

Statement on the Protection of Privacy in UC StatFinder

Incorporated into the design of UC StatFinder are a number of features to protect applicant and student privacy while at the same time minimizing the loss of information that StatFinder provides to users. This document describes these features and explains how the threshold for aggregation of small cells was chosen.

Hardware/Software to Protect Student Data Records

Production servers containing the unit record data processed in UC StatFinder are protected by hardware and software “firewalls”. Extensive tests to penetrate these barriers have been conducted and adjustments have been made to correct known vulnerabilities.

Design Features to Protect Discovery of Small Cells

Federal and state laws--Family Educational Rights and Privacy Act (FERPA) and the California Information Privacy Act (IPA)--have established regulations for the protection of the privacy of student educational records. When student data are reported in aggregate (e.g., statistical data tables, charts and datasets), there is a potential for individual students to be identified when the number of students in the aggregate is small.

A number of design features in StatFinder prevent the discovery of small cells both in the Tables Library and in the StatFinder Query tool. These include the following:

- 1) Display of data elements are in bands where possible. SAT scores, high school GPAs, and parent income are displayed in “ranges”, i.e., “700-800”, “3.60-3.79”, “ and Less than \$40,000”, respectively, which prevents the identification of individual scores.
- 2) Data elements in complex tables may only be displayed in the same order. For example, a request for a breakdown of gender, ethnicity, and high school GPA, will always be displayed with the gender data element first, ethnic group second, and high school GPA third. This prevents discovery of small cells by recombining these data elements in different display orders.
- 3) Bands of data are combined or aggregated with other bands when cell sizes drop below a threshold of “5” per cell (see explanation of how “aggregation” works below).

How Aggregation Works

Aggregation was chosen over other methods of privacy protection (e.g., redaction or perturbation) because this methodology preserves to the greatest extent the original data.

Aggregation occurs when:

- 1) Cell sizes drop below the threshold level of “5” in any row displaying fall applicants, fall admits or fall enrollees. Aggregation will occur if any of these numbers are below the threshold, regardless of whether each of these three counts was requested by the user. A hypothetical example illustrates this process:

<u>American Indian</u>	<u>Fall Applicants</u>	<u>Fall Admits</u>	<u>Fall Enrollees</u>
SAT 200-499	20	10	2
SAT 500-599	50	35	20

In this example, the entire SAT 200-499 band would be aggregated with the SAT 500-599 band because the two American-Indian fall enrollees in the 200-499 SAT band fall below the threshold for reporting. Thus, the result would appear as:

<u>American Indian</u>	<u>Fall Applicants</u>	<u>Fall Admits</u>	<u>Fall Enrollees</u>
SAT 200-499			
SAT 500-599	70	45	22

- 2) The difference between the columns fall applicants and fall admits, or fall admits and fall enrollees, is less than the threshold. For example:

<u>African American</u>	<u>Fall Applicants</u>	<u>Fall Admits</u>	<u>Fall Enrollees</u>
SAT 600-699	100	80	25
SAT 700-800	40	39	12

In this example, the entire SAT 700-800 band is combined with the SAT 600-999 band because the DIFFERENCE between the number of fall applicants and fall admits–1– is less than the threshold for reporting. The bands are combined to protect the identity of the one African-American applicant in the SAT 700-800 band who was NOT admitted (i.e., the student was either denied admission or voluntarily withdrew the application).

<u>African American</u>	<u>Fall Applicants</u>	<u>Fall Admits</u>	<u>Fall Enrollees</u>
SAT 600-699			
SAT 700-800	140	119	37

- 3) In examining fall term applicant, fall term admit, and fall term enrollee counts, bands of data are aggregated based on the smallest cell or difference between columns that falls below the threshold for aggregation even if the user does NOT request that count in reports. So in the second example, if the user chooses NOT to select “Fall Admits” the aggregation of the two rows of data will still take place based on the assumption that all three counts were selected. This prevents users from recovering small cells by selecting different combinations of applicants, admits, or enrollees.

The rules governing how categories of variables are aggregated vary by type of data element:

- a) Data elements consisting of ordered categories (e.g., Parent Income Level, SAT Score Band). For these measures, categories are aggregated from lowest to highest ordered category. If aggregation is needed for the highest category, it is made with the next lowest category. Unknown-Missing is aggregated with the highest ordered category.

- b) Data element consisting of non-ordered categories (e.g., Ethnic Group, Gender, First Language Spoken). Rules vary by data element, with the goal of providing the most information (least amount of aggregation) to requestors.

How the Threshold Level of “5” for Aggregation was Chosen

The determination of the threshold level for the aggregation of small cells in StatFinder was based on the principle of *balancing* protection of student identities with the need to provide the most detailed information to StatFinder users. Moreover, the selection of threshold for aggregation must not make the individual identity of any student or applicant “easily traceable”, as prohibited by FERPA, IPA, and university policy.

The University of California’s draft report, *Guidelines for Release of Aggregated Student Data with Small Cell Sizes*, states:

The University of California considers tables containing aggregated data entries of fewer than 10 individuals to create a situation where a student could be personally identified. Therefore data tables and reports containing cell sizes of fewer than 10 individuals should be carefully reviewed before release to ensure that the identities of students are not easily traceable and create a potential for the invasion of the student’s privacy. In certain cases (based on the audience and use of the data), the release of reports containing cell sizes of fewer than 10 individuals may be defensible.

Given the design features of StatFinder to protect privacy enumerated above, and that StatFinder programming also allows for easy adjustment of the threshold level for aggregation, a series of studies were commissioned to examine the amount of data loss due to aggregation versus the ability to recover small cells for the recommended threshold of “10” and competing lower thresholds. These studies, *UC StatFinder Privacy Reports*, conducted by StatFinder contractor MPR Associates, examined the balance between data loss to aggregation at threshold levels of 10, 5, and 3 versus the ability to recover small cells at threshold levels of 10, 5, 4, and 3.

The statistics with respect to “data loss” due to aggregation in the first *UC StatFinder Privacy Report* show that at a threshold for aggregation of “10” about three-quarters of the rows in complex 3-data element tables would be aggregated, compared to about two-thirds of the cells for a threshold of “5” and just over fifty percent for a threshold of “3”. In 2-data element tables, these percentages were 44% aggregation for threshold “10”, about 33% for threshold “5” and about 24% for threshold “3”. Only in the simplest 1-data element tables was aggregation reduced to under 13% at all threshold levels. Differences in amount of data loss did not vary much by size of campus, though notably, at the UC system’s smallest campus, UC Merced, the amount of aggregation in a 3-data element table reached 83%.

Ability to recover small cells in the first *UC StatFinder Privacy Report* was measured in terms of number of cells recovered and the time-to-recover those cells. The number of small cells recovered and the time to recover these small cells (using the same complex 3-data element tables), did not vary very much between threshold levels of “10”, “5”, “4”, and “3”. However, the sensitivity of the data recovered varied markedly. At the threshold level of “3”, notably, all cells recovered specified the campus of enrollment, including one recovered cell with a cell size of “1”. At threshold levels of “4” and above the cells

recovered either did not identify a campus of enrollment (“university wide data”) or, in the case of threshold “10” consisted of recovering cells with counts of “7”, “9”, or “4” students. No cell sizes of “1” were recovered for threshold levels of “4” and above, save for one cell at the threshold level of “5” for an enrolled student of *unknown race, unknown gender, and no campus specified*. A second *StatFinder Privacy* report conducted using additional years of data found that a few small, identifiable cells could be uncovered at a threshold level of “4”, though only with great effort (6-8 hours of work by an experienced user paid to try and identify these small cells).

While the combinations of data elements tested in the *UC Privacy Report* do not represent the universe of possible combinations of data elements in StatFinder, given the other protections inherent in the system (e.g., banding, fixed order for display of variables in complex tables, and aggregation of implied columns), it was felt, based on study results, that student privacy could still be protected, and data loss minimized, if a threshold of “5” for aggregation were selected.

A review by UC University Counsel of this information concurs with the assessment that a threshold of “5” in the context of the StatFinder system does not make the identity of any student or applicant “easily traceable”¹ (see the letter from University Counsel, *StatFinder and Public Data Display and Privacy Issues*).

¹ Please note: It has also been determined that comparison of data tables generated for separate “populations” in StatFinder, specifically, “All Students” vs. “CA Residents”, or “All Transfers” vs. “California Community College Transfers”, could generate combinations of data that have small cell sizes but that the identities of students in these small cells would NOT be “easily traceable”.