ED 085 146 RC 007 491

AUTHOR Giesecke, G. Lee
TITLE A Comparison of Computerized Techniques for Recognized Spanish Names.
INSTITUTION Human Resources Research Organization, Alexandria, Va.
SPONS AGENCY Department of Defense, Washington, D.C. Assistant Secretary of Defense (Manpower and Reserve Affairs).
REPORT NO DAHC-15-73-C-0131; HumRRO-PP-9-73
PUB DATE Oct 73
NOTE 26p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Classification; Codification; Comparative Analysis; *Computer Science; *Information Processing; *Input Output Analysis; *Spanish Speaking; Tables (Data)

ABSTRACT

The results of various computerized coding techniques used to identify persons of Spanish derivation were compared to questionnaire responses in which people identified themselves as Spanish. The coding techniques classified names as Spanish and non-Spanish using the following: census surnames; Morton surnames; "Broad" Spanish surnames; "Narrow" Spanish surnames; the Buechley technique; "Broad" Spanish first names; "Narrow" Spanish first names; and combinations of these lists. For each coding technique an estimate was calculated for the proportion of persons with Spanish names not classifing themselves as Spanish and persons classifying themselves as Spanish who did not have Spanish names. These estimates were for the United States as a whole and for groupings by geographic area, age, educational attainment, and percentile on the Armed Forces Qualification Test. Data were obtained from the U.S. Air Force Airman Sample Survey of March 1971 and the Air Force Master File of male enlisted personnel (June 1971). It was found that outside of the 5 Southwestern states and at higher educational and aptitude levels, the name recognition procedures include increasing proportions of persons who do not classify themselves as Spanish. Data are presented in tabular form. (NQ)

Professional
Paper
9-73

HumRRO-PP-9-73

HumRRO

A Comparison of Computerized Techniques for Recognizing Spanish Names

G. Lee Giesecke

[illegible]

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street • Alexandria, Virginia 22314

Approved for public release; distribution unlimited.

October 1973

Prepared for

Directorate for Manpower Research
Office of the Assistant Secretary of Defense
(Manpower and Reserve Affairs)
Department of Defense

The Human Resources Research Organization (HumRRO) is a nonprofit corporation established in 1969 to conduct research in the field of training and education. It is a continuation of The George Washington University Human Resources Research Office. HumRRO's general purpose is to improve human performance, particularly in organizational settings, through behavioral and social science research, development, and consultation.

Published
October 1973
by
HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street
Alexandria, Virginia 22314

HumRRO
Professional
Paper
9-73

A Comparison of Computerized Techniques for Recognizing Spanish Names

by

G. Lee Giesecke

October 1973

Prepared for

Directorate for Manpower Research
Office of the Assistant Secretary of Defense
(Manpower and Reserve Affairs)
Department of Defense
Contract DAHC 15-73-C-0131

HumRRO Division No. 7 (Social Science)
Alexandria, Virginia 22314
HUMAN RESOURCES RESEARCH ORGANIZATION

Prefatory Note

This research was conducted under the sponsorship of the Assistant Secretary of Defense (Manpower and Reserve Affairs). The analysis was performed by G. Lee Giesecke, Senior Technician, HumRRO Division No. 7, Human Resources Research Organization, Alexandria, Virginia. The project was conducted under contract DAHC 15-73-C-0131 (HumRRO Project DATA).

A COMPARISON OF COMPUTERIZED TECHNIQUES FOR RECOGNIZING SPANISH NAMES

G. Lee Giesecke

INTRODUCTION

This research makes use of Air Force survey data to show the relationships between those persons who classified themselves as Spanish and those persons whose names would be treated as Spanish by various computerized coding techniques. For each coding technique an estimate is calculated for:

(1) The proportion of persons with Spanish names who did not classify themselves as Spanish.

(2) The proportion of persons who classified themselves as Spanish who did not have Spanish names.

These estimates are calculated for the United States as a whole and for several broad groupings by geographic area, age, educational attainment, and percentile on the Armed Forces Qualification Test (AFQT).

BACKGROUND AND OBJECTIVES OF THE STUDY

Since the Census Bureau's use of Spanish surnames in the 1950 Census, Spanish names have been increasingly used to identify persons of Spanish culture in the United States. The 1960 and 1970 Censuses again made use of Spanish surnames, as did the Census report of *Minority-Owned Busindsses: 1969* (1). An unknown but growing number of research studies (e.g., 2-6) have utilized Spanish surnames as a means of classifying data by ethnic group. In addition, Title 7 of the Civil Rights Act of 1964 requires that employers of 25 or more persons report the number of employees with Spanish surnames in each position held in the company. In some cases (7) the Spanish minorities in the United States are now referred to as "Spanish-surnamed" individuals, as if the name rather than the cultures were the important group-defining characteristic.

Despite this widespread use of the Spanish surname as a surrogate variable, there are no recent studies which attempt to show the degree of correspondence between Spanish-surnamed individuals and individuals who belong to a Spanish cultural group. The present study attempts to remedy this deficiency within certain limitations imposed by the data.

PLAN OF THE STUDY

The data are taken from the U.S. Air Force Airman Sample Survey of March 1971, and the Air Force Master File of male enlisted personnel as of 30 June 1971.

An airman survey is performed triannually to answer a wide variety of questions of interest to the Air Force. The March 1971 survey consisted of a mark sense scanner form which a 5% sample of airmen were asked to complete during duty hours. The questionnaire contained 143 questions and was completed by 29,000 airmen. Excluding airmen on leave, the response to the survey was approximately 90%.

The questionnaire included the following items:

Item	Percent Complete
Social Security Number	97.9
Air Force Specialty Code (Questions #14 and 15)	97.5
Ethnic Question (Question #52)	99.7

The wording of the ethnic question was as follows:

"Which of the following do you consider yourself?"
A. Negro/Black
B. Spanish or Mexican American
C. American Indian
D. Oriental
E. White
F. Other

The individual's name, geographic area, age, and Armed Forces Qualification Test (AFQT) percentile were obtained from the Air Force Master File rather than from the survey data. Linkage to the master file required matches on both the social security number and the Air Force specialty code, as well as a valid ethnic code in the survey data. There remained 22,193 cases for analysis. Eliminating the requirement for a match on the Air Force specialty code would have left 25,351 cases for analysis; however, the quality of the data would have been substantially lower. The items extracted from the master file were as follows:

Item	Percent Complete (for matching cases)
Name	100.0
Educational Level	99.9
Home State	71.1
Armed Forces Qualification Test	63.4
Birth Date	99.4

PROCEDURE

The basic computerized technique for classifying names as Spanish or non-Spanish is to sort the names alphabetically and to compare the sorted cases against entries on a file of Spanish surnames (which is also sorted alphabetically). If an individual's name appears in the Spanish-surname file, his name is classified as Spanish. With this approach the names of the surveyed individuals were classified as Spanish or non-Spanish using each of the following lists:

(1) Census surnames (8).
(2) Morton surnames (9)—a list prepared by Dr. William E. Morton.
(3) "Broad" Spanish surnames—a list prepared by the author and Dr. Santiago Rodriguez[1] by adding to a census list names selected from a file of men separated from the Army. A preliminary selection was made by listing the names of persons who either lived in selected zipcode areas or who had

[1] Dr. Rodriguez is on the staff of the Equal Opportunity Commission.

Spanish first names. The final selection was made manually by Dr. Rodriguez.

(4) "Narrow" Spanish surnames—a subset of the "broad" surnames, developed chiefly by Dr. Rodriguez. Names which occur frequently in non-Spanish cultural groups were excluded.

In addition, an ingenious technique for recognizing Spanish surnames has been developed by Dr. Robert Buechley (10, 11). This technique is based on surname endings and letter combinations. This technique will be referred to as the:

(5) Buechley technique.

Two further procedures classify an individual as Spanish or non-Spanish based upon his first name. These do not require a separate sort of the file, since the list of first names is short enough to be stored in the computer memory and accessed randomly using a search procedure. This approach was used with the following two name lists:

(6) "Broad" Spanish first names—a list of male names developed from a file of Army separatees. The first names of individuals having Spanish surnames were collected. The resulting list was screened by Dr. Rodriguez to eliminate the non-Spanish first names.

(7) "Narrow" Spanish first names—a subset of the "Broad" Spanish first names developed by Dr. Rodriguez. "Broad" first names which occur frequently in non-Spanish cultures were eliminated.

Finally, it is possible to classify an individual as Spanish or non-Spanish based upon different combinations of the above criteria. For example, we might require that an individual have both a narrow Spanish surname and a narrow Spanish first name before classifying the individual as Spanish.

Given these classification schemes and the survey data, it is possible to compare the classification schemes with how the individuals classified themselves.

RESULTS OF THE STUDY

FALSE CLASSIFICATIONS

A comparison of the different classification schemes is given in Table 1. To simplify presentation, it is assumed in the tables that an individual's classification of himself is correct.[2] Those cases "falsely classified as Spanish" in Table 1 are individuals who completed something besides "Spanish or Mexican American" on the ethnic question but whose names were treated as Spanish by a given classification technique. Similarly, those cases "falsely classified as non-Spanish" had entries of "Spanish or Mexican American" on the questionnaire, but their names were not considered Spanish by another classification technique.

INCLUSIVENESS VERSUS EXCLUSIVENESS

In Table 1 it is possible to see obvious tradeoffs between including as many as possible who can reasonably be classified as Spanish and excluding all those who should not be classified as Spanish. For most statistical purposes, the latter is the more important criterion. It is possible to correct for undercounts, but there is no way of correcting a cross-tabulation biased by a substantial number of individuals misclassified by cultural group.

[2] As we will see, the assumption is not always valid.

Table 1

Number and Percent of Persons Falsely Classified as Spanish or Non-Spanish, by Classification Procedure

	Classification Procedure	Number Classified As Spanish	Persons Falsely Classified					
			As Spanish		As Non-Spanish		Total	
			N	%[a]	N	%[b]	N	%[c]
1.	"Broad" Spanish surname	1,025	420	41.0	98	13.9	518	2.3
2.	"Narrow" Spanish surname	814	230	28.3	119	16.9	349	1.6
3.	Census Spanish surname	917	350	38.2	136	19.4	486	2.2
4.	Morton Spanish surname	974	391	40.1	120	17.1	511	2.3
5.	Buechley technique	1,163	550	47.3	90	12.8	640	2.9
6.	Any of the above	1,436	807	56.2	74	10.5	881	4.0
7.	All the above	733	179	24.4	149	21.2	328	1.5
8.	"Broad" Spanish first name	732	393	53.7	364	51.8	757	3.4
9.	"Narrow" Spanish first name	332	78	23.5	449	63.9	527	2.4
10.	Any of the above	1,767	1,119	63.3	55	7.8	1,174	5.3
11.	All the above	246	29	11.8	486	69.1	515	2.3
12.	"Narrow" surname OR ("broad" first name and "broad")surname	822	237	28.8	118	16.8	355	1.6
13.	"Narrow" surname OR ("Nârrow" first name)	885	275	31.1	93	13.2	368	1.7
14.	"Narrow" surname OR ("narrow" first name and "broad" surname)	824	232	28.2	111	15.8	343	1.5
15.	"Narrow" surname OR ("narrow" first name and Morton surname)	825	232	28.1	110	15.7	342	1.5
16.	"Narrow" surname OR ("narrow" first name and Buechley surname)	837	241	28.8	107	15.2	348	1.6

[a] Denominator used for these percentages was the number of persons classified as Spanish by the various coding techniques.

[b] Denominator used for these percentages was the number of persons who classified themselves as Spanish, 703.

[c] Denominator used for these percentages was the number of persons included in the survey, 22,193.

There are, however, limits to how exclusively we can define the Spanish group. The requirement that an individual meet all the name criteria (Table 1, line 11) resulted in only 11.8% misclassified as Spanish. However, only 30.9% of those who considered themselves Spanish were included. It is doubtful that such a small group would be representative. A definition of "Spanish" that requires a Spanish first name is simply too restrictive in the United States. Even among persons having "narrow" Spanish surnames, 48.1% who classified themselves as Spanish did not have Spanish first names.

Of the simple surname classification procedures (Table 1, lines 1-5), the "narrow" Spanish-surname test seems to be the best scheme for general statistical procedures. Fewer persons are misclassified as Spanish and fewer persons are misclassified overall than with the other surname procedures. The results are significant ($p < .01$).

USE OF FIRST NAMES

Attempts to improve the "narrow" surname procedure by additionally coding as Spanish those persons who meet a first name criterion (Table 1, lines 12-16) were not particularly successful. In a few cases, the overall number of misclassifications was reduced; however the differences were too small to justify the additional computational effort and were, in any case, not significant.

Table 1, line 6, suggests that the "narrow" surname procedure could be improved by redefining "narrow" to exclude those surnames not treated as Spanish by the Morton, Census, or Buechley procedures. The difference in overall number of misclassifications was not significant ($\chi^2=1.3$); however, the more exclusive procedure reached significance ($p<.05$) in testing for differences in the proportion of those classified as Spanish who were misclassified ($\chi^2=5.3$).

GEOGRAPHIC DIFFERENCES

When the data are broken out geographically (Table 2) the advantages of a "narrow" surname classification procedure are still apparent. In general, however, all the name classification schemes do rather poorly outside the southwestern United States. This raises the question as to whether persons outside the Southwest who derive from a Spanish-speaking culture are more likely to have been assimilated into the dominant culture or whether such persons are less likely to think of themselves as Spanish regardless of their level of acculturation.

There are not enough cases for further breakdowns within the geographic area.

DIFFERENCES BY EDUCATIONAL LEVEL

It is apparent that geography is not the only issue in determining ethnic classification. Table 3 shows that Spanish-named persons with more than a high school education were less likely to think of themselves as Spanish ($p<.025$).

DIFFERENCES BY AFQT PERCENTILE

Table 4 shows similar differences by percentile score on the Armed Forces Qualification Test (AFQT). At higher AFQT percentiles, persons with Spanish names are less likely to classify themselves as Spanish ($p<.01$). The AFQT is primarily a general aptitude test, rather than an IQ test. It seems reasonable that persons more assimilated into the dominant culture would score higher on the AFQT and also be less likely to classify themselves as Spanish. The chi-square statistics are significant ($p<.01$).

DIFFERENCES BY AGE

Cross-tabulations by age (Table 5) show no clear trend. In the column for persons falsely classified as Spanish, most of the classification schemes show an apparent slight trend whereby younger persons with Spanish names are less apt to classify themselves as Spanish; however, the Buechley technique shows the oposite trend. Using chi-square tests, it appears that none of the relationships is significant.

RATIOS

A related issue is whether the ratio between the numbers of persons classified as Spanish by two different techniques varies substantially for different population

Table 2

Number and Percent of Persons Falsely Classified as Spanish or Non-Spanish, by Geographic Area and Classification Procedure

Geographic Area	Classification Procedure	Number Classified As Spanish	Persons Falsely Classified					
			As Spanish		As Non-Spanish		Total	
			N	%[a]	N	%[b]	N	%[c]
Five southwestern states 2,851 persons (387 persons self-classified as Spanish)	"Narrow" surname	399	58	14.5	46	11.9	104	3.6
	Census surname	403	72	17.9	56	14.5	128	4.5
	Morton surname	411	72	17.5	48	12.4	120	4.2
	Buechley technique	441	88	20.0	34	8.8	122	4.3
New York, New Jersey, Florida 2,252 persons (87 persons self-classified as Spanish)	"Narrow" surname	105	39	37.1	21	24.1	60	2.7
	Census surname	118	54	45.8	23	26.4	77	3.4
	Morton surname	138	72	52.2	21	24.1	93	4.1
	Buechley technique	168	96	57.1	15	17.2	111	4.9
Other areas 10,686 persons (78 self-classified as Spanish)	"Narrow" surname	120	66	55.0	24	30.8	90	0.8
	Census surname	193	138	71.5	23	29.5	161	1.5
	Morton surname	210	153	72.9	21	26.9	174	1.6
	Buechley technique	286	230	80.4	22	28.2	252	2.4
State unknown 6,406 persons (151 self-classified as Spanish)	"Narrow" surname	190	67	35.3	28	18.5	95	1.5
	Census surname	203	86	42.4	34	22.5	120	1.9
	Morton surname	215	94	43.7	30	19.9	124	1.9
	Buechley technique	268	136	50.8	19	12.6	155	2.4

[a] Denominators used for these percentages were the numbers of persons classified as Spanish by each technique in each area.
[b] Denominators used for these percentages were the numbers of persons in each area classifying themselves as Spanish.
[c] Denominators used for these percentages were the numbers of persons in each geographic area.

Table 3

Number and Percent of Persons Falsely Classified as Spanish or Non-Spanish, by Education Level and Classification Procedure

Years of School Completed	Classification Procedure	Number Classified As Spanish	Persons Falsely Classified					
			As Spanish		As Non-Spanish		Total	
			N	%[a]	N	%[b]	N	%[c]
Under 12 years 447 persons (27 self-classified as Spanish)	"Narrow" surname	27	1	—[d]	1	—	2	—
	Census surname	30	5	—	2	—	7	—
	Morton surname	33	7	—	1	—	8	—
	Buechley technique	34	8	—	1	—	9	—
12 years 18,776 persons (624 self-classified as Spanish)	"Narrow" surname	710	196	27.6	110	17.6	306	1.6
	Census surname	803	302	37.6	123	19.7	425	2.3
	Morton surname	854	335	39.2	105	16.8	440	2.3
	Buechley technique	1007	465	46.2	82	13.1	547	2.9
Over 12 2,948 persons (52 self-classified as Spanish)	"Narrow" surname	77	33	42.9	8	—	41	1.4
	Census surname	84	43	51.2	11	—	54	1.8
	Morton surname	87	49	56.3	14	—	63	2.1
	Buechley technique	121	76	62.8	7	—	83	2.8

[a] Denominators used for these percentages were the numbers of persons classified as Spanish by each procedure for each educational level.

[b] Denominators used for these percentages were the numbers of persons who classified themselves as Spanish at each educational level.

[c] Denominators used for these percentages were the numbers of persons at each educational level.

[d] — = percentage based on less than 15 observations.

Table 4

Number and Percent of Persons Falsely Classified as Spanish or Non-Spanish, by AFQT Percentile and Classification Procedure

AFQT Percentile	Classification Procedure	Number Classified As Spanish	Persons Falsely Classified					
			As Spanish		As Non-Spanish		Total	
			N	%[a]	N	%[b]	N	%[c]
AFQT ⩽ 33	"Narrow" surname	147	29	19.7	25	17.5	54	2.1
2,577 persons (143 self-	Census surname	157	44	28.0	30	21.0	74	2.9
classified as Spanish)	Morton surname	168	49	29.2	24	16.8	73	2.8
	Buechley technique	205	82	40.0	20	14.0	102	4.0
AFQT 34-67	"Narrow" surname	212	60	28.3	39	20.4	99	1.9
5,084 persons (191 self-	Census surname	241	94	39.0	44	23.0	138	2.7
classified as Spanish)	Morton surname	260	108	41.5	39	20.4	147	2.9
	Buechley technique	310	147	47.4	28	14.7	175	3.4
AFQT > 67	"Narrow" surname	124	47	37.9	13	—[d]	60	1.0
6,414 persons (90 self-	Census surname	160	84	52.5	14	—	98	1.5
classified as Spanish)	Morton surname	170	94	55.3	14	—	108	1.7
	Buechley technique	231	151	65.4	10	—	161	2.5
AFQT Unknown	"Narrow" surname	331	94	28.4	42	15.1	136	1.7
8,118 persons (279 self-	Census surname	359	128	35.7	48	17.2	176	2.2
classified as Spanish)	Morton surname	376	140	37.2	43	15.4	183	2.3
	Buechley technique	417	170	40.8	32	11.5	202	2.5

[a] Denominators used for these percentages were the numbers of persons classified as Spanish by each procedure for each AFQT group.

[b] Denominators used for these percentages were the numbers of persons who classified themselves as Spanish for each AFQT group.

[c] Denominators used for these percentages were the numbers of persons in each AFQT group.

[d] — = percentage based on less than 15 observations.

Table 5

Number and Percent of Persons Falsely Classified as Spanish or Non-Spanish, by Age and Classification Procedure

Age	Classification Procedure	Number Classified As Spanish	Persons Falsely Classified					
			As Spanish		As Non-Spanish		Total	
			N	%[a]	N	%[b]	N	%[c]
Under 25	"Narrow" surname	211	69	32.7	27	16.0	96	1.7
5,656 persons (169 self-	Census surname	228	90	39.5	31	18.3	121	2.1
classified as Spanish)	Morton surname	258	114	44.2	25	14.8	139	2.5
	Buechley technique	270	120	44.4	19	11.2	139	2.5
25 through 34	"Narrow" surname	226	65	28.8	26	13.9	91	1.5
6,022 persons (187 self-	Census surname	261	101	38.7	27	14.4	128	2.1
classified as Spanish)	Morton surname	268	107	39.9	26	13.9	133	2.2
	Buechley technique	309	141	45.6	19	10.2	160	2.7
Over 34	"Narrow" surname	372	93	25.0	65	18.9	158	1.5
10,437 persons (344 self-	Census surname	422	155	36.7	77	22.4	232	2.2
classified as Spanish)	Morton surname	443	167	37.7	68	19.8	235	2.3
	Buechley technique	579	286	49.4	51	14.8	337	3.2

[a] Denominators used for these percentages were the numbers of persons classified as Spanish by each procedure for each age group.

[b] Denominators used for these percentages were the numbers of persons who classified themselves as Spanish in each age group.

[c] Denominators used for these percentages were the numbers of persons in each age group.

subgroups. If the ratios do vary substantially, then care must be exercised in correcting estimates of the number of Spanish-named persons according to one classification scheme so that they are comparable with the number of Spanish-named persons based upon a second classification scheme.

As may be seen from Table 6, there are some differences. The ratio of "narrow" surnamed persons to Census Spanish-surnamed individuals is particularly low outside New York, New Jersey, Florida, and the five southwestern states. The ratios of Spanish-surnamed persons by the Buechley technique to Census Spanish-surnamed persons vary considerably. Outside the southwest the ratios are particularly high. These ratios also depend upon AFQT percentile ($p<.01$) and age ($p<.05$).

The ratios of persons with Morton surnames to persons with Census surnames differ very little by population subgroup. This does not seem surprising when one considers that Morton used the 1960 Census surnames as a starting point for building his name list and that the Census Bureau subsequently reintroduced many of Morton's additions into the 1970 Census list of Spanish surnames. Although the number of names on Morton's list is still much larger than those on the 1970 Census list, Morton's additional names occur infrequently; thus the ratio of persons with Morton surnames to Census surnames is only slightly larger than 1.

The ratios of persons classified as Spanish to those who classify themselves as Spanish are shown in Table 7. The ratios depend upon geographic area ($p<.01$), educational level ($p<.05$), and AFQT level ($p<.01$).

DISCUSSION

CHOOSING A SUITABLE CLASSIFICATION TECHNIQUE

Anyone who has built a list of Spanish surnames has probably faced the embarrassment of finding obvious Spanish surnames not on his list. Perhaps for this reason most classification schemes err on the side of being too inclusive.

It seems clear from these data that for general statistical purposes the best computerized procedure for classifying names as Spanish or non-Spanish is a procedure based on a "narrow" definition of Spanish. This leads to fewer overall misclassifications and, more importantly, the Spanish group includes a smaller portion of persons who are not actually Spanish.

Three caveats should be attached to this conclusion. First, it should be pointed out that computerized coding is not the only alternative. In theory, manual coding can have fewer misclassifications than the computerized techniques involved here, since additional information such as accent marks or names of relatives can be utilized. A manual coder can also accept name variations (the Buechley technique can normally handle name variations, but the other surname techniques cannot). However, comparing the results for five southwestern states (Table 2) with Buechley's California results (11), it appears that manual coding using the 1970 Census list is less accurate than computerized coding. The problem was not in falsely classifying non-Spanish as Spanish. The results in Buechley's study and in the present study were not significantly different in this respect (Table 8).

The manual techniques appear, however, to misclassify substantially more Spanish as non-Spanish, as shown in Table 9. Buechley notes that clerical coding errors of this type are especially common with names that do not "look" very Spanish.

The second caveat is that "narrow" surname classification is best only at this point in time. It is quite possible that the Buechley technique may be improved so that the high proportion of those falsely classified as Spanish may be reduced.[3] The Buechley

[3] Buechley is, in fact, planning a revised version of his Spanish-surname recognition program.

Table 6

Ratios of Number of Persons Classified as Spanist According to Different Classification Procedures, by Population Subset

Population Subset	Persons With "Narrow" Surnames To Persons With Census Surnames	Persons With Morton Surnames To Persons With Census Surnames	Persons With Buechley Surnames To Persons With Census Surnames	Persons With Buechley Surnames To Persons With ' Narrow" Surnames
Five southwestern states	.99	1.02	1.09	1.11
New York, New Jersey, Florida	.90	1.17	1.42	1.60
Other areas	.62	1.09	1.48	2.38
State unknown	.94	1.06	1.32	1.41
Persons with 12 or fewer years of education	.88	1.06	1.25	1.41
Persons with over 12 years of education	.92	1.04	1.44	1.57
AFQT Percentile ≤ 33	.94	1.07	1.31	1.39
AFQT Percentile 34-67	.88	1.08	1.29	1.46
AFQT Percentile > 67	.78	1.06	1.44	1.86
AFQT Unknown	.92	1.05	1.16	1.26
Under 25 years of age	.93	1.13	1.18	1.28
Ages 25-34	.87	1.03	1.18	1.37
Over 34 years of age	.88	1.05	1.37	1.56
Total	.89	1.06	1.27	1.43

Table 7

Ratios of Persons Classified as Spanish to Those Classifying Themselves as Spanish, by Population Subset and Coding Technique

Population Subset	Coding Technique			
	"Narrow" Surname	Census Surname	Morton Surname	Buechley Technique
Five southwestern states	1.03	1.04	1.06	1.14
New York, New Jersey, Florida	1.21	1.36	1.59	1.93
Other areas	1.54	2.47	2.69	3.67
State unknown	1.26	1.34	1.42	1.77
Persons with 12 or fewer years of education	1.13	1.28	1.36	1.60
Persons with over 12 years of education	1.48	1.61	1.67	2.33
AFQT Percentile ≤ 33	1.03	1.10	1.17	1.43
AFQT Percentile 34-67	1.11	1.26	1.36	1.62
AFQT Percentile > 67	1.38	1.78	1.89	2.57
AFQT unknown	1.19	1.29	1.35	1.49
Under 25 years of age	1.25	1.35	1.53	1.60
Ages 25-34	1.21	1.39	1.43	1.65
Over 34 years of age	1.08	1.23	1.29	1.68
Total	1.16	1.30	1.39	1.65

Table 8

Number of Persons Falsely Classified[a] as Spanish by Study and Coding Techniques

Coding Technique	Buechley's Study (Manual Coding Using Census Surnames)	Present Study (Computerized Coding Using Census Surnames)
Buechley technique	46	88
Census surnames	38[b]	72

$\chi^2 = 0.0$ (1 *df*); $p < .01$.

[a] In Buechley's study, a false classification was determined by inspection of the names classified as Spanish.

[b] Buechley gives this number as 40, since he believed that two of the names on the 1970 census list were not Spanish. In order to get a valid comparison of manual and computerized techniques, it is necessary to not count these as errors.

Table 9

Number of Persons Falsely Classified[a] as Non-Spanish by Study and Coding Technique

Coding Technique	Buechley's Study (Manual Coding Using Census Surnames)	Present Study (Computerized Coding Using Census Surnames)
Buechley technique	52	34
Census surnames	223	56

$\chi^2 = 41.7$ (1 *df*): $p < .01$.

[a] In Buechley's study a false classification was determined by inspection of the names classified as Spanish.

technique already has two advantages in that it does not require an alphabetic sort of the surnames to be classified, and it has fewer misclassified as non-Spanish.

The third caveat is that for some purposes it may be desirable to use Spanish names only as a means of restricting attention to a group who *may* be "Spanish". The definitive assessment of ethnicity is determined by a follow-up of individuals whose names are treated as Spanish by the computerized coding technique. In this case, a more inclusive coding technique (e.g., the Buechley technique) has clear advantages.

DEFINITIVE LIST OF SPANISH SURNAMES

It should be mentioned that the list of "narrow" Spanish surnames used here or any known list cannot be considered definitive. There probably are names not on the list which should be, and vice versa.

Interestingly, there is a simple and completely automated procedure for building a definitive list. Unfortunately the procedure requires a very large magnetic tape file of the names of persons living in the United States. The definitive list could be constructed simply by accepting only those surnames possessed by persons who in a high percentage of cases have Spanish first names.

USE OF SPANISH SURNAMES OUTSIDE THE SOUTHWEST

Outside the Southwest, the proportion of Spanish-surnamed persons in the study who did not classify themselves as "Spanish or Mexican American" was so large that one must ask whether Spanish-surname classification in those areas has any merit at all. If, for example, a study were conducted to determine the income levels of Spanish-surnamed college graduates in Minneapolis, probably only a small percent of the study group would be culturally Spanish.

Whether the situation is as serious as the figures in Table 2 suggest is not clear. It would seem that the Air Force sample represents a more assimilated group than the population of Spanish-surnamed persons living in the United States. Also, there are culturally Spanish persons, particularly Puerto Ricans, who would not want to classify themselves as "Spanish or Mexican Americans." Nevertheless, the apparent number of misclassifications is so large that one must proceed with caution, at least until further studies can examine the backgrounds of Spanish-surnamed persons living outside the Southwest.

The Bureau of the Census, incidentally, has long contended that Spanish-surname classification would not hold up outside the Southwest. This situation may change, however, as more Hispaños inhabit those areas. Even now there are undoubtedly local areas outside the Southwest where the correspondence between Spanish surname and Spanish culture is strong.

Also it should be mentioned that there are study designs where the poor specificity of Spanish surname classification can be tolerated. For example, if in certain areas one finds employers of blue collar workers who have no persons of Spanish surname on their payrolls, there would be good evidence of discriminatory employment practices. The poor specificity of Spanish-surname classification, in such a case, becomes a problem in the opposite direction. It is possible to have discriminatory employment practices, and still employ a substantial number of persons with Spanish surnames.

EFFECT OF BIAS AND OTHER PROBLEMS IN THE DATA

The problems of ethnic classification using Spanish surnames are serious enough that it may well be asked whether some idiosyncrasies in our data or its treatment might have magnified the problems.

The most obvious bias in the data occurs because Air Force enlisted men are not an unbiased sample of the U.S. population. Further bias arises from non-response to the survey and from the requirement to match the master file on social security number and Air Force Specialty Code.

The bias caused by requiring a match on the Air Force Specialty needs no speculation. The results with and without the Air Force Specialty Code match are shown in Table 10. This match eliminated persons who were not conscientiously completing their forms and possibly a small set of miscoded social security numbers which found matching cases in the master file. Without the match, the apparent problems in the use of Spanish-surname classifications would increase.

The effect of most other forms of bias would cause the Spanish-named persons among the survey respondents to represent a more assimilated group than people in the general population. The only effect of the bias is to restrict the range of assimilation in the survey data. This could have the effect of increasing the proportion of persons misclassified as Spanish in the study, but it should not create the differences observed between population subgroups.

Table 10

Comparison of Misclassification by Coding Technique With and Without Matching on the Air Force Specialty Code (AFSC)

	Percent Falsely Classified as Spanish[a]		Percent Falsely Classified as Non-Spanish[b]	
Coding Technique	With AFSC Match	Without AFSC Match	With AFSC Match	Without AFSC Match
"Narrow" surname	28.3	30.3	16.9	38.8
Census surname	38.2	40.1	19.4	41.1
Broad surname	40.1	41.8	17.1	39.2
Buechley technique	47.3	48.7	12.8	35.9

[a] The denominators used for these percentages were the numbers of persons classified as Spanish by each coding technique.

[b] The denominators used for these percentages were the numbers of persons who classified themselves as Spanish.

The most ticklish problem in the data occurs because of the late appearance of the ethnic question in the survey—the 52nd question in a survey of 143 questions. It is possible that by this stage a sizeable portion of persons were not conscientiously completing the questionnaire.

Nonconscientious marking would in effect create noise in the data. This noise should not create the geographic differences in the proportion of Spanish-surnamed persons who classified themselves as Spanish; however, it could have a substantial effect on the proportion of those marking "Spanish or Mexican American" on the questionnaire who did not have Spanish names. The difference is that the Spanish-surnamed population does not depend on the survey results for its definition; however, the population of those indicating Spanish on the survey does depend on survey results.

The type of effects that rote marking might have on the results may best be seen from a separate Air Force survey. In the airman survey of July 1971 the same ethnic question was asked as the 105th of 150 questions, a placement much later than the 52nd of 143 questions in the March survey. A comparison of the results of the two surveys is shown in Table 11. While the percent falsely classified as Spanish is approximately the same in both surveys, the percent falsely classified as non-Spanish differs substantially.

To provide a more realistic estimate of the persons misclassified as non-Spanish, it is necessary to correct the tabulations in some way. This was done by assuming that among S_1 (the set of persons identifying themselves as Spanish) and S_2 (the subset of S_1 having "narrow" Spanish surnames), Pf (the proportion of persons having "narrow" Spanish first names) should be the same. Any deficit of Pf_1 in S_1 under Pf_2 in S_2 would be attributed to careless marking. The number N_c classifying themselves as Spanish through carelessness may then be estimated by:

$$N_c = N_1 \ (1 - \frac{Pf_1}{Pf_2})$$

where N_1 is the number of cases in S_1. By subtracting N_c from both numerator and denominator, adjusted estimates may be calculated for the percent of persons falsely classified as non-Spanish. The same procedure may be followed within each geographic area, AFQT group, and educational level. The results are shown in Tables 12 and 13.

Table 11

Percent of Persons Falsely Classified as Spanish and Non-Spanish by Population Subset and Survey

Population Subset	Percent Falsely Classified as Spanish[a]		Percent Falsely Classified as Non-Spanish[b]	
	March	July	March	July
All areas	28.3	28.0	16.9	27.3
Five southwestern states	14.5	16.1	11.9	13.2
New York, New Jersey, Florida	37.1	36.4	24.1	33.0
Other areas	55.0	61.1	30.8	59.3
State unknown	35.3	32.5	18.5	37.9
AFQT-33	19.7	22.6	17.5	25.3
AFQT 34-67	28.3	28.5	20.4	32.0
AFQT-67	37.9	32.8	14.4	31.2
AFQT unknown	28.4	28.8	15.1	21.7
Years of school ⩽ 12	26.7	26.8	17.1	27.2
Years of school > 12	42.9	39.2	15.4	28.4

[a] The denominators used for these percentages were the numbers of persons in the population subsets who had a "narrow" Spanish surname.

[b] The denominators used for these percentages were the numbers of persons in the population subsets who classified themselves as Spanish on the survey.

Table 12 shows that the adjustment procedure does a credible job of explaining differences between the unadjusted results of the March and July surveys.

Table 13 shows that while the Buechley technique is still the most inclusive of the Spanish-surname classification procedures, it nevertheless misses almost 8% of those persons classifying themselves as Spanish. The 8% estimate is, if anything, low, since it assumes that those who do not have Spanish surnames are as apt to have Spanish first names as those who do have Spanish surnames. The assumption may not be entirely true.

Tables 12 and 13 also show that the proportion misclassified as non-Spanish depends upon the geographic area ($p<.005$) but does not depend on either the AFQT or educational levels ($p>.05$). One must, of course, view these results cautiously because of the indirect procedure used in creating Tables 12 and 13.

SUMMARY AND CONCLUSIONS

Several computerized procedures for classifying names as Spanish or non-Spanish were compared, using Air Force survey data. The results of each classification procedure were compared with the classifications selected by respondents to the survey. The conclusions were as follows:

(1) Outside five southwestern states, Spanish name classifications included enough persons who did not consider themselves Spanish that the usefulness of the technique for these areas is seriously reduced.

(2) At higher educational levels and AFQT percentiles, all the surname classification procedures included increasing proportions of persons who did not consider themselves Spanish.

Table 12

Adjusted Number and Percent of Persons Falsely Classified as Non-Spanish by Population Subset and Survey

	March Survey			July Survey			March and July Combined		
	Number			Number			Number		
Population Subset	Self-Classified as Spanish	Without Spanish Surname[a]	Percent	Self-Classified as Spanish	Without Spanish Surname[a]	Percent	Self-Classified as Spanish	Without Spanish Surname[a]	Percent
Five southwestern states	364.51	23.51	6.5	367.86	38.86	10.6	731.58	61.58	8.4
New York, New Jersey, Florida	78.00	12.00	(15.4)	76.12	13.12	(17.2)	153.89	24.89	16.2
Other area	61.71	7.71	(12.5)	40.08	3.08	(7.7)	101.50	10.50	(10.3)
State unknown	150.33	27.33	18.2	133.16	23.16	17.4	283.38	50.38	17.8
AFQT ≤ 33	128.35	10.35	(8.1)	141.94	14.94	(10.5)	269.96	24.95	9.2
AFQT 34-67	175.38	23.38	13.3	170.84	17.85	10.4	345.86	40.85	11.8
AFQT > 67	80.08	3.08	(3.8)	113.52	27.52	24.2	192.34	29.34	15.3
AFQT unknown	267.25	30.26	11.3	194.29	21.29	11.0	461.57	51.57	11.2
Years of School ≤ 12	603.08	63.08	10.5	555.80	66.80	12.0	1158.59	129.59	11.2
Years of School > 12	47.14	3.14	(6.7)	59.29	11.29	(19.0)	106.84	14.84	(13.9)
Total	654.55	70.55	10.8	617.22	78.22	12.7	1270.35	147.35	11.6

[a] These columns give the adjusted number of persons who marked "Spanish or Mexican American" on the survey but who did not have a "narrow" Spanish surname.

Note: Parentheses ( ) indicate that percentage is based upon a numerator less than 15.

Table 13

Adjusted Number and Percent of Persons Falsely Classified as Non-Spanish by Population Subset and Coding Technique[a]

(March and July Surveys Combined)

	Census Surnames			Morton Surnames			Buechley Technique		
	Number			Number			Number		
Population Subset	Self-Classified as Spanish	Without Spanish Surname	Percent	Self-Classified as Spanish	Without Spanish Surname	Percent	Self-Classified as Spanish	Without Spanish Surname	Percent
Five southwestern states	728.60	73.60	10.1	735.48	54.48	7.4	743.03	40.03	5.4
New York, New Jersey, Florida	150.84	28.84	19.1	150.32	24.32	16.2	160.20	21.20	13.2
Other area	102.62	10.62	(10.3)	103.73	10.73	(10.3)	104.85	10.85	(10.3)
State unknown	280.29	62.29	22.2	280.95	49.95	17.8	279.00	31.00	11.1
AFQT ⩽ 33	274.52	41.52	15.1	278.80	32.80	11.8	273.38	18.38	6.7
AFQT 34-67	343.37	51.37	15.0	345.04	38.04	11.0	356.03	28.03	7.9
AFQT > 67	195.06	33.06	16.9	194.70	29.70	15.3	194.02	23.02	11.9
AFQT unknown	447.50	47.50	10.6	450.77	37.77	8.4	463.67	33.67	7.3
Years of school ⩽ 12	1149.76	151.76	13.2	1160.42	119.42	10.3	1177.86	94.86	8.1
Years of School > 12	108.00	21.00	19.4	105.60	17.60	16.7	104.82	5.82	(5.6)
Total	1258.63	171.63	13.6	1267.14	136.14	10.7	1284.88	100.88	7.9

[a] See Table 12 for comparable figures when "narrow" surname classification is used.

Note: Parentheses ( ) indicate that percentage is based upon a numerator less than 15.

(3) Even for the most inclusive surname classification technique, the portion of Spanish persons who are missed is estimated to be 8% or higher.

(4) There is some evidence that more persons self-identified as Spanish are missed by the surname classification procedures outside the Southwest.

(5) The best classification procedure for general statistical purposes, the "narrow" surname technique, required a more exclusive list of Spanish surnames than has generally been used. This procedure had fewer overall misclassifications and the resulting Spanish group contained fewer persons who did not consider themselves Spanish.

(6) Future research efforts are outlined to:

(a) Produce a more definitive list of Spanish surnames.

(b) Explore improvements in Buechley's technique of classifying Spanish names.

(c) Further examine persons of Spanish surname and culture who do not classify themselves as "Spanish or Mexican Americans."

LITERATURE CITED

1. Bureau of the Census. *Minority-Owned Businesses: 1969*, August 1971.

2. Weaver, Charles N. "A Comparative Study of Job Performance of Spanish-Surnamed Police Officers in San Antonio, Texas," *Phylon*, No. 30, 1969, pp. 27-33.

3. Morton, William E. "Demographic Redefinition of Hispaños," *Public Health Reports*, No. 85, 1970, pp. 617-623.

4. Christiansen, T., and Livermore, G. "A Comparison of Anglo-American and Spanish-American Children on the WISC," *Journal of Social Psychology*, No. 81, 1970, pp. 9-13.

5. Bates, W., Dubeck, P., and Redlinger, L. "The Social Context of Urban Narcotic Use," *Proceedings of the Southwestern Sociological Association*, No. 19, 1969, pp. 199-205.

6. Jackson, E.J., and Buechley, Robert W. "Study of Vital Statistics in California Through Identification of Spanish Surnames," *Boletin de la Officina Sanitaria Panamerica*, No. 69, 1970, p. 436.

7. Cabinet Committee on Opportunity for the Spanish Speaking. *Spanish Surnamed American College Graduates, 1971-1972*, 1972.

8. Bureau of the Census. *1970 Census General Coding Procedures Manual*, Attachment J2.

9. Morton, William E. *U.S. Bureau of the Census List of Spanish Surnames, Revised as of October 20, 1967*, University of Oregon Medical School, January 1968.

10. Buechley, Robert W. "A Reproducible Method of Counting Persons of Spanish Surname," *Journal of the American Statistical Association*, No. 56, 1961, pp. 88-97.

11. Buechley, Robert W. "A Computer Program that Distinguishes Spanish Surnames" (unpublished).

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER HumRRO-PP-9-73	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A COMPARISON OF COMPUTERIZED TECHNIQUES FOR RECOGNIZING SPANISH NAMES		5. TYPE OF REPORT & PERIOD COVERED Professional Paper 6. PERFORMING ORG. REPORT NUMBER Professional Paper 9-73
7. AUTHOR(S) G. Lee Giesecke		8. CONTRACT OR GRANT NUMBER(S) DAHC 15-73-C-0131
9. PERFORMING ORGANIZATION NAME AND ADDRESS Human Resources Research Organization (HumRRO) 300 North Washington Street Alexandria, Virginia 22314		10. PROGRAM ELEMENT PROJECT TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Directorate for Manpower Research Office of the Assistant Secretary of Defense (Manpower and Reserve Affairs)		12. REPORT DATE October 1973 13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Research performed by HumRRO Division No. 7 (Social Science). Report also issued as Manpower Research Report 73-2, Office of the Secretary of Defense, October 1973.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Classification procedures	Ethnic classification
Coding techniques	Name classification
Computer technology	Spanish surnames

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study was performed to show the validity, or lack of it, of various coding techniques used to identify persons of Spanish derivation. The results of computerized methods to identify Spanish names are compared with responses to questionnaires in which people identified themselves as Spanish. Outside of five southwestern states and at higher educational and aptitude levels, the name recognition procedures include increasing proportions of persons who do not classify themselves as Spanish. This

(Continued)

20. (Continued)

problem is mitigated by using a more restrictive list of Spanish surnames than has been used previously.