native Hawaiian or Other Pacific Islander; White)
 - c) Checking all reasons for missing school
 - d) Checking all eight beverages, four or more times per day
 - e) Checking all gender identities
 - f) Checking all gambling activities as daily
 - g) Checking all e-cig sources
 - h) Checking all alcohol sources
 - i) Checking all marijuana sources

11. If smoking a cigarette A81A is “1” (0 days), A84A (flavored) = missing.
12. If smoking a cigar A81B is “1” (0 days), A84B (flavored) = missing.
13. If using chewing tobacco A81C is “1” (0 days), A84C (flavored) = missing.
14. If using a hookah A81E is “1” (0 days), A84D (flavored) = missing.
15. If biological sex = female and respondent indicated they did not drink in the last 30 days or the last 12 months and A89 is missing (How many times have you had 4 or more drinks in the past 30 days), we converted A89 to 1 (0 days).
16. If biological sex = male and respondent indicated they did not drink in the last 30 days or the last 12 months and A90 is missing (How many times have you had 4 or more drinks in the past 30 days), we converted A90 to 1 (0 days).
17. If A85 (drinking an alcoholic beverage in the last 30 days) is missing, but A87 (drinking an alcoholic beverage in the last 12 months) is “0 days”, we converted A85 to 1 (“0 days”).
18. If grade = 8, a student should not have received questions A97D (driving a motor vehicle after using alcohol or other drugs). In the few cases this happened, the responses to this question by the 8th grade student was converted to missing.
19. Answering “No method was used” (A112A) and also marked one or more methods of birth control was used the last time they had sex (A112B-A112F) -- in these cases we converted the responses to A112A-F to missing.

Fixing Denominators

With the use of skip logic on the survey to decrease the time it takes to complete the survey (e.g., lead question: “Have you had dental problems during the past 12 months”, if the answer is “No”, the follow-up questions about dental problems are skipped), the percentages reported in the follow-up questions may be misleading if the first question is not taken into account. That is, if looking at the number/percentage of students with “decayed teeth or cavities” the resulting percentage reported without fixing the denominator may be alarming because the only students answering this question are the students who first reported they had dental problems in the past 12 months. To rectify this situation, we base the percentage on the whole population of students who answered the first question, if the student reported no dental problems, their response to “decayed teeth or cavities” would be counted as a “0” or “No” rather than missing.

Items with fixed denominators include (note, the “R” at the end of the variable name indicates the denominator was adjusted):

1. A40AR – toothaches or pain
2. A40BR – decayed teeth or cavities

3. A40CR – swollen, painful or bleeding gums
4. A40DR – could not eat certain foods because of a dental problem
5. A40ER – missed one or more days because of a dental problem
6. A36AR – A36IR – during a typical week, do you participate in the following activities... (A-I)
7. A78AR-A78ER – gambling activities (A-E)
8. A81AR-A81ER – using cigarettes, cigars, chewing tobacco, vape, hookah
9. A85R – drinking alcohol last 30 days
10. A87R – drinking alcohol last 12 months
11. A88R – how much do you drink at one time
12. A89R – 4 or more drinks in a row
13. A90R – 5 or more drinks in a row
14. A91R – how many days used marijuana
15. A108R – number of sexual partners

Creating New Variables

A number of variables were created based on the responses to certain questions to create a variable or a summary composite item. These include:

1. **COUNTYNAME** – based on the district information in MDE’s database and the county the district belongs to.
2. **Race** – Race exclusive groups based on responses to question A3A-A3G. This variable provides the number/percentage of students who marked only one group (indicated by “only” after the label, e.g., “Asian or Asian American only”). For those who marked more than one Race, they are classified as “Multiple Races (checked more than one)”.
3. **VapFlav** – If vaped in past 30 days, student used a flavored e-cigarette (based on items: A83B-A83I).
4. **SmokFlav** – if smoked a cigarette in past 30 days, student smoked a flavored (menthol) cigarette (based on item: A84A).
5. **CigarFlav** – if smoked cigar in past 30 days, student smoked a flavored cigar (based on item A84B).
6. **Slitflav** – if used smokeless tobacco in past 30 days, student used flavored smokeless tobacco (based on item A84C).

7. **HookahFlav** – if used hookah in past 30 days, student used flavored hookah (based on item A84D).
8. **Tobmonth** – use of conventional tobacco products (cigarettes, cigars, smokeless tobacco (based on items: A81A-A81C).
9. **Anytobmonth** – use of any tobacco products, including e-cigarettes and hookah (based on items: A81A-A81E).
10. **FlavTob** – Student used a flavored tobacco product in the past 30 days (based on new variables above: VapFlav, SmokFlav, CigarFlav, Slftlavm HookahFlav and Anytobmonth).
11. **Freqbingedrk** – frequent binge drinking, 5 or more drinks at a time on 10 or more occasions during the past year (based on items: A87 and A88).
12. **Binge** -- Binge drinking 4 or more drinks in a row (females) or 5 or more drinks in a row (males) within a couple of hours (based on items: biosex and A89 and A90).
13. **Combined_monthly** – bullied once or more in the last 30 days (based on items: A32A-A32E for Level 2 and 3 surveys; A32A, A32B, A32C, A32E for Level 1). Items greater than or equal to 2 for these items.
14. **Combined_weekly** – bullied weekly in the last 30 days (based on items: A32A-A32E for Level 2 and 3 surveys; A32A, A32B, A32C, A32E for Level 1). Items greater than or equal to 3 for these items.
15. **Combined_daily** – bullied daily in the last 30 days (based on items: A32A-A32E for Level 2 and 3 surveys; A32A, A32B, A32C, A32E for Level 1). Items equal to 5 for these items.
16. **Region** – 7-County metro area (Anoka, Carver, Dakota, Hennepin, Ramsey, Scott and Washington) vs. Greater Minnesota (other Minnesota Counties not listed above).
17. **FiveFV** – servings of fruits, fruit juice, and vegetables in the past 7 days (based on items: A52A, A52B, A52C).
18. **ACEsGrp** – ACEs score of 0 (none) to 4 (Four or more), using the same scoring method as past administrations. Based on items: (PRISON (A67), SXABUSE (A75/A76), A69A, A69B, A69C, A72, A73, A74), we count the number of ACES (0-8) and convert to the ACEsGrp short scale as noted above (0 to 4 or more).
19. **Weight** – based on height and weight variables and converted to a BMI score. The weight variable takes into account the BMI score, biosex, age, and grade. New in 2025 we included “underweight” as a category in addition to “overweight”, “obese”, and “normal” weight.
20. **Aduseany** – Any alcohol or drug use during the past year (excluding tobacco). Based on items: A122A, A122C, A122D-122P, A85, A87, A91.) Depending on the drug(s) selected by the student they can fall into 5 separate categories: 0 = No alcohol or marijuana or other drug use in the past year, 1= Used only alcohol, 2= Used alcohol and marijuana, but not other drugs, 3= Used marijuana or other drugs, but not alcohol, 4= Used alcohol and marijuana and other drugs in the past year.

21. **A122AR-A122PR** – this is the list of possible drugs used by students in the past year. Based on the programming of the question, students could select none of the drugs listed, some of the drugs listed, or all of the drugs listed. If they do select at least one drug, the program will code it as a “1” if selected and a “0” if not selected, however this only occurs if something was selected, if a student did not select anything (which is valid), all the drugs will be coded as “0”. The issue is if a student did not get this far into the survey, all of the responses will be coded as “0”, thereby producing misleading results and basing the percentage of drug use on all students, even those that did not get this far in the survey. Our solution to this is to assume if a student did not get to the substance use module of the survey, they can be counted as missing in A122A-P, but if they did get as far as the substance use module and did not select any of the drugs in A122A-A122P, we assume they did not use any of these drugs and therefore A-P should be coded as “0” or did not use. Items A122AR-A122PR reflects this coding and is how we reported percentages for A122A-P in the reports.