Demographic Aspects of Surnames from Census 2000

David L. Word, Charles D. Coleman, Robert Nunziata and Robert Kominski

ACKNOWLEDGEMENTS

We would like to thank Peter Morrison of the RAND Corporation for initial encouragement to work on this project and for his comments; Signe Wetrogan, John Long, and Nancy Gordon for enabling this work; Maureen Lynch, Bert Kestenbaum (Social Security Administration), James Farber, and Matthew Falkenstein for providing data; Emmett Spiers for help on modifying the Lynch-Winkler string comparator program to enable Edit #2; Susan Love for providing the definition of data-defined person; Rodger Johnson, Campbell Gibson and Frank Hobbs for demographic review; Robert Fay for comments, information, and revisions; Gregg Robinson for comments; and Marjorie Hanson for editorial review of this report.

1. INTRODUCTION

A person's name is one of the most basic pieces of information that describes them. Moreso than a person's race, sex or age, we most often recognize people by their name. But names are not divorced from other aspects of an individual. Often, by knowing a name, we can infer many other things about the person. Names also have a historical context, ebbing and increasing over time with changes in popular culture.

This report documents both the overall frequency of surnames (last names), as well as some of the basic demographic characteristics that are associated with surnames. **The presentation of data in this report focuses on summarized aggregates of counts and characteristics associated with surnames, and, as such, do not in any way identify any specific individuals.**

The data for this project were taken from records from the 2000 decennial census of population. The primary purposes of the U.S. decennial census of population are to provide data with geographic detail on the population for use in reapportionment and redistricting, and administering governmental programs. However, for decades, decennial census data have been used by government agencies, researchers, academicians, businesses, the news media, and many others to describe and understand demographic trends and patterns in the U.S. population.

In releasing any data or information from the decennial census, the U.S. Census Bureau has a legal obligation under Title 13 of the U.S. Code to protect the confidentiality of individuals' information. In this regard, individual questionnaires of any specific census, (generally of interest for genealogical and historical research), are not released by the National Archives until 72 years after that specific census has been taken. Additionally, no public-use microdata files of any type contain name information

This report has been undertaken to provide a better understanding of the overall distribution of surnames in the population, and to provide some idea of the relationship between surnames and basic demographic characteristics such as gender, race and ethnicity. Even in this highly aggregate form, this information may be helpful in genealogical, marketing, and cultural research, as well as a variety of other applications. As such, it is useful information in helping to understand the ever-changing nature of the cultural mosaic that helps to define our nation.

2. THE BASE DATA

While Census 2000 is the first decennial census that permits examining demographic detail with names, this report is by no means the first to present tabulations of names. The Social Security Administration has published counts of frequently occurring surnames numerous times (SSA, 1957, 1964, 1975, 1985). Their tabulations consist of surnames of all people who had obtained Social Security Numbers as of the dates of these reports. The number of distinct surnames reported have ranged from about 1,500 (SSA, 1957) to over 8,000 (SSA, 1985). These names, however, have been limited to six characters. Six characters are certainly sufficient to uniquely identify shorter names like SMITH, BROWN and JONES. On the other hand, a name such as MARTIN could be MARTIN, or, it could be something like MARTINI, or MARTINEZ. The Social Security Administration has had ongoing data releases on the first names of newborns for each year since 1990 (SSA, 2003). SSA's first compilation of newborns first names was released in Shackleford (1998). These data, however, lack race and ethnicity information and are limited to the 1,000 most frequent male first names and the 1,000 most frequent female first names.

In July 1995, the Census Bureau placed summary information on male and female first names and last names on its website (Census Bureau, 1995). The data released in 1995 were created from a sample of 7.2 million census records (about 3 percent of the population) developed as part of the 1990 Post-Enumeration Survey (PES) operation, following the 1990 decennial census. Word and Perkins (1996) have used these same data to develop a Spanish surname list, also available from the Census Bureau

This report uses name responses from almost 270 million people with valid name information in Census 2000. As part of the Census 2000 form, individuals were asked to print their name, as well as the names of all other persons enumerated at a given address. All information on the census forms, including written information such as names, was captured in an optical scanning process conducted at four census processing centers around the country. After scanning, the original forms were shredded and destroyed. The scanned forms were then converted into strings of characters data, using optical character recognition software (OCR). These strings of characters become the base data for use in this report. More discussion about the process of converting the written-in names to data, including the assumptions used to define and edit names, will be discussed in the section, "Methodology of Measuring Names".

3. CHARACTERISTICS OF SURNAMES

3.1 How many names are there?

Even after applying various edits and acceptance criteria to the names, there are a sizable number of unique names in the population. Over 6 million last names were identified. Many of these names were either unique (occurred once) or nearly so (occurred 2-4 times) raising questions about the actual validity of the name. Cursory examination of the data indicates that many of these unique names were probably the entire name of the person (first and last, or first, middle initial and last) concatenated into a single continuous string, with some other information. At this time, it is not possible to easily break a fully concatenated name back into its' constituent parts. Doing so, however, would have reduced the counts of unique names sizably, while only slightly increasing the numbers of person with more common names. While a relatively large proportion of all names relate to only one person or a few people, a large proportion of the entire population can be identified with a relatively small proportion of all names. Table 1 better explains this phenomenon.

Table 1 shows the frequency of last names and the numbers of people who are defined by them. Seven last names are held by a million or more people. The most common last name reported was SMITH, held by about 2.3 million people, or about .9 percent of the population. Another 6 names with over a million respondents (JOHNSON, WILLIAMS, BROWN, JONES, MILLER and DAVIS), along with SMITH, account for about 4 percent of the population, or one in every 25 people. There are another 268 last names each occurring at least 100,000 times, but less than 1 million times. Together, these 275 last names, just 4/100,000 of all reported last names, account together for 26 percent of the population, or about one of every four people. On the flip side of this distribution, about 65 percent (or 4 million) of all captured last names were held by just one person, and about 80 percent (or 5 million) were held by no more than 4 people.

Table 1

Last Names by Frequency of Occurrence and Number of People: 2000

Last Names				People with these Names		
Frequency of Occurrence	Number	Cumulative Number	Cumulative Proportion (percent)	Number	Cumulative Number	Cumulative Proportion (percent)
1,000,000+	7	7	0.0	10,710,446	10,710,446	4.0
100,000-999,999	268	275	0.0	60,091,601	70,802,047	26.2
10,000-99,999	3,012	3,287	0.1	77,657,334	148,459,381	55.0
1,000-9,999	20,369	23,656	0.4	58,264,607	206,723,988	76.6
100-999	128,015	151,671	2.4	35,397,085	242,121,073	89.8
50-99	105,609	257,280	4.1	7,358,924	249,479,997	92.5
25-49	166,059	423,339	6.8	5,772,510	255,252,507	94.6
10-24	331,518	754,857	12.1	5,092,320	260,344,827	96.5
5-9	395,600	1,150,457	18.4	2,568,209	262,913,036	97.5
2-4	1,056,992	2,207,449	35.3	2,808,085	265,721,121	98.5
1	4,040,966	6,248,415	100.0	4,040,966	269,762,087	100.0

3.2 Characteristics of surnames

Table A-1 shows the distribution of the top 50 last names in terms of numeric count, crosstabulated by Race/Hispanic origin. As Section 4.4.7 explains, race data in this analysis is constructed so that any person identified as Hispanic is placed in that classification, regardless of reported race. As such, race identification is used only for those persons who are not Hispanic.

As can be seen, many surnames have race/Hispanic distributions that appear to be quite distinct from the race/Hispanic distribution of the population as a whole. Especially in the case of the Hispanic population, which constitutes about 12 percent of the overall population in this study, it is clear that there are names which might be characterized as strongly "Hispanic" last names. In Table A1 this includes such names as GARCIA, RODRIQUEZ, MARTINEZ, HERNANDEZ, LOPEZ, GONZALEZ, and several others. Each of these surnames have race/Hispanic proportions which are over 90 percent Hispanic.

While other surnames have strong associations with specific race groups, none show the kind of strength in association as with these Hispanic-related names. The name MILLER, for example belongs about 86 percent of the time to persons classified as White, while Whites make up about 70 percent of this population. BAKER is another

surname with a higher-than average percentage of White ownership, at 82 percent. Among Black persons there appear to be high-than-expected occurrences for names such as WILLIAMS, JACKSON, HARRIS AND ROBINSON, for example.

Large differentials for persons in the race categories of American Indian/Alaskan Native, Asian/Pacific Islander and persons choosing two or more races, are less clear in the short list of the fifty highest occurring last names. For this reason, the list of the 1000 most frequently occurring last names was examined with a view toward identifying those last names that are held by the highest concentration of a single race/Hispanic group.

Table 2 shows, for each race/Hispanic group, the ten last names with the highest relative concentration for that group. Included in this table is the name, the overall rank of that name out of the top 1000 last names, the total number of persons with that last name, its frequency per 100,000 people in the population, and the percentage of people holding that name that occupy the race/Hispanic group in which it is shown.

Table 2. Last names with greatest likelihood by race and Hispanic origin groups

NAME	RANK	COUNT	per 100K	% in this group
WHITE				
YODER	707	44245	16.4	98.1
KRUEGER	863	36694	13.6	97.1
MUELLER	467	64305	23.8	97.0
KOCH	657	47286	17.5	96.9
SCHWARTZ	330	84699	31.4	96.8
SCHMITT	898	35326	13.1	96.8
NOVAK	899	35282	13.1	96.8
SCHNEIDER	272	100553	37.3	96.7
SCHROEDER	450	66412	24.6	96.7
HAAS	941	34032	12.6	96.7
BLACK				
WASHINGTON	138	163036	60.4	89.9
JEFFERSON	594	51361	19.0	75.2
BOOKER	902	35101	13.0	65.6
BANKS	278	99294	36.8	54.2
JACKSON	18	666125	246.9	53.0
MOSLEY	699	44698	16.6	52.8
DORSEY	763	41104	15.2	51.8
GAINES	739	42369	15.7	50.3
RIVERS	879	35980	13.3	50.2
JOSEPH	356	80030	29.7	48.8
API				
ZHANG	963	33202	12.3	98.2
HUANG	697	44715	16.6	96.8
CHOI	872	36390	13.5	96.5
LI	519	57786	21.4	96.4
HUYNH	790	40011	14.8	96.2
YU	874	36285	13.5	96.2
NGUYEN	57	310125	115.0	95.9
PHAM	498	59949	22.2	95.9
WU	683	45815	17.0	95.9
TRAN	188	136095	50.5	95.6

NAME	RANK	COUNT	per 100K	% in this group
AIAN				
LOWERY	752	41670	15.4	4.4
HUNT	157	151986	56.3	3.9
SAMPSON	844	37234	13.8	3.8
JACOBS	233	115540	42.8	3.7
LUCERO	945	33922	12.6	3.1
MOSES	858	36814	13.6	2.9
BIRD	944	33962	12.6	2.6
JAMES	80	233224	86.5	2.5
ASHLEY	852	37021	13.7	2.4
PROCTOR	918	34682	12.9	2.3
TWO OR MORE RACES				
ALI	876	36079	13.4	17.5
KHAN	665	46713	17.3	15.6
SINGH	396	72642	26.9	15.3
SHAH	831	37833	14.0	5.9
PATEL	172	145066	53.8	5.8
JOSEPH	356	80030	29.7	5.3
COSTA	900	35227	13.1	5.2
ANDRADE	666	46702	17.3	5.0
SILVA	214	126164	46.8	4.8
VANG	982	32333	12.0	4.8
HISPANIC				
BARAJAS	989	32147	11.9	96.0
OROZCO	690	45289	16.8	95.1
ZAVALA	938	34068	12.6	95.1
VELAZQUEZ	789	40030	14.8	94.9
IBARRA	662	46895	17.4	94.7
JUAREZ	429	68785	25.5	94.7
MEZA	835	37662	14.0	94.7
HUERTA	959	33348	12.4	94.6
CERVANTES	520	57685	21.4	94.5
VAZQUEZ	328	84926	31.5	94.5

Last names which are 'dominated' by a single race/Hispanic group are not necessarily names which occur in high relative frequency in the population. Note that many of the names shown in Table 2 rank near the lower end of the 1000-name list. However, there are exceptions. JACKSON, which is held 53 percent of the time by Blacks, is the 18th most common last name. NGUYEN, held 96 percent of the time by Asian/Pacific islanders, is the 57th most frequent name.

Some groups, notably, Whites, Asian/Pacific Islanders, and Hispanics, demonstrate very high 'ownership' of some names – at levels exceeding ninety percent. These situations constitute places where a given name can, with high certainty, be assumed to be held by a person of a single specific race. Note that the relative distribution of the group in the overall population is somewhat irrelevant to this – Whites constitute 70 percent of the overall population in these data, and Hispanics are 12 percent, but Asian/Pacific Islanders are only 3.7 percent, yet dominate the ten names most singly-associated with their race group.

4. METHODOLOGY OF MEASURING NAMES

4.1 Turning names into data

Turning a written name on a census form into usable data for tabulation purposes is a task which involves a number of assumptions and decision rules. This section describes both the operational and logical decision rules used to turn 'names into data'.

As was discussed earlier, the data for this research comes from the written-in names persons provided when they filled out their census form. In most cases, answers to questions in the census are made by marking an appropriate box among a list of answers. For example, to designate one's sex, the respondent chooses from two boxes, one labeled 'male', the other labeled 'female', and marks the box that best describes them. In the process of transforming the marks into data, the mark is 'read' in the scanning process, and is then assigned a value by the OCR software (such as 1 for male, 2 for female).

Some items, such as race and relationship, allow a respondent to either mark a box, or if they feel no box is appropriate, to write in a response. Other items, such as language, are write-in only. In cases such as these, all write-ins must eventually either be assigned a numeric value, or excluded as an unacceptable or inappropriate response In the case of language, for example, the Census Bureau codes about 360 unique languages, each with its own numeric coding value (so, for example, language code # 625 corresponds to the language "Spanish"). Obviously, many factors enter into the process of turning a written response on paper into a numerically–coded value. For instance, alternate spellings, including incorrect spellings, random marks on the paper, abbreviations, etc., must all have decision rules associated with them, in order to code them. Doing so makes it possible to have a reasonable number of coded languages: while the Census has 380

distinct language codes, it received well over 100,000 unique character strings or "words" that people wrote down as their language.

Transformation of names into data follows much the same route, but is much more complex, engaging a larger set of rules and procedures. In addition to the possible sources of error already mentioned, a number of other issues relevant to the character strings defining names come into play. This includes things such as: 1) scanning or reading errors by the OCR software, 2) mis-keying, 3) respondents entering data into incorrect locations in questionnaires, 4) respondents entering no name or an invalid name, and 5) concatenation of multiple name parts (e.g, first and initial) when they are written in the space for a single name. Each of these problems must be addressed either with some kind of editing or resolution rule, or the name must be left as is. It is due to many of these issues that there are a large number of 'names' which occur only once or twice – many of these 'unique' names are variants of more common names which short of inspection on a one-by-one basis, cannot be "corrected" to the character string they actually are supposed to represent.

4.2 Definition of a name

For purposes of these tabulations, a captured name from Census 2000 is considered to be "valid" if it satisfies the following two criteria:

1. Both the first and last names must have at least two alphabetic characters.

2. A first or last name may also be considered valid if it has support in the Social Security Administration's 1998 NUMIDENT file. The NUMIDENT file used in this research is a 5 percent sample of all people who had been issued Social Security numbers as of November 1998. The advantage of using the NUMIDENT file as a benchmark is that individuals receive Social Security cards with the names they provide on their applications. If an individual receives a card with an error in their name, they have an incentive to report the error and have it corrected. Thus, it is assumed that the NUMIDENT file is the most current and correct source for validating names. The NUMIDENT file contains the most recent name (first and last) and demographic characteristics (sex, date of birth, and race) of each individual on it, as of the date of its production. The NUMIDENT file was not screened to eliminate deceased individuals nor those living outside the United States in November 1998. Twenty million people are represented in the sample, about one-fourteenth of the number of people enumerated in Census 2000. Thus, on average, the ratio of the Census 2000 count for a name (last, male first, or female first) to the NUMIDENT sample count should be about 14 to 1. Ratios far larger than 14 to 1 indicate possible "invalid" names. In this project, a census-captured name with a ratio of approximately 50 or more to 1 triggered an investigation into its validity and whether it should be changed or deleted.

Records with names that did not satisfy the two criteria were deleted for purposes of these tabulations. For example, names such as A LINCOLN or MISTER T would be

ruled invalid because neither the first name A nor the last name T contains two letters, thus violating criterion 1. Another example of an invalid name under criterion 1 would include a string like PETERJDAVIS as a first name, along with a blank last name. In this case, while it is possible that the correct name is PETER J DAVIS, the action of the respondent in writing the entire name in the space allotted for the first name only means there is no name in the space for the last name, thus violating criterion 1.

Some persons originally had blanks in either or both of their name fields, and thus failed criterion 1. Titles such as JR, SR and III were removed from the last name field, but not from the first name field. Intervening blanks and hyphens were also deleted. So, for example, the last names of LOPEZ-GARCIA, LOPEZ GARCIA and LOPEZGARCIA were all retained as LOPEZGARCIA. A similar example for the first name MARY ANN would show that it appears in the file as the concatenated entry MARYANN.

Examples of invalid names under criterion 2 are strings such as PERSON (as a first name), ADULT, BABY, HOUSEHOLDER, or SPOUSE. Thus, a census record with the name BABY MILLER would have been dropped from the analysis, because BABY is not supported as a valid name in the NUMIDENT file. In general, criterion 2 was applied independently to first names and last names. However, the "complete names" (that is, first and last name in combination) of JOHN DOE and JANE DOE were ruled invalid because the proportion of the people who reported the last name DOE and who also reported a first name of JOHN or JANE is far higher in the Census records file than in the NUMIDENT file. Other possibly invalid names like ELMER FUDD and MICKEY MOUSE remain on the file because we did not subject all complete names to the process used for JOHN and JANE DOE.

As names become less frequent, the possibility that a string of letters is not a valid name increases. As has been noted, this is due to many factors – misplaced or mis-scanned letter(s), bad spelling, and a variety of other causes. All names occurring 300 or more times were reviewed for validity, using a series of rules described in section 4.4.5, although certainly this process did not delete or modify all invalid names. However, names occurring fewer than 300 times were not examined at all, because of their large relative volume.

The initial data file used for this report contains 279,132,770 data-defined person records (census records with at least two data fields with valid responses) from an intermediate file created during the processing for Census 2000. After applying the criteria and the edits we developed for improving names, the final number of records for analysis comprises 269,762,087 people, or approximately 96 percent of all people counted in Census 2000.

4.3 Editing names

A variety of edits were developed to improve the quality of the name strings. Simply summarized, the edits attempted to identify and resolve a series of basic problems in name strings. These edits include:

1) elimination of some characters to yield a cleaner name
2) resolving similar but inconsistent last names within households
3) switching transposed first and last names

Not every possible edit that might be imagined to 'clean up' names was developed. Two examples of edits not made are noted here. One edit not developed would 'break apart' those names entered by respondents as single string. It is not uncommon for people to write in their full name in the last name space, realize they have made an error, and then write in their first name again in the first name space. Based on visual inspection of subsets of single-occurrence names, we believe this error may account for the large number of "unique" names. An edit to break apart compound names was considered too intensive a task for this research, and since compound names only account for a small fraction of the total population, it is unlikely they would change the overall distribution of names in the population in any substantively significant way.

The list of edits undertaken and resultant impact on data records is detailed in Appendix A.

4.4 Edits

4.4.1 The preliminary edit

The preliminary edit performed several minor operations to produce cleaner names. It implemented the Criterion 1 rule that both the first and last names of a respondent needed to contain at least two letters for retention. Titles, such as JR, SR, III, and IV, which were either separated from first names by spaces or erroneously concatenated, were removed. JOHN DOE and JANE DOE records were deleted. Names with embedded numerals (e.g., HEN6RY) had the numerals deleted. Finally, CHRISTOPHEJR and CHRISTOPHESR were edited to CHRISTOPHER. (First names are limited to 13 captured characters. The problem – truncation by the respondent in order to fit in a "JR/SR" title -- occurred to the name CHRISTOPHER with great frequency, nearly 15,000 times.)

The count of first names edited or deleted was 5,637,813. The vast majority of these changes were the removal of titles

4.4.2 Edit 1: Removing dangling initials

As noted above, the modifications applied to the basic file removed most embedded blanks. Because many middle initials were entered in either the first or last name field and not in the box marked "M.I.," the preliminary edit deleted the intervening blanks, causing these middle initials to be concatenated to names. In order to trim the dangling initials, the ratio of the name as captured (e.g., JOHNL) to its stem name (e.g., JOHN) was computed and compared to the similar ratio from the NUMIDENT file. For each name type (last, male first, or female first), a threshold value for each letter appended to the stem name was created. If a name in the census file had fewer than a prespecified threshold value of occurrences (generally in the range of 1 to 2 per thousand), the name was contracted back to the stem name. For example, JOHNL, SMITHB, and JENNIFERG were edited to JOHN, SMITH, and JENNIFER, respectively, while CAROL(E), ROBERT(O), and BROWN(E) are retained as CAROLE, ROBERTO, and BROWNE, respectively. In other words, when the ratio of the count of a name with an appended initial to its stem is small, it is assumed that the appended name is the actual name.

Dangling initials were eliminated from 2,226,434 last names.

4.4.3 Edit 2: Making last names agree within households

While we recognize that Mr. SMITH and Ms. JOHNSON may reside in the same household, generally within many households one would expect most of the last names to agree. However, when Mr. SMITH and Ms. SNITH share a household, it is likely that one of the two last names was captured incorrectly. Working on the assumption that people within households share last names, edit 2 operated on all two-or-more-person households where at least two individuals had differently-spelled last names. This is done using a method called a "string comparator". Simply put, a string comparator "score" quantifies the degree to which two strings are the same. Factors such as length of the string, as well as similarity in specific characters, enter into the score assigned to two compared strings. If two strings are exactly the same, they have a comparator score of 1.0. If they are highly dissimilar (no letters in common) they have a value of zero. Table 3 shows six examples of name pairs with their comparator scores.

Table 3

Selected Name Pairs and Their Comparator Scores

Name 1	Name 2	Comparator Score
PADERAWSKI	PADEREWSKI	.9800
BROWN	BROVN	.9347
WORD	WARD	.8950
WOOD	WORD	.8667
WOOD	WARD	.7450
KNIGHT	DAY	.0000

Table 3 shows that KNIGHT and DAY are very dissimilar, as one would expect, so their value is 0. Slight variations in short strings, such as WORD, WARD and WOOD, lead to lower scores than a small variation in a longer string, such as BROWN. Relatively high scores accrue to longer strings with only small variations, such as PADEREWSKI/PADERAWSKI.

In this research, the Lynch-Winkler (1994) string comparator was used to help measure the similarity between each pair of last names whose lengths differed by no more than one letter. Two names were judged to be similar based on a combination of the comparator score and the frequency of the less frequent name. When the string comparator score of two last names exceeded the prespecified threshold shown in Table 4, we computed the initial frequencies of the two last names and changed the less frequent last name to the more frequent one. Table 4 shows the criteria for editing last names. If the string comparator returns the score in the left column, the less frequent name is edited to the more frequent name, when the former's frequency is less than the value shown in the right column. Thus, in a household with a SMITH and a SNITH, the SNITH entry was edited to SMITH. For a fuller description of this algorithm, see the discussion in Coleman, Word, and Nunziata (2003).

Table 4

Criteria for Editing Last Names

Comparator Score	Less Common Name Frequency
0.9 or more	unbounded
0.8-0.8999	100
0.7-0.7999	10

The number of records edited in this step was 6,721,444 (last names only).

4.4.4 Edit 3: Correcting transposed names

Despite extensive testing and development of the census questionnaire, not every respondent answers questions in the way we expect them to. Particularly in the case of write-in information, respondents often fail to detect directions or cues that are intended to assist them in providing their response in a correct manner. One common error made by respondents in entering their name is the tendency to transpose their name, that is, to write their first name in the space provided for their last name, and to write their last name in the space provided for their first name. A preliminary examination of the data showed that respondents did occasionally transpose their names on their Census 2000 forms. That is, they wrote their last name in the first name field and vice versa. Edit 3 sought to reduce these errors while not introducing new transposition errors. Analysis of a sample of data after Edit 2 suggested that about 1 in 170 respondents transposed their names, or a probability of transposition of about 0.6 percent. Thus, the 0.6 percent probability was used to create a threshold to trigger transposition of the captured name.

One intuitively believes that a male named "SMITH JOHN has most likely transposed his name, given that SMITH and JOHN are the most common last and male first names, respectively and that these names rarely occur in the reverse order. However, intuition cannot determine whether or not a male captured as THOMAS JAMES should be transposed to JAMES THOMAS, as both THOMAS and JAMES frequently occur as both last names and male first names. Edit 3 operationalizes these intuitions by creating an odds ratio that a name as captured was transposed. The nature of names makes it possible to define an odds ratio measure to detect whether a name as captured was likely to have been transposed.

The odds that a first name, such as JOHN, is correct as captured is essentially the ratio count of JOHN as a first name for the given sex to the count of JOHN as a last name for the same sex. The computation of the odds for a last name, such as SMITH, is done similarly. Two complications occur. The first is simply the possibility of a zero in the numerator or denominator of the odds ratio. To remedy this, we add an arbitrary small number, 0.5, to the initial counts of both the first and last names. The second is the possibility of spuriously increased last name counts for a sex due to contamination by transpositions of first names. For this reason, we use the minimum count by sex for last names. For example, although JENNIFER is a legitimate but infrequent last name, the number of females with the last name JENNIFER greatly exceeds the number of males with the last name JENNIFER, since JENNIFER as a first name is almost exclusively female. Transposition by females spuriously increases the count of females with last name JENNIFER, while males with the last name JENNIFER are virtually unaffected. Taking the minimum of the two counts gets closer to the true count of females (or males) with last name JENNIFER.

The odds ratio, R, is the product of the odds that the first name as given is correct, R_1, and the odds that the last name as given is correct, R_2. Table 5 shows the four ways that a name can be captured.

Table 5

The Four Ways That a Name Can Be Captured

	Sex	
Position	Male	Female
First	*w*	*x*
Last	*y*	*z*

The variables *w*, *x*, *y* and *z* are the counts for a given name in the four cells.

Given a first name, last name pair and a sex, we can compute the components of the odds ratio *R*. Let the subscripts 1 and 2 indicate the first and last names, respectively. For females, $R_1 = (x_1 + 0.5) / (\min(y_1, z_1) + 0.5)$ and $R_2 = (\min(y_2, z_2) + 0.5) / (x_2 + 0.5)$. Likewise, for males, $R_1 = (w_1 + 0.5) / (\min(y_1, z_1) + 0.5)$ and $R_2 = (\min(y_2, z_2) + 0.5) / (w_2 + 0.5)$.

Now, given a respondent's full name after Edit 2 (first_name, last_name) and his/her sex *S*, which can take the values *M* for male and *F* for female, the odds ratio *R* that the name as captured is correct, as opposed to being transposed, is computed as

$$R = \frac{\text{first\_name}_{1S} + 0.5}{\min(\text{first\_name}_{2M}, \text{first\_name}_{2F}) + 0.5} \times \frac{\min(\text{last\_name}_{2M}, \text{last\_name}_{2F}) + 0.5}{\text{last\_name}_{1S} + 0.5}, \qquad (1)$$

where first_name_{tX} and last_name_{tX} are the total counts of first_name and last_name as names of type *t*, where *t* = 1, 2 denote first and last names, respectively, for sex *X*, where *M* is male and *F* is female. The first term in equation (1) is the odds that the first name as captured is correct, assuming independence between first and last names. The second term is the odds that the last name as captured is correct, again assuming independence.

Table 6 illustrates the results of applying this algorithm. For four name pairs, it shows the odds ratio that the entries as captured were not transposed and the probabilities that they are correct. For purposes of Table 6, last names are assumed to be divided equally between the two sexes.

Table 6

Name Pairs as Captured and Their Transposition Odds and Probabilities

First Name	Last Name	Sex	Odds Ratios			Probability of Being Correct
			R_1	R_2	R	
KING	JENNIFER	Female	0.007	0.069	0.00045	<0.001
JONES	WILLIAM	Male	0.004	0.012	0.00005	<0.001
LINDA	SMITH	Female	102.083	2861.176	292075.92159	1.000
THOMAS	JAMES	Male	4.890	0.125	0.61284	0.380

In the first two pairs, the odds that the captured names are transposed are overwhelming, much less than the 1:170 (i.e., $R < 0.006$) threshold needed to trigger transposition by Edit 3. In the JENNIFER KING example, Edit 3 would transpose the two name entries. LINDA SMITH, on the other hand, has almost certainly been captured correctly, given the massively large odd ration it produces. THOMAS JAMES is a more interesting situation. The *R* value is far above .006, translating to a probability of correctness of 0.38, which means not even a 50-50 chance that it is transposed. Thus, when a male gives his name as THOMAS JAMES, we have insufficient evidence to transpose it. Likewise, a male giving his name as JAMES THOMAS also lacks sufficient evidence of transposition.

The count of edits via transposition was 1,352,881. The numbers of first and last names changed are equal.

4.4.5 Edit 4: Edits of large occurrence names

At the completion of the preedit and edits 1 through 3, all first and last names which had 300 or more occurrences were further reviewed. During this process three additional edits were implemented.

4.4.5.1 Edit 4, Part a: Editing or deleting invalid first and last names

The first part of this edit implemented the Criterion 2 rule (validation from the SSA NUMIDENT file), and compared counts of all names with at least 300 occurrences to their respective counts in the 5 percent sample of the Social Security NUMIDENT file. For each name of each type (last, male first, and female first), we computed the ratio of Census file occurrences after Edit 3 to the NUMIDENT sample occurrences. Names which either did not occur at all in the NUMIDENT sample or had ratios of 50:1 or greater were then determined to be either invalid (BABY, BOY, GIRL), mis-scanned (HAII for HALL), or misspelled (JOESPH for JOSEPH). Mis-scanned and misspelled names were edited to the most likely alternative; invalid names were deleted.
This edit resulted in 97,129 deletions.

4.4.5.2 Edit 4, Part b: Deleting spurious multiplicates

The second part of this edit involved an examination of records where the exact name and date of birth occurred six or more times. While the majority were due to coincidence (e.g., LINDA SMITH born September 30, 1947 and JOSE HERNANDEZ born March 19, 1963), some seemed invalid – multiple occurrences on the same form, or in the same specific geographic place. These spurious records may have resulted from respondents entering their names multiple times on the same form, or from a respondent completing multiple forms. All identified spurious multiplicate records were deleted.
The counts of these edits and deletions were: first names, 711,027; last names, 215,743.

4.4.5.3 Edit 5: Spurious male first initials

A final examination revealed a few male first names where it appeared that a single letter was attached to the beginning of a valid name. For example, if F. Scott Fitzgerald gave his first name as F SCOTT, the file would have retained it as FSCOTT. This edit was similar to Edit 1, with the difference that the first letter of various names was trimmed. This edit was only applied to male first names, as no female first name occurring more than 300 times had this type of error.

The number of male first names edited was 124,118.

4.4.6 Edits summary

In all, 8,701,943 changes were made to first names and 10,266,502 to last names. In every instance, the original name had fewer occurrences than the name to which it was changed. Since every change had to result from the application of a replicable edit rule, many errors in names were not corrected, and as noted earlier, likely result in many of the unique occurrence names in the file. Nevertheless, we believe the quality of the resulting name files is considerably higher than that of the initial, unedited data.

4.4.7 Definitions of race and Hispanic origin used in this report

The race categories shown in these files are the modified race categories used in the Census Bureau's population estimates program. The Census Bureau's population estimates program modified the Census 2000 race data to eliminate the "Some Other Race" category to be more consistent with the race categories that appear on the administrative records used to produce population estimates.

The race modification generally conforms to the Office of Management and Budget's (OMB, 1997) standards for collecting and presenting data on race and ethnicity. The OMB (1997) standards identified five race categories: 1) White, 2) Black or African American, 3) American Indian and Alaska Native, 4) Asian, and 5) Native Hawaiian and Other Pacific Islander. Additionally, OMB (1997) recommended that respondents be able to select one or more races to indicate their racial identity. For respondents unable to identify with any of the five race categories, OMB (1997) approved including a sixth category - "Some Other Race" – on the Census 2000 questionnaires.

About 18.5 million people checked "Some Other Race" alone or in combination with one or more other races. These people were primarily of Hispanic origin and many wrote in a specific Hispanic-origin type (e.g., Mexican or Puerto Rican) as their race. To conform to the U.S. Census Bureau's population estimates program's race definitions, responses that were only "Some Other Race" were modified by blanking the "Some Other Race" response and imputing one or more OMB standard races. Standard Census Bureau procedures achieved the imputation by assigning the race response(s) of another Census respondent who had the same response to the Hispanic-origin question. Reponses that

included both "Some Other Race" and one or more OMB standard races were modified just by blanking the "Some Other Race" response.

For purposes of this report, all people were categorized into six mutually exclusive racial and Hispanic origin groups. People indicating that they were Hispanic were categorized as Hispanic, regardless of race. The remaining non-Hispanic population was collapsed into five non-Hispanic race categories: 1) White only, 2) Black only, 3) American Indian and Alaskan Native only, 4) Asian and Pacific Islander only, and 5) Two or More Races. Native Hawaiians and Other Pacific Islanders were combined with Asians because of the former group's small total – only 398,835 people reported these races, out of a total U.S. population of 281,421,906. This combination is also consistent with the OMB race classification used before 1997.

5. DATAFILE DOCUMENTATION

5.1 General

This report is accompanied by two datafiles containing summary data on last name frequencies. Tabulations with demographic characteristics are available for all surnames occurring at least 100 times in Census 2000. The number 100 was chosen to eliminate those names with very low counts, thus assuring confidentiality. Placing a floor of 100 on the frequencies protects individuals privacy, but does eliminate data on millions of names. Information is provided for 151,671 surnames, covering about 89.8 percent of the population. [1]

The data associated with this report are contained in the Appendix and in two electronic files (A and B). Appendix Table A-1 contains a printed list on the 50 most frequently occurring last names. File A consists of a CSV (comma separated values) file of the 1,000 most frequently occurring last names, similar in format to Appendix Table A-1. File B consists of a CSV file of all names occurring at least 100 times, constituting 151,671 records. In all tables, cell values between 1 and 4 were suppressed to maintain confidentiality; resulting percentages may, therefore, not sum to 100.

The data include two fields identified as "proportion per 100,000 people" and "cumulative proportion per 100,000 people." The number 100,000 is convenient for expressing these proportions, as it reduces the need to show many digits after the decimal point. The term "cumulative proportion" indicates the proportion of the total population covered by that name and all more frequent names.

5.2 File A: Excel File of 1,000 Most Frequently Occurring Surnames

[1] The percentages reported in this paragraph use the total number of people (total, male, female) with valid names as determined by this report as the denominators.

File A is similar in format to Appendix Table A-1, with the difference being that the first row contains field labels and the subsequent 1,000 rows contain data. The file is in Excel format, suitable for import into many software packages.

File A: Surnames: Counts and Distribution in Percent by Non-Hispanic Race and Hispanic Origin: 2000 - Top 1000 Names

Field	Description
name	Last name
rank	Rank
count	Number of occurrences
prop100k	Proportion per 100,000 people for name
cum_prop100k	Cumulative proportion per 100,000 people
pctwhite	Percent Non-Hispanic White Only
pctblack	Percent Non-Hispanic Black Only
pctapi	Percent Non-Hispanic Asian and Pacific Islander Only
pctaian	Percent Non-Hispanic American Indian and Alaskan Native Only
pct2prace	Percent Non-Hispanic of Two or More Races
pcthispanic	Percent Hispanic Origin

Record count: 1000 names
Fields suppressed for confidentiality are assigned the value (S).

5.3 File B: CSV File of All Surnames Occurring 100 or More Times

File B contains the same data as the previous files, but for all names occurring at least 100 times. Users working with these files are urged to confirm that they can match the percentages in File A before proceeding to any set of routines involving percentages.

File B: Surnames: Counts, Total and by Non-Hispanic Race and Hispanic Origin: 2000 – All Names of Count 100 or Greater

Field	Description
name	Last name
rank	Rank
count	Number of occurrences
prop100k	Proportion per 100,000 people for name
cum_prop100k	Cumulative proportion per 100,000 people
pctwhite	Percent Non-Hispanic White Only
pctblack	Percent Non-Hispanic Black Only
pctapi	Percent Non-Hispanic Asian and Pacific Islander Only
pctaian	Percent Non-Hispanic American Indian and Alaskan Native Only
pct2prace	Percent Non-Hispanic of Two or More Races
pcthispanic	Percent Hispanic Origin

Record count: 151,671 names
Fields suppressed for confidentiality are assigned the value (S).

6. REFERENCES

Coleman, Charles D., David L. Word and Robert Nunziata. 2003. "Algorithms for Clustering Elements of a Vector Equipped with a Similarity Relation," internal U.S. Census Bureau manuscript.

Lynch, Maureen P. and William E. Winkler. 1994. "Improved String Comparator," Technical report, Statistical Research Division, U.S. Census Bureau.

Passel, Jeffrey S. and David L. Word. 1980. "Constructing the List of Spanish Surnames for the 1980 Census: An Application of Bayes' Theorem." Paper presented to the 1980 meetings of the American Statistical Association, Denver, CO.

Shackleford, Michael W. 1998. "Name Distributions in the Social Security Area, August 1997." Social Security Administration Actuarial Note number 139, Office of the Chief Actuary, Social Security Administration, June 1998, <http://www.ssa.gov/OACT/NOTES/note139/original_note139.html>, accessed December 24, 2003.

U.S. Census Bureau. 1995. "Frequently Occurring First Names and Surnames From the 1990 Census." Available at <http://www.census.gov/genealogy/www/freqnames.html>.

U.S. Office of Management and Budget. 1997. "Revisisions to the Standards for the Classification of Federal Data on Race and Ethnicity," *Notice*, Vol. 62, No. 210, October 30, 1997.

U.S. Social Security Administration. 1957. *Preliminary Report of Distribution of Surnames in the Social Security Number File*.

U.S. Social Security Administration. 1964. *Report of Distribution of Surnames in the Social Security Number File*.

U.S. Social Security Administration. 1975. *Report of Distribution of Surnames in the Social Security Number File, September 1, 1974*, SSA Publication Number 034-75.

U.S. Social Security Administration. 1985. *Report of Distribution of Surnames in the Social Security Number File, September 1, 1984*. SSA Publication Number 42-004, April 1985.

U.S. Social Security Administration. 2003. "Popular Baby Names." <http://www.ssa.gov/OACT/babynames/>, accessed December 24, 2003.

Word, David L. and R. Colby Perkins, Jr. 1996. "Building a Spanish Surname List for the 1990s: A New Approach to an Old Problem," U.S. Census Bureau Population Division Technical Working Paper No. 13. Available at <http://www.census.gov/population/documentation/twpno13.pdf

TABLE A-1

Top 50 Surnames: Counts and Distribution in Percent by Non-Hispanic Race and Hispanic Origin: 2000

Name	Rank	Count	Proportion per 100,000 Individual Name	Proportion per 100,000 Cumulative	Percent Non-Hispanic White	Black	American Indian/ Alaskan Native	Asian/ Pacific Islander	Two or More Races	Hispanic (of any race)
TOTAL		269,762,087			69.8	11.8	0.7	3.7	1.7	12.3
SMITH	1	2,376,206	880.9	880.9	73.3	22.2	0.8	0.4	1.6	1.6
JOHNSON	2	1,857,160	688.4	1569.3	61.5	33.8	0.9	0.4	1.8	1.5
WILLIAMS	3	1,534,042	568.7	2138.0	48.5	46.7	0.8	0.4	2.0	1.6
BROWN	4	1,380,145	511.6	2649.6	60.7	34.5	0.8	0.4	1.9	1.6
JONES	5	1,362,755	505.2	3154.7	57.7	37.7	0.9	0.3	1.9	1.4
MILLER	6	1,127,803	418.1	3572.8	85.8	10.4	0.6	0.4	1.3	1.4
DAVIS	7	1,072,335	397.5	3970.3	64.7	30.8	0.8	0.4	1.7	1.6
GARCIA	8	858,289	318.2	4288.5	6.2	0.5	0.6	1.4	0.5	90.8
RODRIGUEZ	9	804,240	298.1	4586.6	5.5	0.5	0.2	0.6	0.4	92.7
WILSON	10	783,051	290.3	4876.9	69.7	25.3	1.0	0.5	1.7	1.7
MARTINEZ	11	775,072	287.3	5164.2	6.0	0.5	0.6	0.6	0.5	91.7
ANDERSON	12	762,394	282.6	5446.8	77.6	18.1	0.7	0.5	1.6	1.6
TAYLOR	13	720,370	267.0	5713.9	67.8	27.7	0.7	0.4	1.8	1.6
THOMAS	14	710,696	263.5	5977.3	55.5	38.2	1.0	1.6	2.0	1.7
HERNANDEZ	15	706,372	261.8	6239.2	4.6	0.4	0.3	0.6	0.3	93.8
MOORE	16	698,671	259.0	6498.2	68.9	26.9	0.7	0.4	1.7	1.5
MARTIN	17	672,711	249.4	6747.5	77.5	15.3	0.9	0.7	1.6	4.0
JACKSON	18	666,125	246.9	6994.5	41.9	53.0	1.0	0.3	2.2	1.5
THOMPSON	19	644,368	238.9	7233.3	72.5	22.5	1.2	0.4	1.8	1.6
WHITE	20	639,515	237.1	7470.4	67.9	27.4	1.0	0.4	1.8	1.5
LOPEZ	21	621,536	230.4	7700.8	5.8	0.6	0.5	1.0	0.5	91.5
LEE	22	605,860	224.6	7925.4	40.1	17.4	1.0	37.8	2.3	1.3
GONZALEZ	23	597,718	221.6	8147.0	4.8	0.4	0.2	0.4	0.3	94.0
HARRIS	24	593,542	220.0	8367.0	53.9	41.6	0.7	0.4	2.0	1.5
CLARK	25	548,369	203.3	8570.3	76.8	18.5	0.9	0.4	1.6	1.7
LEWIS	26	509,930	189.0	8759.3	61.0	33.8	1.1	0.5	2.0	1.6
ROBINSON	27	503,028	186.5	8945.8	51.3	44.1	0.5	0.4	2.0	1.7
WALKER	28	501,307	185.8	9131.6	61.2	34.2	0.8	0.4	1.8	1.6
PEREZ	29	488,521	181.1	9312.7	5.9	0.5	0.3	1.2	0.5	91.6
HALL	30	473,568	175.6	9488.2	75.1	20.8	0.6	0.5	1.6	1.4
YOUNG	31	465,948	172.7	9661.0	68.9	23.8	0.7	2.9	1.9	1.7
ALLEN	32	463,368	171.8	9832.7	70.2	25.1	0.8	0.4	1.8	1.6
SANCHEZ	33	441,242	163.6	9996.3	5.8	0.5	0.5	1.0	0.5	91.8
WRIGHT	34	440,367	163.2	10159.6	68.3	27.4	0.7	0.4	1.8	1.5
KING	35	438,986	162.7	10322.3	72.8	22.0	1.0	0.9	1.7	1.6
SCOTT	36	420,091	155.7	10478.0	62.6	32.3	1.2	0.4	1.9	1.7
GREEN	37	413,477	153.3	10631.3	59.3	36.2	0.6	0.3	1.8	1.7
BAKER	38	413,351	153.2	10784.5	82.1	13.6	0.8	0.5	1.5	1.5
ADAMS	39	413,086	153.1	10937.6	76.2	19.2	0.8	0.5	1.6	1.8
NELSON	40	412,236	152.8	11090.5	80.3	14.9	1.1	0.5	1.5	1.7
HILL	41	411,770	152.6	11243.1	66.8	28.4	0.9	0.4	1.8	1.6
RAMIREZ	42	388,987	144.2	11387.3	4.4	0.3	0.3	1.0	0.4	93.7
CAMPBELL	43	371,953	137.9	11525.2	76.5	19.1	0.6	0.4	1.7	1.7
MITCHELL	44	367,433	136.2	11661.4	63.5	31.5	1.0	0.4	1.9	1.6
ROBERTS	45	366,215	135.8	11797.1	79.6	15.9	0.8	0.5	1.7	1.6
CARTER	46	362,548	134.4	11931.5	60.5	35.0	0.7	0.4	1.9	1.5
PHILLIPS	47	351,848	130.4	12062.0	79.0	16.4	1.0	0.4	1.7	1.6
EVANS	48	342,237	126.9	12188.8	70.7	25.0	0.7	0.4	1.7	1.5
TURNER	49	335,663	124.4	12313.3	66.7	29.3	0.6	0.3	1.7	1.4
TORRES	50	325,169	120.5	12433.8	6.0	0.6	0.3	1.4	0.5	91.2

Race/Hispanic groups are mutually exclusive; see report for details

Appendix A. Documentation of Edit Steps and Records Affected

Total persons: Initial File	279,132,770	
Number of persons removed by DSCMO operations	5,878,113	
Number of persons removed by the preliminary edits	5,775,577	^1
Number of persons with less than 2 characters in either the first or last name	5,750,926	^1
Number of first names with less than 2 characters	4,736,292	
Number of last names with less than 2 characters	4,207,119	
Number of persons with the name JOHN/JANE DOE	24,588	^1
Number of persons with a numeral in first or last name	63	^1
Number of first names containing a numeral	51	
Number of last namescontaining a numeral	16	
Total persons: Census 2000 - post preliminary edits	269,768,216	
Number of person records changed by any preliminary edit	5,637,813	
Number of person records changed to remove titles	5,509,745	
Number of first names with titles removed	5,509,745	
Number of last names with titles removed	0	
Number of person records changed to CHRISTOPHER	14,957	
Number of person records changed to concatenate LYN/LYNN	29,740	
Number of person records changed to truncate characters after the first occurrence of a space	85,903	
Edit 1 - post preliminary edit	269,768,216	
Total number of person records changed by edit 1	3,097,416	
Number of first names changed to remove dangling character	826,104	
Number of last names changed to remove dangling character	2,226,434	
Edit 2 - post edit 1	269,768,216	
Number of last names changed	6,721,444	
Edit 3 - post edit 2	269,768,216	
Number of names transposed	1,352,881	
Edit 4 - post edit 3	269,768,216	
Total number of person records removed by edit 4a	97,129	^2
Total persons dropped by edit 4a - first names	96,046	^2
Total persons dropped by edit 4a - last names	1,083	^2
Total number of person records changed by edit 4b (excluding records dropped in edit 4a)	925,684	
Total number of first names changed	711,027	
Total number of last names changed	215,743	
Edit 5 - post edit 4	269,762,087	
Number of first names changed to remove first character	124,118	

^1 - mutually exclusive: adds to the total number of persons dropped during the preliminary edit

^2 - mutually exclusive: adds to the total number of persons dropped during edit 4a