SMARVUS Data Processing Summary

The following steps were taken to clean, translate, and anonymise the dataset, as it is now presented on the OSF (see the codebook for detailed processing information about each variable):

Student Survey

1. Data from individual surveys was checked and cleaned.
    a. Key identifiers were added
    b. Maths education questions were renamed
    c. Eligibility question responses were adapted to represent respective ‘yes’ or ‘no’ responses
    d. Question numbers that appeared differently to the master survey were renamed for consistency
2. Data from the individual surveys was combined with other data from a given country and/or in a given language (depending on how many surveys there were in each category) and checked and cleaned again.
    a. Display order and any empty columns were removed
    b. Questions where responses options had a column per response (e.g., ethnicity and SpLDs) were merged
3. Data from the combined surveys were then combined into larger batches and checked and cleaned again. At this stage, some translation of free-text responses was also undertaken.
    a. Display order and any empty columns were removed (where missed in/different to 2a
    b. Questions where responses options had a column per response (e.g., ethnicity and SpLDs) were merged (where missed in/different to 2b)
    c. Most categorical responses were translated into English
4. All files were combined and final checks, cleaning, and final translation was done.
    a. Unique ID was generated
    b. All text was removed from the measurement scale responses so only numeric data remained
    c. Display order and any empty columns were removed (where missed in/different to 3a)
    d. Questions where responses options had a column per response (e.g., ethnicity and SpLDs) were merged (where missed in/different to 3b)
    e. Any translations missed in 3c were corrected
    f. University names were adapted to the English variants for consistency
    g. Free text responses (specifications where ‘other’ was chosen) were merged with the primary variable, as appropriate
    h. Free text responses were translated to English (using external spreadsheets that were read into R and re-joined with the data)
    i. Responses to the age variable were converted to age in years as required (e.g., some responses were dates of birth, or age in years and months)
    j. Variables were renamed to meaningful labels
    k. Variables were reordered

Instructor Survey

1. Qualtrics meta-data was removed
2. Corrections were applied as required (some researchers informed me of errors in their survey responses)
3. Incorrect responses were removed (e.g., those referring to more than one module, for modules prior to the student survey data collection period, or pertaining to postgraduate modules)

4. Empty responses were removed
5. Variables were renamed to meaningful labels

Grade Data

1. Each grade data file was wrangled into three columns, where possible – the student identifier (name and/or ID), grade (for the module students were taking at the time of data collection), and grading_scale
2. In some cases, there were grades for more than one module (i.e. previous statistics grades) so there were columns for each

Data Combining

1. Variables were created in the student survey data that represented the corresponding university name and statistics module name in the instructor survey
2. The instructor survey was joined to the student survey data using the university names and statistics module names
3. A variable was created in the student survey that represented the statistics module the grade in the grade data referred to (grade_module_name)
4. The grade data was joined to the already combined student and instructor surveys, using the student identifier (name and/or id) and grade_module_name (if necessary to extract the correct grade where there were multiple grades available)

Checking and Editing

1. Although checks were conducted throughout each stage of processing thus far, a further round of checks were completed
2. Any required edits were made (such as correcting typos and recoding miscoded variables

Data Paper Processing/Summaries

1. Duplicate entries were identified and removed
2. Empty cases were identified and removed
3. Summaries and supplemental materials were produced for the data paper at this stage

Anonymisation

1. The following variables were (re)categorised:
    a. degree_major
    b. degree_minor
    c. degree_level_maths_specify
    d. age
    e. spld
2. The ethnicity and related variables were removed
3. Free-text responses were checked for anonymity (e.g., those indicating mature or international student status) and redacted as required

Final Edits

1. Empty columns were removed (*n* = 170, all were from either the maths education or instructor survey questions)
2. Variables representing questions added by individual research teams were removed (e.g. ‘Is Dutch your first language?’)
3. Final renaming of variables (e.g., ‘stats_course’ to ‘stats_module’ for consistency with the data paper)
4. Final relocating of columns

Post Data Paper Edits

The following additions have been made since the publication of the data paper and are not in versions of the data or codebook prior to those with the suffix `_250124` (ddmmyy).

1. A Master's student manually cleaned the self-reported statistics grade data, which was manually checked by the project lead. The raw self-reported statistics grades are still available. The cleaned versions have the suffix `_edited`.
2. Where students provided a 0 for their grade, it was mostly unclear whether this represented a true zero grade (i.e., a fail) because some respondents used 0 to also represent NA. As such, we have recoded `0` as Zero so that researchers can handle these cases as they feel most appropriate.
3. Where students provided a grade that was inconsistent with either their other grades (e.g., suddenly specifying a `4` when all their other grades were in the 70s), with their classmates' grades (e.g., specifying a letter grade when the others gave numeric grades), their answers to other questions (e.g., they specified they hadn't taken statistics but gave a grade), or with the question (e.g., provided an answer that wasn't a single grade), these have been recorded as 'Inconsistent'.
4. Note that we only recoded variables as inconsistent with strong evidence. If the provided grade was a plausible outlier (i.e., it was at an extreme end of the scale and other students weren't reporting grades in the same range, but was technically possible), we did not recode it.
5. A new variable was created called `stats_edu_grade_data_scale` which contains the scales that the students reported their grades on. This did not always match the scales provided by their instructors. We suspect this ma, at least in part, be due to differences between the use/translations of the words 'grades' and 'marks' in different contexts and is something that should be kept in mind when using this data. Note that we only populated values in this variable where the grade data was present and usable.
6. Where a student reported grades that were clearly on a different scale to others on their course, we recorded their `stats_edu_grade_data_scale` as `Inconsistent`.