De-identification Procedures

In order to protect participants' personally identifiable information (PII) and adhere to applicable privacy regulations, we included removal of PII as one of our data processing phases. As part of this process, we removed PII from student assignments and associated file names. ETS follows the general definition of PII as any information that can identify a natural person. To that end, we removed students' names, email addresses, physical addresses, instructor names, university names, and other personal information that, in combination with other identifying information and depending on the particular context, may identify the student. We replaced them with non-real names, addresses, etc., as appropriate for the purpose and context. In some cases, PII was removed without replacement (e.g., when the PII was included in a document header for the instructor). Further details are provided below.

Step 1: Automated Header Detection and Removal

The first step in de-identifying the study writing data was to leverage an existing automated writing evaluation (AWE) engine to automatically detect and remove writing assignment headers, which frequently contain student names and instructor names.

Step 2: Replacing/Removing Participant Names and Contact Info

Next, three research assistants (RAs) examined the output from the automated de-identification round, which included new versions of the student texts (with the headings automatically removed) and a file containing the deleted heading text. During this phase, the RAs manually removed and/or replaced key identifiable information that went undetected in step one. PII indicating participants' real identities was *replaced* if it was embedded within the content of the essay and/or important to the genre (for example, in the case of a cover letter). Participant name and contact information replacements were made as follows:

PII	**Replacement**
Participant full name	Sam Doe
Participant first name	Sam
Participant last name	Doe
Phone number	(555) 555-5555
Email address	samdoe@uni.edu
Address	123 Main Street Anytown, NJ USA 12345
LinkedIn Profile URL	www.linkedin.com/in/sam-doe

Step 3: Removing additional autobiographical details

In a further effort to protect participant confidentiality, the RAs did an additional review of the data for other autobiographical details that could be deemed quasi-identifiers (i.e., information that in combination with other identifying details in the writing assignment, may identify the student). During this third round of de-identification, the RAs replaced the names of real people, places, and institutions connected to the student with pseudonyms or generic nouns (see the table below for examples).

Information	Replacement Protocol	Examples
Real people connected to the student • Includes: Names of family members, friends, classmates, instructors • Excludes: Public figures, authors' names, individuals quoted in cited sources	Replace with pseudonyms	• Person 1: Casey Smith • Person 2: Riley Johnson • Person 3: Jaime Miller • Person 4: Kerry Jones • Person 5: Skyler Williams • Person 6: Pat Brown • Person 7: Harper Miller • Person 8: Bailey Davis
Real cities connected to the student (e.g., hometown, city of residence)	Replace with pseudonyms	• City 1: Anytown • City 2: Anyville • City 3: Springfield • City 4: Franklin
Real counties connected to the student (e.g., high school named after county)	Replace with pseudonyms	• County 1: Washington County • County 2: Jefferson County
Institution	Replace with generic noun (and add the appropriate determiner, if needed)	• The university • My university • College • My school district
Course materials containing instructor's name	Replace with generic noun (and add the appropriate determiner, if needed)	• Class lecture • Course syllabus
Bibliographic information that may reveal participant's university	Replace with generic noun (and add the appropriate determiner, if needed) or pseudonym	• "Article Title." • http://removed-this-url.com/
Key dates (e.g., birthdates, dates of family member's deaths)	Replace with alternate (non-existent) date	• February 29, 2017

Step 4: CTRL-F Searches

As a final step in de-identification, we used the "CTRL-F" function to search the essay texts for participants' surnames. This involved concatenating the writing assignment texts into a single file, conducting CTRL-F searches for participants' last names, and reviewing the search results. If it was not possible to CTRL-F search for a student's last name (i.e., if the name corresponded to a common string that yielded the message "That shows up a lot!" in Microsoft Word), then separate searches were conducted with the student's first name and full name. In cases where student names generated search results, an RA reviewed the results to check that these names referred to public figures or fictional characters, rather than participants. If the names referred to actual participants or people that the participant knows, they were removed or replaced in accordance with the guidelines outlined above.