NATIONAL STUDENT CLEARINGHOUSE®
RESEARCH CENTER™

Gender Imputation

March 7, 2013

NATIONAL STUDENT CLEARINGHOUSE RESEARCH CENTER

2300 Dulles Station Blvd., Suite 300, Herndon, VA 20171

Contents

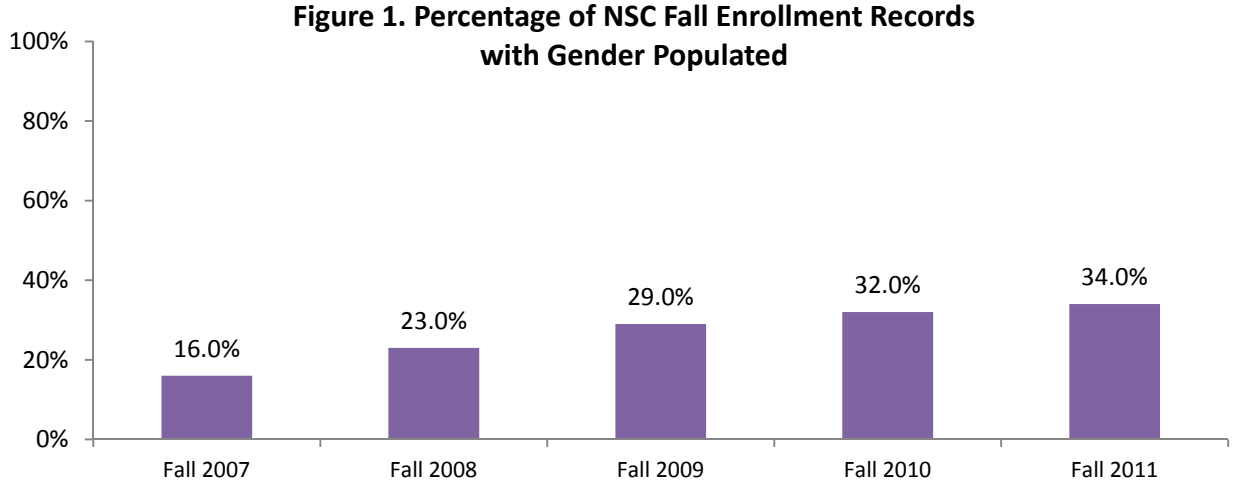
This report was supported by a grant from the Lumina Foundation.

Lumina Foundation, an Indianapolis-based private foundation, is committed to enrolling and graduating more students from college — especially 21st century students: low-income students, students of color, first-generation students and adult learners. Lumina's goal is to increase the percentage of Americans who hold high-quality degrees and credentials to 60 percent by 2025. Lumina pursues this goal in three ways: by identifying and supporting effective practice, through public policy advocacy, and by using our communications and convening power to build public will for change. For more information, log on to www.luminafoundation.org.

Introduction

In late 2007, the National Student Clearinghouse (NSC) expanded its Enrollment Reporting service to include several additional data elements (commonly referred to as the "A2" or "expanded" data elements). One of these expanded data elements is student gender. Although gender is potentially important to a number of research projects, it is currently an optional data element in Clearinghouse Enrollment Reporting. Consequently, postsecondary institutions have populated this field in a limited percentage of enrollment records. However, as Figure 1 illustrates, this percentage has been growing steadily since 2007.

Figure 1. Percentage of NSC Fall Enrollment Records with Gender Populated



While the NSC Research Center expects this trend to continue, in order to make *immediate* use of the gender data element for research, it is necessary to impute its value for the majority of Clearinghouse enrollment records. To meet this need, the Research Center developed an imputation process in which previously submitted name-gender pairs are used to determine the probability of any first name being associated with either gender. To increase the accuracy of the imputation process, it also draws on name-gender data from the Social Security Administration and the U.S. Census Bureau. The Research Center's report on Fall 2012 Current Term Enrollment Estimates, released in December 2012, marked the first application of this gender imputation process in a Clearinghouse research project.

Imputation Method

Data Sources

The imputation is based on data from three sources:

1) National Student Clearinghouse: As of June 2012, approximately 70 million submissions of enrollment records to the Clearinghouse included gender. These submissions comprise over 14 million students and nearly 600,000 unique first names.

2) Social Security Administration: The SSA provides national data on the gender frequencies of first names. The SSA dataset consists of approximately 90,000 unique first names from over 320 million U.S. births between 1880 and 2010. The listing is restricted to names with at least five occurrences.

3) U.S. Census Bureau: The Census Bureau also provides gender frequencies for first names. The Census dataset contains a little over 5,000 names obtained from the decennial census of 1990.

Table 1 shows summary characteristics of all three data sources. Clearinghouse data contain many more unique first names than the other sources, owing to the fact that the Clearinghouse collects transactional data, which sometimes contains nicknames, typos, abbreviations, and other idiosyncrasies. International students may also contribute to the large variety of first names submitted to the Clearinghouse. Because of this variety, many first names occur only once or twice in Clearinghouse data, a factor that must be taken into account when assessing the confidence level of name-gender associations. The depth of the SSA and Census data provide balance for the breadth of names observed in the Clearinghouse data.

Table 1: Characteristics of NSC, SSA, and Census Data

	NSC	SSA	Census
Number of unique individuals in dataset	14,065,847	322,402,727	5,569,651
Number of unique names in dataset	583,654	88,496	5,163
Average number of individuals per name	24	3,643	1,079
Names exclusively associated with females	355,747 (61%)	51,754 (58%)	3,944 (76%)
Names exclusively associated with males	183,318 (31%)	27,090 (31%)	888 (17%)
Names associated with more than one gender	44,589 (8%)	9,652 (11%)	331 (6%)

Assessment of Various Rule Sets

For each data source (SSA, Census, and Clearinghouse) an imputation subset was created by determining how frequently a first name occurred in combination with either gender. The first name was then assigned to the gender with which it occurred more frequently (if the count of female submissions was equal to that of male submissions, the name was assigned to the female gender). For a name-gender pair to be utilized in the imputation process, the name-gender assignment had to be consistent across every imputation subset in which it occurred. In this way, the SSA and Census data served to increase the confidence level of the imputations.

The actual imputation dataset represented the union of the imputation subsets constructed from each data source. Different rule sets (providing varied levels of confidence) produced imputation datasets of different sizes. Larger imputation datasets made it possible to impute gender for more records, but with less confidence. Table 2 shows results for the most extreme rule sets, along with results for the rule set that was ultimately selected. In weighing the results, it was determined that the optimal balance of quantity and confidence was achieved by limiting the imputation dataset to names that occurred at least two times within a single data source, and that occurred in combination with the same gender 95% of the time.[1]

Table 2: Results of various imputation rule sets

	Fall 2007	Fall 2008	Fall 2009	Fall 2010	Fall 2011
No Imputation					
Percentage of fall enrollments with gender populated	16%	23%	29%	32%	34%
Male-to-female ratio	0.73	0.72	0.73	0.74	0.74
Highly Aggressive *(no restrictions; imputation based only on Clearinghouse data)* *582,771 name-gender pairs in resulting imputation dataset*					
Percentage of fall enrollments with gender after imputation	98%	98%	98%	98%	98%
Male-to-female ratio	0.74	0.74	0.75	0.75	0.75
Highly Conservative *(100% name-gender consistency within a data source; at least 5 occurrences of name within a data source; name-gender consistency across all data sources in which name occurs)* *43,033 name-gender pairs in resulting imputation dataset*					
Percentage of fall enrollments with gender after imputation	20%	27%	33%	36%	38%
Male-to-female ratio	0.61	0.64	0.66	0.67	0.68
Optimal *(95% name-gender consistency within a data source; at least 2 occurrences of name within a data source; name-gender consistency across all data sources in which name occurs)* *157,785 name-gender pairs in resulting imputation dataset*					
Percentage of fall enrollments with gender after imputation	91%	91%	91%	92%	91%
Male-to-female ratio	0.74	0.74	0.75	0.75	0.75

[1] The 95% rule only had to be met for at least one of the data sources in which the name occurred. However, the name had to be assigned to the same gender across every data source in which it occurred.

Validation

The use of SSA and Census data provided partial external validation. As described previously, the SSA and Census datasets were used in two ways: to ensure that name-gender pairs were consistent across every dataset in which they occurred, and to enhance the imputation process by contributing name-gender pairs that did not occur in Clearinghouse data.

Further validation was provided by comparing the gender distributions from the Fall 2012 Current Term Enrollment Estimates, the first Clearinghouse report to make use of the imputation process, to the most recent postsecondary student gender distributions available from NCES (Fall 2011). As Table 3 indicates, the distributions were very similar at the national level and within institutional sectors.

Table 3: Gender distributions from NCES Digest of Education Statistics and NSC Current Term Enrollment Estimates

Sector	Gender	Digest Of Education Statistics, Fall 2011 Enrollments	NSC Current Term Enrollment Estimates, Fall 2012
All Institutional Sectors	men	43%	43%
	women	57%	57%
Four-year, public	men	45%	45%
	women	55%	55%
Four-year, Private	men	42%	42%
	women	58%	58%
Four-year, for-profit	men	36%	35%
	women	64%	65%
Two-year, public	men	43%	42%
	women	57%	58%

Finally, Table 4 provides summary statistics on gender imputation from the Fall 2012 Current Term Enrollment Report. Across all four years included in the report, gender was provided by the schools in 32% of enrollment records. After imputation, 91% of the records included a value for gender.

Table 4: Summary statistics on gender imputation from Fall 2012 Current Term Enrollment Estimates

	Fall 2009	Fall 2010	Fall 2011	Fall 2012
Count of Enrollment Records	18,456,177	19,050,433	19,298,378	19,044,344
% of records with school-provided gender	27.3%	31.3%	34.6%	35.8%
Frequency of agreement between school-provided and imputed gender	99.6%	99.6%	99.5%	99.5%
% of records where imputation could not be applied	11.7%	12.1%	12.7%	13.3%
% of records with gender after imputation (either school-provided or imputed)	91.3%	91.5%	91.5%	91.3%

References

National Center for Education Statistics. Digest of Education Statistics, 2012. Retrieved on 2/11/2013 from http://nces.ed.gov/programs/digest/index.asp

National Student Clearinghouse Research Center. Fall 2012 Current Term Enrollment Estimates. Retrieved on 3/7/2013 from http://research.studentclearinghouse.org/files/TermEnrollmentEstimate-Fall2012.pdf

Social Security Administration. Beyond the Top 1000 Names. Retrieved on 5/1/2012 from http://www.ssa.gov/OACT/babynames/limits.html

United States Census Bureau. Genealogy – Name Files. Retrieved on 5/1/2012 from http://www.census.gov/genealogy/names/names_files.html