RESEARCH REPORT SERIES
(Statistics #2012-02)

Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations

Patrick M. Joyce
Donald Malec[1]
Roderick A. Little
Aaron Gilary

[1] National Center for Health Statistics

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: January 24, 2012
Revised: February 16, 2012

Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations

Patrick M. Joyce[1], Donald Malec[2], Roderick A. Little[1], Aaron Gilary[1]

[1]U.S. Census Bureau
[2]National Center for Health Statistics

February 16, 2012

Abstract

Section 203 of the Voting Rights Act relates to provisions requiring the use of election materials in languages other than English for states or political subdivisions when their voting age populations have certain characteristics. Data from the 2010 Census and the American Community Survey (ACS) provide counts and estimates for these characteristics. The ACS is a general purpose sample survey designed to produce a large volume of estimates across the spectrum of the nation's geographic areas and subgroups of the population. This document describes a small area model which was developed and applied to provide improved estimates of characteristics for states and political subdivisions relating to citizenship, limited-English proficiency, and illiteracy. Methods of constructing point estimates under an empirical Bayes procedure are detailed. Methods for estimating the precision of the estimates are described in an appendix. This technical report presents models and methodology. Descriptions of data and analysis in support of and in assessment of modeling decisions made in producing the estimates will be supplied in a future report.

This page is intentionally left blank.

Contents

ACKNOWLEDGMENTS

The modeling development and resulting estimates were carried out by Patrick M. Joyce and Donald Malec with assistance from Roderick A. Little and Aaron Gilary. Estimation procedures were verified by Michael Starsinic and Rolando Rodriguez. The technical and policy details were reviewed by a Steering Committee: Alfredo Navarro, Roderick A. Little, William R. Bell, David W. Whitford, Melissa Creech, and Catherine McCully. Assistance for drafting this report was provided by Tommy Wright with comments from Joseph Schafer. Final review was provided by Alfredo Navarro, Lynn Weidman, Eric Slud, and Catherine McCully. The work of this project could not have been completed without the assistance of Mark Asiala, Michael Beaghen, Robert Creecy, Alina Kline, Keith Albright, John Jordan, Edward Castro, and Julie Tsay.

1 Introduction: Section 203 of the Voting Rights Act of 1965

This paper describes a statistical model to provide a method to estimate population quantities of interest relating to the Section 203 provisions of the Voting Rights Act. These provisions are stipulated in the U.S. Code, Title 42, Chapter 20, Subchapter 1-B, §1973aa-1a, entitled "Bilingual election requirements." Selected provisions that relate to actions to be taken by the U.S. Census Bureau are as follows:

> ***(a) Congressional findings and declaration of policy***
>
> *The Congress finds that, through the use of various practices and procedures, citizens of language minorities have been effectively excluded from participation in the electoral process. Among other factors, the denial of the right to vote of such minority group citizens is ordinarily directly related to the unequal educational opportunities afforded them resulting in high illiteracy and low voting participation. The Congress declares that, in order to enforce the guarantees of the fourteenth and fifteenth amendments to the United States Constitution, it is necessary to eliminate such discrimination by prohibiting these practices, and by prescribing other remedial devices.*
>
> ***(b) Bilingual voting materials requirement***
>
> > ***(1) Generally***
> >
> > *Before August 6, 2032, no covered State or political subdivision shall provide voting materials only in the English language.*
> >
> > ***(2) Covered States and political subdivisions***
> >
> > > ***(A) Generally***
> > >
> > > *A State or political subdivision is a covered State or political subdivision for the purposes of this subsection if the Director of the Census determines, based on the 2010 American Community Survey census data and subsequent American Community Survey data in 5-year increments, or comparable census data, that–*
> > >
> > > > *(i)(I) more than 5 percent of the citizens of voting age of such State or political subdivision are members of a single language*

minority and are limited-English proficient;

(II) more than 10,000 of the citizens of voting age of such political subdivision are members of a single language minority and are limited-English proficient; or

(III) in the case of a political subdivision that contains all or any part of an Indian reservation, more than 5 percent of the American Indian or Alaska Native citizens of voting age within the Indian reservation are members of a single language minority and are limited-English proficient; and

(ii) the illiteracy rate of the citizens in the language minority as a group is higher than the national illiteracy rate.

(B) Exception

The prohibitions of this subsection do not apply in any political subdivision that has less than 5 percent voting age limited-English proficient citizens of each language minority which comprises over 5 percent of the statewide limited-English proficient population of voting age citizens, unless the political subdivision is a covered political subdivision independently from its State.

(3) Definitions

As used in this section–

(A) the term "voting materials" means registration or voting notices, forms, instructions, assistance, or other materials or information relating to the electoral process, including ballots;

(B) the term "limited-English proficient" means unable to speak or understand English adequately enough to participate in the electoral process;

(C) the term "Indian reservation" means any area that is an American Indian or Alaska Native area, as defined by the Census Bureau for the purposes of the 1990 decennial census;

(D) the term "citizens" means citizens of the United States; and

(E) the term "illiteracy" means the failure to complete the 5th primary grade.

(4) Special Rule

The determinations of the Director of the Census under this subsection shall be effective upon publication in the Federal Register and shall not be subject to review in any court.

(c)-(d)[Not given here. See U.S. Code, Title 42, Chapter 20, Subchapter 1-B, §1973aa-1a.]

(e) Definitions

For purposes of this section, the term "language minorities" or "language minority group" means persons who are American Indian, Asian American, Alaska Native, or of Spanish heritage.

2 Statistical Inference Goals

The statutes of Section 203 require that the Census Bureau use data to determine if a particular jurisdiction (political subdivision) qualifies for language assistance (i.e. additional "voting materials"). If a jurisdiction qualifies for language assistance, it is said that the jurisdiction has coverage under Section 203. To determine coverage, estimates of interest relating to proportions of U.S. citizenship status, limited-English proficiency status, and illiteracy status for the jurisdiction's voting age population will be used. These terms are defined using the following working definitions:

A person is in the ***voting age population*** if that person is at least 18 years old. Alternatively, references may be made to ***voting age persons***. This is measured by the American Community Survey as well as the 2010 Census.

A person's U.S. citizenship is measured by the American Community Survey.

A person is ***English language proficient*** if the person speaks English "very well." Conversely, a person is ***limited-English proficient*** if the person speaks English at less than "very well." These concepts are measured by the American Community Survey.

A person is said to be ***illiterate*** if the person has less than a fifth grade education, i.e., that the person has only completed the fourth grade or lower. This is measured by the American Community Survey.

In the analysis, references will be primarily made to U.S. citizens (CIT), limited-English proficient (LEP) persons, and illiterate (ILL) persons of the voting age population (VAP). References to citizens of the voting age population (CVAP) will also be made.

In order to fulfill Section 203, we will make estimates for language minority groups (LMGs), which are race or ethnic groups that speak certain languages. See Table 1 for a list of LMGs. While the interest is to identify language minority groups, the data sources identify race and ethnicity. In particular, focus is placed on persons who are of Hispanic or Spanish heritage, Asian American, American Indian, or Alaska Native.

To frame the general problem, let N_{gjcli} denote the number of voting age persons such that

g is a language minority group designation. (See Table 1)
j is a jurisdiction. ($j = 1, ..., 7892$)
c represents U.S. citizenship status. (Yes=1 or No=0)
l represents limited-English proficiency status. (Yes=1 or No=0)
i represents illiteracy status. (Yes=1 or No=0)

This paper is concerned with assessing which jurisdictions are covered under Section 203, that is, which jurisdictions are required to provide language assistance for particular LMGs. Jurisdictions can qualify for language coverage at three levels: state level, jurisdiction level, and American Indian Area/Alaska Native Area (AIA/ANA) levels. We discuss each level in detail in the next three subsections.

2.1 State Level Coverage

State level Section 203 coverage is assessed by state level calculations. To be covered for language assistance for LMG g at the state level, the number of limited-English proficient CVAP LMG

Table 1: Language Minority Groups for which Tabulations and Determinations are Required Under Section 203

Hispanic	Houma
Asian Indian	Iroquois
Bangladeshi	Kiowa
Cambodian	Lumbee
Chinese	Menominee
Filipino	Mexican American Indian
Hmong	Navajo
Indonesian	Osage
Japanese	Ottawa
Korean	Paiute
Laotian	Pima
Malaysian	Potawatomi
Pakistani	Pueblo
Sri Lankan	Puget Sound Salish
Thai	Seminole
Vietnamese	Shoshone
Other Asian	Sioux
Apache	South American Indian
Arapaho	Spanish American Indian
Blackfeet	Tohono O'Odham
Canadian and French Indian	Ute
Central American Indian	Yakama
Cherokee	Yaqui
Cheyenne	Yuman
Chickasaw	All other AI tribes
Chippewa	AI tribes, not specified
Choctaw	Alaska Athabascan
Colville	Aleut
Comanche	Inupiat
Cree	Tlingit-Haida
Creek	Tsimshian
Crow	Yup'ik
Delaware	Alaskan Native Tribes, not specified
Hopi	AI or AN tribes, not specified

persons in state S, which we denote by $N_{gS11\cdot}$, divided by the CVAP population, denoted by $N_{\cdot S1\cdot\cdot}$, must be greater than five percent. In other words,

$$\frac{N_{gS11\cdot}}{N_{\cdot S1\cdot\cdot}} > 0.05 \tag{1}$$

which is an aggregate of jurisdictions in state S. It is also necessary that the percentage of illiterate persons among the voting age citizen limited-English proficient persons of LMG g in the state exceeds the national percentage of illiteracy amongst voting age citizens. In other words, we must also have

$$\frac{N_{gS111}}{N_{gS11\cdot}} > \frac{N_{\cdot\cdot1\cdot1}}{N_{\cdot\cdot1\cdot\cdot}} \tag{2}$$

Note that S stands for all jurisdictions in state S while j stands for a single jurisdiction in a state.

If the thresholds in (1) and (2) are met, the state qualifies for coverage, as does any subordinate jurisdiction j, unless

$$\frac{N_{gj11\cdot}}{N_{\cdot j1\cdot\cdot}} < 0.05. \tag{3}$$

2.2 Jurisdiction Level Coverage

Jurisdiction level Section 203 coverage is assessed by calculations at the jurisdiction level. A jurisdiction refers to each county within a state except for the following states: Connecticut, Maine, Massachusetts, Michigan, New Hampshire, Rhode Island, Vermont, and Wisconsin where a jurisdiction refers to a minor civil division (MCD). MCDs within these states typically represent cities and towns within the state. Additionally, Kalawao County, Hawaii is to be aggregated as part of Maui County, Hawaii.

For jurisdiction j, let $N_{\cdot j1\cdot\cdot}$ represent the number of voting age citizens, let $N_{gj11\cdot}$ represent the number of voting age citizens belonging to LMG g who are also limited-English proficient, and let N_{gj111} represent those voting age citizens who have limited-English proficiency and are illiterate who belong to LMG g. Then, jurisdiction j qualifies for Section 203 coverage for LMG g if the following holds:

$$\frac{N_{gj11\cdot}}{N_{\cdot j1\cdot\cdot}} > 0.05 \quad \text{OR} \quad N_{gj11\cdot} > 10,000 \tag{4}$$

and

$$\frac{N_{gj111}}{N_{gj11\cdot}} > \frac{N_{\cdot\cdot1\cdot1}}{N_{\cdot\cdot1\cdot\cdot}}. \tag{5}$$

That is to say, jurisdiction j must provide coverage for LMG g in the event that 1) the percentage of limited-English proficient members of the LMG g amongst voting age citizens exceeds five percent, or that the total number of voting age citizen limited-English proficient persons of LMG g exceeds 10,000, and 2) the percentage of illiterate persons amongst the voting age citizen limited-English proficient persons of LMG g exceeds the national percentage of illiteracy amongst voting age citizens.

2.3 American Indian Area & Alaska Native Area Level Coverage

American Indian Area and Alaska Native Area (AIA/ANA) level Section 203 coverage is assessed by making calculations across the particular AIA/ANA tribal area. Further, coverage is only assessed amongst AIA/ANA persons in each of these areas. For tribal area a, let A stand for the American Indian/Alaska Native (AIAN) population. A specific AIA/ANA tribal area a qualifies for coverage for LMG g (of those AIAN LMGs) if the following holds:

$$\frac{N_{ga11\cdot}}{N_{Aa1\cdot\cdot}} > 0.05 \tag{6}$$

and

$$\frac{N_{ga111}}{N_{ga11\cdot}} > \frac{N_{\cdot\cdot1\cdot1}}{N_{\cdot\cdot1\cdot\cdot}} \tag{7}$$

where the right hand side of (7) is the national CVAP illiteracy rate as computed in (2) and (5). From the U.S. Code for Section 203, it then follows that if an AIA/ANA qualifies for language coverage, then so does any jurisdiction which contains a part of that particular AIA/ANA. Note that the population of tribal areas for AIAN persons are related by the use of subscripts a and A; this descriptive notation is not utilized elsewhere within the document.

3 Methodological Overview

In this section, we give an overview of the technical details in Sections 4, 5, and 6.

If all of the counts and proportions specified in (1)-(7) were known for each of the 7,892 jurisdictions and each of the 50 states plus the District of Columbia, then the Director of the Census Bureau could easily determine Section 203 coverage.

The 2010 Census does give the count of voting age persons by jurisdictions and states for various language minority groups. For the census date, the number of voting age persons $N_{gj...}$ for $j = 1, ..., 7892$ in each of the 50 states plus the District of Columbia and the number of persons belonging to all language minority groups g are known. However, the 2010 Census does not provide the more detailed counts required by Section 203. Specifically, the 2010 Census does not collect: **(1)** U.S. citizenship status, **(2)** limited-English proficiency status, nor **(3)** illiteracy status. Hence, we do not know counts and proportions of voting age persons by jurisdiction (including AIA/ANA) and state for each language minority group for the specified detailed characteristics.

The American Community Survey (ACS) provides estimates of these quantities for all states and jurisdictions. However, the sampling error or uncertainty of the estimates of the characteristics needed for Section 203 is a weakness particularly for jurisdictions with small (ACS) samples within the period 2005-2009. This motivated the development of model-based estimates of the rates given above, which are detailed in Section 4.

The Need for Minority Estimation Groups

Even though we know the number of persons in each language minority group for each jurisdiction, some individuals belong to more than one language minority group (Table 1). This double-counting does not cause any serious problems. However, models that incorporate multiple language minority group membership for such persons would be very complex, and in fact, Census and ACS information about the numerous combinations of multiple language minority group membership is limited.

If we could assign each person to one and only one language minority group for the purpose of modeling, this modeling challenge goes away. For modeling purposes only, we introduce the concept of "minority estimation groups" which are mutually exclusive and exhaustive for the population. The seventy-four (74) minority estimation groups are given in Table 3. The method of assignment to a minority estimation group is described in Section 5.3.

Grouping Jurisdictions Before Modeling

Because some jurisdictions have more sample than others and because of the detailed information on U.S. citizenship and limited-English proficiency required for Section 203, direct estimates for jurisdictions are improved by combining them with estimates from groups of similar jurisdictions.

Groupings of similar jurisdictions are based on models of voting age persons that predict the following aspects of Section 203 requirements:

- the probability an individual is a U.S. citizen,
- the probability an individual is limited-English proficient given the person is a U.S. citizen, and
- the probability an individual is illiterate, given the person is a U.S. citizen who is limited-English proficient.

Detailed information about this is given in Section 4.1. The modeling is enhanced by using groupings of similar jurisdictions (see Section 6), as the model will be able to capture similarity in the probabilities which will increase the precision of the estimates.

The methodological details of the elements described in this section follow in Sections 4, 5, and 6.

4 Model Formulation

In order to make the Section 203 determinations, estimates are needed for the number of voting age U.S. citizens, voting age limited-English proficient U.S. citizens, and voting age illiterate limited-English proficient U.S. citizens. Information about these estimates is not collected in the 2010 Census but is collected of a sample of households and group quarters residents as part of the American Community Survey (ACS). Using estimates from the ACS, together with detailed counts of the voting age population from the 2010 Census, the Section 203 determinations can be made for all relevant language minority groups within political jurisdictions.

This section specifies the model that incorporates data from the 2005-2009 ACS 5-year data file to estimate the relevant rates of U.S. citizenship status, limited-English proficiency status, and illiteracy status that are then applied to the 2010 Census counts for making the determinations.

This section discusses what needs to be estimated and further describes the model that makes use of partitions of Census and ACS records based on race/ethnicity domains and jurisdictions. Lastly, we describe the application of the model to language minority groups in order to make Section 203 determinations.

Table 2: Identification of the Four Situations (Labeled 'A', 'B', 'C', and 'D' Which Represent Population Counts) Needed to Make the Determinations for Each Language Minority Group

	Citizen		Non-Citizen	
	LEP	Non-LEP	LEP	Non-LEP
ILL	A	C	D	
Non-ILL	B			

Three types of estimates for the voting age population are needed from the ACS based on the four situations illustrated in Table 2. From Table 2, we see **(1)** the proportion of the population that consists of U.S. citizens is represented by (A+B+C)/(A+B+C+D); **(2)** the proportion of U.S. citizens who have limited-English proficiency (LEP) is represented by (A+B)/(A+B+C); and, lastly **(3)** the proportion of the U.S. citizen LEP group who are illiterate (ILL) is represented by A/(A+B).

For the voting age population, we are interested in U.S. citizens; within those who are U.S. citizens, we are interested in those who are limited-English proficient; and within those who are U.S. citizens and limited-English proficient, we are interested in those who are illiterate. Likewise, we can express any of these parts with a relevant combination of terms. This will allow us to relate the three estimates of interest through a sequence of conditional probabilities. Combining this information with the specific language minority group information available at the jurisdiction level from the 2010 Census will shape the model described in Sections 4.1 and 4.1.3.

4.1 Models for U.S. Citizenship Status, LEP Status, and Illiteracy Status

The models presented in this section are based on the actual jurisdictions that will be used for the determinations. However, the actual language minority groups are not used directly in the model

because one person may belong to a number of different language minority groups. Instead, mutually exclusive and exhaustive race/ethnicity groups, termed "minority estimation groups" are used and given in Table 3.

Each minority estimation group is modeled separately. One consequence of modeling separately is that outcomes from one minority estimation group will have no effect on the estimates for another. However, the models are built to borrow information from other jurisdictions in the same minority estimation group. In order to borrow from only similar jurisdictions, classes of jurisdictions are formed and data are combined within but not across these classes (see Section 6). The classes formed vary in their jurisdictional composition based on each specific minority estimation group. Hierarchical models that relate to the classes and their jurisdictions are specified in Section 4.1.3. The first level of the model is at the jurisdiction by minority estimation group level. The second level of the model takes the jurisdictions and pools them by class. We provide some notation and details in Sections 4.1.1 and 4.1.2.

4.1.1 Notation

The subscript d denotes minority estimation group, and subscript j denotes jurisdiction (county-defined for some states and MCD defined for the remaining states). Note that (d, j) partitions both 2005-2009 ACS 5-year data and 2010 Census data.

Each of the models for U.S. citizenship status, LEP status, and illiteracy status begins with a specification of the following conditional probabilities (or proportions) at the individual level:

$$\begin{aligned} p_{dj}^{(C)} &= \text{Prob(a person is a U.S. citizen | voting age and } (d,j)) \\ p_{dj}^{(L)} &= \text{Prob(a person has limited-English proficiency | voting age, } (d,j), \text{ and U.S. citizen)} \\ p_{dj}^{(I)} &= \text{Prob(a person is illiterate | voting age, } (d,j), \text{ U.S. citizen, and limited-English proficient)} \end{aligned} \tag{8}$$

We can link $p_{dj}^{(C)}$, $p_{dj}^{(L)}$, and $p_{dj}^{(I)}$ to the 2010 Census population counts for each minority estimation group d, jurisdiction j, and race/ethnicity index $\underline{\nu}$ as defined in Table 4 for a given (ν, j).

Table 3: Hispanic, Asian, and AIAN LMG Minority Estimation Group Assignment Codes for Person Records.

Minority Estimation Group Assignment Codes			
Code		Code	
1	Mexican	38	Delaware
2	Puerto Rican	39	Hopi
3	Cuban	40	Houma
4	Central American, Dominican Republic	41	Iroquois
5	Latin/South American	42	Kiowa
6	Other Hispanic	43	Lumbee
7	Asian Indian	44	Menominee
8	Bangladeshi	45	Mexican American Indian
9	Cambodian	46	Navajo
10	Chinese	47	Osage
11	Filipino	48	Ottawa
12	Hmong	49	Paiute
13	Indonesian	50	Pima
14	Japanese	51	Potawatomi
15	Korean	52	Pueblo
16	Laotian	53	Puget Sound Salish
17	Malaysian	54	Seminole
18	Pakistani	55	Shoshone
19	Sri Lankan	56	Sioux
20	Thai	57	South American Indian
21	Vietnamese	58	Spanish American Indian
22	Other Asian	59	Tohono O'Odham
23	Apache	60	Ute
24	Arapaho	61	Yakama
25	Blackfeet	62	Yaqui
26	Canadian and French Indian	63	Yuman
27	Central American Indian	64	All other AI tribes
28	Cherokee	65	AI tribes, not specified
29	Cheyenne	66	Alaska Athabascan
30	Chickasaw	67	Aleut
31	Chippewa	68	Inupiat
32	Choctaw	69	Tlingit-Haida
33	Colville	70	Tsimshian
34	Comanche	71	Yup'ik
35	Cree	72	Alaskan Native Tribes, not specified
36	Creek	73	AI or AN tribes, not specified
37	Crow	74	No LMG Membership

Table 4: Notation for Language Minority Group/Minority Estimation Group Modeling

G:	total number of language minority groups (Table 1) and Hispanic subgroups needed for the determinations (Table 3).
$\underline{\nu}$:	$= (\nu_1, \ldots, \nu_G)$, a vector indicating presence, $\nu_g = 1$, or absence, $\nu_g = 0$, of language minority group (or Hispanic subgroup), g.
$d(\underline{\nu}, j)$:	The unique minority estimation group for a person with a language minority group pattern, $\underline{\nu}$, in jurisdiction j.
$c_{\underline{\nu}j}$:	The 2010 Census count of the voting age population in jurisdiction, j, with language minority group pattern, $\underline{\nu}$.

Using the notation of Table 4, ACS person records and 2010 Census person records are linked to the minority estimation groups, and then use the probabilities in (8) to link the data to the items of interest.

4.1.2 Models Linking Observed ACS Outcomes with $p_{dj}^{(C)}$, $p_{dj}^{(L)}$, and $p_{dj}^{(I)}$

Let n_{dj} be the total number of voting age person records observed for (d, j) within the ACS data. Suppose a simple random sample of n_{dj} voting age persons had been selected from the minority estimation group and jurisdiction combination (d, j), then let $m_{dj}^{(C)}$, $m_{dj}^{(L)}$, and $m_{dj}^{(I)}$ denote the sampled number of voting age U.S. citizens, the sampled number of voting age limited-English proficient U.S. citizens, and the sampled number of voting age illiterate limited-English proficient U.S. citizens, respectively. The proportions $p_{dj}^{(C)}$, $p_{dj}^{(L)}$, and $p_{dj}^{(I)}$ could be estimated from ACS data using the following simple binomial models:

$$\begin{aligned} m_{dj}^{(C)} &\sim \text{Binomial}(n_{dj}, p_{dj}^{(C)}) \\ m_{dj}^{(L)} &\sim \text{Binomial}(m_{dj}^{(C)}, p_{dj}^{(L)}) \\ m_{dj}^{(I)} &\sim \text{Binomial}(m_{dj}^{(L)}, p_{dj}^{(I)}). \end{aligned} \tag{9}$$

However, the ACS data are not collected via simple random sampling, but instead are collected via a stratified systematic sampling design with a two-phase component for initial non-responders and additional post-survey weighting adjustments. To address this complication, the counts $\{n_{dj}, m_{dj}^{(C)}, m_{dj}^{(L)}, m_{dj}^{(I)}\}$ are replaced by adjusted counts $\{\tilde{n}_{dj}, \tilde{m}_{dj}^{(C)}, \tilde{m}_{dj}^{(L)}\}$ given in (10) and $\{m_{dj}^{*(C)}, m_{dj}^{*(L)}, m_{dj}^{*(I)}\}$ given in (11) below that reflect these complex design features.

To describe the adjustments $\{\tilde{n}_{dj}, \tilde{m}_{dj}^{(C)}, \tilde{m}_{dj}^{(L)}\}$, let $\bar{p}_{dj}^{(C)}$, $\bar{p}_{dj}^{(L)}$, and $\bar{p}_{dj}^{(I)}$ denote the design-based ACS estimates of the three underlying proportions for each minority estimation group d in jurisdiction j, and let $\{\bar{p}_{d}^{(C)}$, $\bar{p}_{d}^{(L)}$, and $\bar{p}_{d}^{(I)}\}$ denote the corresponding design-based national estimates aggregated over jurisdictions. These design-based estimates follow the standard ACS estimation methodology [U.S. Census Bureau, 2009] where the response value (1, in this case) is multiplied by the weight and summed over the domain of interest, minority estimation group d. Also let $\bar{V}_{d}^{(C)}$, $\bar{V}_{d}^{(L)}$, and $\bar{V}_{d}^{(I)}$ denote the corresponding design-based variance estimates of $\{\bar{p}_{d}^{(C)}, \bar{p}_{d}^{(L)}, \bar{p}_{d}^{(I)}\}$, as in U.S. Census Bureau [2009]. The ACS variance estimation methodology uses a replication-weight based procedure with 80 replicates.

Adjusted sample sizes, $\tilde{n}_{dj}$, $\tilde{m}_{dj}^{(C)}$, and $\tilde{m}_{dj}^{(L)}$, are then calculated as follows:

$$\tilde{n}_{dj} = \begin{cases} n_{dj} \frac{\bar{p}_{d}^{(C)}(1-\bar{p}_{d}^{(C)})}{n_d \bar{V}_{d}^{(C)}} & \text{when } \bar{p}_{d}^{(C)}(1-\bar{p}_{d}^{(C)}) > 0 \\ n_{dj} & \text{when } \bar{p}_{d}^{(C)}(1-\bar{p}_{d}^{(C)}) = 0, \end{cases}$$

$$\tilde{m}_{dj}^{(C)} = \begin{cases} m_{dj}^{(C)} \frac{\bar{p}_{d}^{(L)}(1-\bar{p}_{d}^{(L)})}{m_{d}^{(C)} \bar{V}_{d}^{(L)}} & \text{when } \bar{p}_{d}^{(L)}(1-\bar{p}_{d}^{(L)}) > 0 \\ m_{dj}^{(C)} & \text{when } \bar{p}_{d}^{(L)}(1-\bar{p}_{d}^{(L)}) = 0, \end{cases} \tag{10}$$

and

$$\tilde{m}_{dj}^{(L)} = \begin{cases} m_{dj}^{(L)} \frac{\bar{p}_{d}^{(I)}(1-\bar{p}_{d}^{(I)})}{m_{d}^{(L)} \bar{V}_{d}^{(I)}} & \text{when } \bar{p}_{d}^{(I)}(1-\bar{p}_{d}^{(I)}) > 0 \\ m_{dj}^{(L)} & \text{when } \bar{p}_{d}^{(I)}(1-\bar{p}_{d}^{(I)}) = 0, \end{cases}$$

where n_d, $m_{d}^{(C)}$, and $m_{d}^{(L)}$ are the number of observed voting age ACS person records, the number of observed voting age U.S. citizens ACS person records, and the number of observed voting age U.S. citizens with limited-English proficiency ACS person records, respectively, at the national level for minority estimation group d. Note that in (10), whenever $\bar{p}(1-\bar{p}) = 0$, the actual sample size is used in order to avoid possibly extreme sample size adjustments in a case where the observed data are homogeneous.

Further, using $\bar{p}_{dj}^{(C)}$, $\bar{p}_{dj}^{(L)}$, and $\bar{p}_{dj}^{(I)}$, the corresponding design-based estimates for (d, j), set:

$$\begin{aligned} m_{dj}^{*(C)} &= \tilde{n}_{dj} \bar{p}_{dj}^{(C)} \\ m_{dj}^{*(L)} &= \tilde{m}_{dj}^{(C)} \bar{p}_{dj}^{(L)} \\ m_{dj}^{*(I)} &= \tilde{m}_{dj}^{(L)} \bar{p}_{dj}^{(I)}. \end{aligned}$$

The triple $(m_{dj}^{*(C)}, m_{dj}^{*(L)}, m_{dj}^{*(I)})$ then serves as the data from ACS for our model. The approximate "design adjusted" binomial models used for all estimation are

$$\begin{aligned} m_{dj}^{*(C)} &\sim \text{Binomial}(\tilde{n}_{dj}, p_{dj}^{(C)}) \\ m_{dj}^{*(L)} &\sim \text{Binomial}(\tilde{m}_{dj}^{(C)}, p_{dj}^{(L)}) \\ m_{dj}^{*(I)} &\sim \text{Binomial}(\tilde{m}_{dj}^{(L)}, p_{dj}^{(I)}). \end{aligned} \tag{11}$$

The adjustments of the sample sizes in (10) and the adjustment of binomial response in (11) are constructed so that we are able to match design-based point estimates and variances. In (10) the intent is to make a variance adjustment for each minority estimation group d in jurisdiction j for each portion of the binomial model treating each level separately. The most straight-forward way to do this given the data challenges that exist for some minority estimation groups is to make a national-level design-effect adjustment and then apply these to the local jurisdictions. The national-level gives us a guarantee that we have adequate information for a design-effect whereas lower levels of geography may not have enough information to reliably provide a design-effect for certain minority estimation groups. In (11) we want to use and match the design-based point estimate for the minority estimation group d in jurisdiction j so that we use the information provided by differential weights within the jurisdiction. As a result, we have made an adjustment so that the sample size reflects the variance and the approximate binomial response reflects jurisdictional level information.

4.1.3 The Hierarchical Model

The ACS was not designed to provide the precise design-based language minority group by jurisdiction estimates needed for the Section 203 determinations described in Section 2. To address this problem, the jurisdictions are grouped into classes of similar jurisdictions, as described in Section 6. Hierarchical models are used to combine the direct "design-based estimates" within each class with pooled estimates from the set of similar jurisdictions in order to improve the quality of the estimates. The hierarchical models used here extend the jurisdiction level model of minority estimation groups specified in equation (11).

For each minority estimation group d, jurisdictions may be partitioned into classes in different ways, as the notation will show. For minority estimation group d, define:

$B_d^{(b)}$: number of classes of jurisdictions formed (Section 6) based on minority estimation group d for the model level b, for $b = C$, L, and I.

$h_b(d, j) \in \{1, \ldots, B_d^{(b)}\}$: identifies the class that jurisdiction j is in (Section 6), with respect to the estimation of minority estimation group d for the model level b, for $b = C$, L, and I.

The hierarchical models are defined by the binomial models (11) for the design-adjusted counts where the probabilities are assumed to follow Beta distributions:

$$\begin{aligned} p_{dj}^{(C)} &\sim \text{Beta}(\mu_{dh_C(d,j)}^{(C)} M_{dh_C(d,j)}^{(C)}, (1 - \mu_{dh_C(d,j)}^{(C)}) M_{dh_C(d,j)}^{(C)}) \\ p_{dj}^{(L)} &\sim \text{Beta}(\mu_{dh_L(d,j)}^{(L)} M_{dh_L(d,j)}^{(L)}, (1 - \mu_{dh_L(d,j)}^{(L)}) M_{dh_L(d,j)}^{(L)}) \\ p_{dj}^{(I)} &\sim \text{Beta}(\mu_{dh_I(d,j)}^{(I)} M_{dh_I(d,j)}^{(I)}, (1 - \mu_{dh_I(d,j)}^{(I)}) M_{dh_I(d,j)}^{(I)}) \end{aligned} \tag{12}$$

where μ represents the mean of the Beta distribution and M represents a scaling parameter such that the variance is $\frac{\mu(1-\mu)}{M+1}$. The parameters μ and M are related to the standard parameters α and β from the Beta(α, β) distribution by the expressions $\mu = \frac{\alpha}{\alpha+\beta}$ and that $M = \alpha+\beta$.

Note that although the use of the $h_b(d, j)$ notation identifies which jurisdictions share the same class, we can establish a less burdensome notation by defining $h_b = h_b(d, j)$ and rewriting the models in (12) as:

$$\begin{aligned} p_{dj}^{(C)} &\sim \text{Beta}(\mu_{dh_C}^{(C)} M_{dh_C}^{(C)}, (1 - \mu_{dh_C}^{(C)}) M_{dh_C}^{(C)}), \\ p_{dj}^{(L)} &\sim \text{Beta}(\mu_{dh_L}^{(L)} M_{dh_L}^{(L)}, (1 - \mu_{dh_L}^{(L)}) M_{dh_L}^{(L)}), \\ p_{dj}^{(I)} &\sim \text{Beta}(\mu_{dh_I}^{(I)} M_{dh_I}^{(I)}, (1 - \mu_{dh_I}^{(I)}) M_{dh_I}^{(I)}). \end{aligned} \tag{12$'$}$$

4.2 Empirical Bayes Estimates of Proportions for Minority Estimation Groups

The empirical Bayes estimates for the proportions $p_{dj}^{(C)}$, $p_{dj}^{(L)}$, and $p_{dj}^{(I)}$ given in (8) using the models and data of (11) and (12$'$) have the following forms:

$$\begin{aligned} \hat{p}_{dj}^{(C)} &= \omega_{dj}^{(C)} \bar{p}_{dj}^{(C)} + (1-\omega_{dj}^{(C)}) \hat{\mu}_{dh_C}^{(C)} \\ \hat{p}_{dj}^{(L)} &= \omega_{dj}^{(L)} \bar{p}_{dj}^{(L)} + (1-\omega_{dj}^{(L)}) \hat{\mu}_{dh_L}^{(L)} \\ \hat{p}_{dj}^{(I)} &= \omega_{dj}^{(I)} \bar{p}_{dj}^{(I)} + (1-\omega_{dj}^{(I)}) \hat{\mu}_{dh_I}^{(I)} \end{aligned} \tag{13}$$

where: $\bar{p}_{dj}^{(C)} = \frac{m_{dj}^{*(C)}}{\tilde{n}_{dj}}$, $\bar{p}_{dj}^{(L)} = \frac{m_{dj}^{*(L)}}{\tilde{m}_{dj}^{(C)}}$, and $\bar{p}_{dj}^{(I)} = \frac{m_{dj}^{*(I)}}{\tilde{m}_{dj}^{(L)}}$ are the design-based estimates for each U.S. citizenship rate, LEP rate, and illiteracy rate, respectively, and, correspondingly, $\omega_{dj}^{(C)} = \frac{\tilde{n}_{dj}}{\hat{M}_{dh_C}^{(C)} + \tilde{n}_{dj}}$, $\omega_{dj}^{(L)} = \frac{\tilde{m}_{dj}^{(C)}}{\hat{M}_{dh_L}^{(L)} + \tilde{m}_{dj}^{(C)}}$, and $\omega_{dj}^{(I)} = \frac{\tilde{m}_{dj}^{(L)}}{\hat{M}_{dh_I}^{(I)} + \tilde{m}_{dj}^{(L)}}$. Lastly, $\hat{\mu}_{dh_C}^{(C)}$, $\hat{\mu}_{dh_L}^{(L)}$, $\hat{\mu}_{dh_I}^{(I)}$, $\hat{M}_{dh_C}^{(C)}$, $\hat{M}_{dh_L}^{(L)}$, and $\hat{M}_{dh_I}^{(I)}$ are the maximum likelihood estimates (MLEs) based on the distributions specified in (11) and (12′) such that

$$(\hat{\mu}_{dh}, \hat{M}_{dh}) = \arg\max_{\mu, M} \prod_{j:h_b(d,j)=h} \int_0^1 \frac{\Gamma(\mu M)\Gamma((1-\mu)M)}{\Gamma(M)} p_{dj}^{\mu M + x_{dj} - 1} (1-p_{dj})^{((1-\mu)M + y_{dj} - x_{dj} - 1)} dp_{dj} \tag{14}$$

where y_{dj} stands for the Binomial trial sizes on the right-hand side of (11) and x_{dj} stands for the numbers m^* on the left-hand side of (11).

Note that the estimates in (13) must be adjusted when $\tilde{n}_{dj} = 0$, $\tilde{m}_{dj}^{(C)} = 0$, or $\tilde{m}_{dj}^{(L)} = 0$ because this results in cases where $\bar{p} = \frac{0}{0}$. Specifically

$$\begin{aligned} & \text{if} \quad \tilde{n}_{dj} = 0 \quad \text{then we set} \quad \hat{p}_{dj}^{(C)} = \hat{\mu}_{dh_C}^{(C)}, \\ & \text{if} \quad \tilde{m}_{dj}^{(C)} = 0 \quad \text{then we set} \quad \hat{p}_{dj}^{(L)} = \hat{\mu}_{dh_L}^{(L)}, \\ \text{and} \quad & \text{if} \quad \tilde{m}_{dj}^{(L)} = 0 \quad \text{then we set} \quad \hat{p}_{dj}^{(I)} = \hat{\mu}_{dh_I}^{(I)}. \end{aligned}$$

The empirical Bayes estimates of the proportions (13) combine the design-based estimates with the corresponding model-based estimates of the means within the same class. The design-based estimate receives relatively more weight in (13) if the sample size (e.g., $\tilde{m}_{dj}^{(C)}$) is large or if the within class variability among jurisdictions (e.g., $M_{dh_L}^{(L)}$) is small.

For more background on the use of empirical Bayes estimates and for a more detailed account of the derivation of the estimates in (13), see Section 3.3.2 of Carlin and Louis [2000].

4.3 Empirical Bayes Estimates of the Counts for the Determinations of Specific Language Minority Groups

In Section 4.1, we noted that a person may belong to more than one language minority group. Each minority language group vector $\underline{\nu}$ within each jurisdiction j is mapped to a unique minority

estimation group $d = d(\underline{\nu}, j)$, as discussed in Section 4.1.1. Estimates for a particular language minority group then use the empirical Bayes estimated proportions for the minority estimation group to which that language minority group is mapped. Specifically, for language minority group vector $\underline{\nu}$ in jurisdiction j, we apply the U.S. citizenship parameter $p^{(C)}_{d(\underline{\nu},j)j}$, limited-English proficiency given U.S. citizenship parameter $p^{(L)}_{d(\underline{\nu},j)j}$, and illiteracy given U.S. citizenship and limited-English proficiency parameter $p^{(I)}_{d(\underline{\nu},j)j}$ where $d = d(\underline{\nu}, j)$ is the minority estimation group to which $(\underline{\nu}, j)$ is mapped.

The counts needed for determinations in jurisdiction j are **(1)** $NCIT_j$, the number of voting age citizens, **(2)** $NLEP_{gj}$ and $NILL_{gj}$, respectively the number of limited-English proficent and illiterate voting age citizens in language minority group g. Given the mapping between minority language group and minority estimation group of the previous paragraph, we can think of these counts needed for the Section 203 determinations as being related to the 2010 Census counts $c_{\underline{\nu}j}$ for language minority group vector $\underline{\nu}$ in jurisdiction j and the proportions estimated in Section 4.2 by the following:

$$\begin{aligned} NCIT_j &= \sum_{\underline{\nu}} c_{\underline{\nu}j} p^{(C)}_{d(\underline{\nu},j)j}, \\ NLEP_{gj} &= \sum_{\underline{\nu}:\nu_g=1} c_{\underline{\nu}j} p^{(C)}_{d(\underline{\nu},j)j} p^{(L)}_{d(\underline{\nu},j)j}, \\ \text{and} \quad NILL_{gj} &= \sum_{\underline{\nu}:\nu_g=1} c_{\underline{\nu}j} p^{(C)}_{d(\underline{\nu},j)j} p^{(L)}_{d(\underline{\nu},j)j} p^{(I)}_{d(\underline{\nu},j)j}. \end{aligned} \tag{15}$$

Here $\sum_{\underline{\nu}:\nu_g=1}$ denotes the sum over all possible values of $\underline{\nu}$ where $\nu_g = 1$, that is for language minority group g (see notation defined in Table 4); The counts $NCIT_j$, $NLEP_{gj}$ and $NILL_{gj}$ are estimated by replacing the proportions $p^{(C)}_{d(\underline{\nu},j)j}$, $p^{(L)}_{d(\underline{\nu},j)j}$, and $p^{(I)}_{d(\underline{\nu},j)j}$ by the empirical Bayes estimates (13), yielding

$$\begin{aligned} \widehat{NCIT}_j &= \sum_{\underline{\nu}} c_{\underline{\nu}j} \hat{p}^{(C)}_{d(\underline{\nu},j)j}, \\ \widehat{NLEP}_{gj} &= \sum_{\underline{\nu}:\nu_g=1} c_{\underline{\nu}j} \hat{p}^{(C)}_{d(\underline{\nu},j)j} \hat{p}^{(L)}_{d(\underline{\nu},j)j}, \\ \text{and} \quad \widehat{NILL}_{gj} &= \sum_{\underline{\nu}:\nu_g=1} c_{\underline{\nu}j} \hat{p}^{(C)}_{d(\underline{\nu},j)j} \hat{p}^{(L)}_{d(\underline{\nu},j)j} \hat{p}^{(I)}_{d(\underline{\nu},j)j}. \end{aligned} \tag{16}$$

These estimates then provide the ratios and counts needed for the determinations for language

minority group g, as described in (1)-(7) . Specifically:

$$\begin{array}{ll} \text{LEP count:} & \widehat{LEP}_{gj} = \widehat{NLEP}_{gj}, \\ \text{LEP rate:} & \widehat{RLEP}_{gj} = \frac{\widehat{NLEP}_{gj}}{\widehat{NCIT}_j}, \\ \text{Illiteracy rate:} & \widehat{RILL}_{gj} = \frac{\widehat{NILL}_{gj}}{\widehat{NLEP}_{gj}}. \end{array}$$

All count estimates in these expressions are rounded to whole numbers. Determinations for American Indian/Alaska Native areas are computed in a similar manner to (16) noting that the jurisdiction is the small area in modeling and so the estimates of counts will be sums of $\hat{p}^{(C)}_{d(\nu,j),j}$, $\hat{p}^{(L)}_{d(\nu,j),j}$, and $\hat{p}^{(I)}_{d(\nu,j),j}$ over the appropriate ν and j within the AIA/ANA.

5 Data Preparation and Application

The purpose of this section is to outline the methods and procedures used to go from ACS and Census data to estimates produced from the models. As stated in a previous section, we have two main sources of data: the 2005-2009 American Community Survey (ACS) 5-year and the 2010 Census. The challenge is to take this information and produce estimates. In order to do so, various data tabulations and transformations are required.

5.1 ACS Data and Background

The American Community Survey (ACS) [U.S. Census Bureau, 2009] is designed as a continuous collection of data for statistical estimates at various levels of geography and for various time periods of one-year, three-years, and five-years. The ACS is designed to provide data to answer questions of various policy and statutory concern for persons, households, and group quarters. As a result, the methods of the ACS data collection and estimation are designed to apply to a wide variety of estimates.

In this project, we use 2005-2009 ACS 5-year data. This dataset contains approximately 22.6 million records of which 17.3 million are persons who are 18 years of age or older, and hence in the voting age population. Table 5 gives the particular ACS data variables of interest in this work. For a more complete definition see U.S. Census Bureau [2011].

Table 5: Definitions of ACS Variables of Interest

Variable	Description
AGE	Age.
CIT	Citizenship. This value represents whether a person is a citizen born in the United States, a citizen born in Puerto Rico, a citizen born elsewhere to American parents, naturalized, or not a citizen of the United States.
ENG	English Ability. The levels of this variable are speaks English "very well", "well", "not well", or "not at all". This presupposes that a person speaks a language other than English. If a person only speaks English, then this variable is encoded as 'NA'.
SCHL	Educational Attainment. This variable represents the highest level of education attained by the person.
HSGP	Hispanic Origin Group. This variable gives a single value designation for a person's Hispanic origin. This data was recoded to the ACS/2010 Census HSGP levels in order to match a related 2010 Census tabulation.
RCCx	Detailed Race Code. These codes relate the racial status of a person.
RCCxN	Numeric Detailed Race Code. The codes serve the same function as RCCx except that they are pure numeric values whereas RCC values are alpha-numeric.
ST	State. This variable gives a person's state of current residence.
CTY	County. This variable gives a person's county of current residence.
MCD	Minor Civil Division. This variable gives a person's MCD of current residence.
PWGT	Person Weight. This variable gives the survey sampling-based person weight associated to the person record.
REPW1-REPW80	Replication Person Weight. For each person, these 80 variables give the weights assigned to the person record for the purpose of computing sampling variances.

Given the 2005-2009 ACS 5-year data, we first remove all records with AGE < 18, as all our inferences relate to voting age persons. After that, we construct indicator variables for U.S. citizenship, limited-English proficiency, and illiteracy. See Table 6.

Table 6: Definitions of CIT_IND, LEP_IND, and ILL_IND.

Indicator	Value	Definition
CIT_IND	1	A U.S. Citizen.
	0	Not a U.S. Citizen.
LEP_IND	1	Speaks English less than "very well".
	0	Speaks English "very well" or does not speak a language other than English.
ILL_IND	1	Having completed at most the fourth grade as indicated by the SCHL variable
	0	Having completed a grade level higher than the fourth grade as indicated by the SCHL variable.

After the records are removed, indicator variables relating to the various language minority groups are formed. We take the RCCxN and HSGP codes, and we convert them into indicator variables for Hispanic status (HISP1-HISP6), Asian language minority status (ASIANG1-ASIANG16), and AIAN language minority status (AIANG1-AIANG51). The conversion of race and ethnicity codes into individual indicators allows us to partition the records into individual minority estimation groups.

5.2 Census Data and Background

As described in Section 4, we use the 2010 Census data as a control on the specific number of persons belonging to a particular race and/or ethnicity. In order to do this, we must be able to tabulate persons amongst the several language minority group combinations, geographic areas, and AIA/ANA tribal areas. Because the 2010 Census data provides race and ethnicity information, the task becomes relatively straight-forward: provide an extensive tabulation which gives the number of persons and voting age persons for each unique combination of language minority group membership (even if that membership is "other") for each unique combination of jurisdiction and AIA/ANA tribal area. The specifications for the tabulation are given by Beaghen [2011].

To further describe this tabulation, let us consider one case where we have a jurisdiction with Hispanic (HISP), Asian Indian (ASG1), and Bangladeshi (ASG2) persons. We could certainly have cases where a person belongs to multiple language minority groups. When this occurs, any particular combination of language minority groups will have its own tabulation record which gives

the number of persons (TOTAL) and adult persons (VOTAG) of that particular combination.

Table 7: Census Language Minority Group Tabulation Example

ν	HISP	AS	ASG1	ASG2	$\cdots$	CTY	AIANNHCE	TOTAL	VOTAG
1	1	0	0	0		XXX	9999	142	103
2	0	1	1	0		XXX	9999	54	41
3	0	1	0	1		XXX	9999	19	12
4	1	1	1	0		XXX	9999	3	1
5	0	1	1	1		XXX	9999	5	0

As we see in the rows of Table 7, we have examples of persons who are Hispanic alone, Asian Indian alone, Bangladeshi alone, Hispanic-Asian Indian in combination, and Asian Indian-Bangladeshi in combination in the county with FIPS[1] code XXX and AIANNHCE code 9999. (AIANNHCE represents American Indian, Alaska Native, and Hawaiian Homelands areas. In the case of this tabulation, codes for Hawaiian Homelands are not included and thus are not separated from non-tribal areas. The code 9999 represents persons that are not located within tribal areas.) When we summarize the 2010 Census in this fashion, this allows us to get an accounting of persons for all the particular groups of membership. Further, it allows us to partition the 2010 Census completely noting that if a person does not belong to any language minority group, that person would have an indicator for "OTHER". We should note that the tabulation example given in Table 7 is a simplified case as we omitted variables which are useful for tabulation. Likewise, we simplified the tabulation of Hispanic groups. Hispanic persons are recorded with a Hispanic indicator variable as described in the Table 6, but they are also given an indicator variable related to their specific HSGP value as noted earlier in Table 5.

5.3 Assignment of Multiple Language Minority Group Persons to Minority Estimation Groups

As has been described (Section 3), a person can belong to none, one, or multiple language minority groups, so formulating a model which is idealized to the unit (person) level can be difficult. However, if we are to model each person on an individual basis by applying different models based on that person's characteristics, then it is rather convenient to assign that person

[1]See FIPS PUB 5-3 for states, FIPS PUB 6-4 for counties, and FIPS PUB 55-3 for places and MCDs.

to one and only one group for modeling. For development of a model, persons who are in multiple language minority groups will only be assigned to one group, which we call a minority estimation group. The minority estimation group assignment needs to be done with some amount of care given the sensitive nature of the assumption. However, the assumption must be made in order to further facilitate estimation procedures.

The ideals of the assignment procedure assume that if a person happens to belong to multiple language minority groups, then that person will always belong to those constituent language minority groups for the purpose of tabulation. Any minority estimation group assignment will be based on local information as much as possible. If there is not enough local information, then we look to larger geographies to make a minority estimation group assignment. Persons can only be assigned to a minority estimation group if they already belong to that particular language minority group. To illustrate, assume we have a Japanese-Korean adult. This person, for the purpose of tabulation, is both Japanese and Korean. For the purpose of estimation, however, this person will be assigned to either the Japanese estimation group or the Korean estimation group based upon local population information. If there are more Japanese than Koreans locally, then the individual will be assigned to the Japanese estimation group, and the individual's characteristics will be estimated from the Japanese minority estimation group.

For the sake of modeling as stated, we separate Hispanics into six mutually exclusive groups (Mexican; Puerto Rican; Cuban; Central American, Dominican Republic; Latin/South American; Other Hispanic) as defined in Beaghen [2011]. Likewise, Hispanics who also belong to additional language minority groups will be assigned on the basis of the size of their specific Hispanic sub-group population and not based upon the overall Hispanic group population.

The procedure or algorithm for assigning each person to a specific minority estimation group for modeling is as follows:

- Assign all persons who belong to either none or one language minority group to their corresponding minority estimation group. All remaining unassigned persons are those who have two or more language minority group assignments.
- Assign all persons who are at least in part Mexican American Indian, Central American Indian, South American Indian, and Spanish American Indian first. If a person happens

to belong to only one of these language minority groups, then directly assign that person to that minority estimation group. If the person happens to belong to more than one of these language minority groups, we follow the general algorithm for assignment methodology detailed in Section 5.3.1 in order to decide amongst these minority estimation groups.

- Assign all remaining persons by the general algorithm for assignment detailed in Section 5.3.1.

The general purpose of this algorithm is to make sure that the Central, Latin/South, Mexican, and Spanish American Indians are assigned to some group and that this group is of non-trivial size. If too many persons are assigned to Hispanic minority membership, then the estimates constructed from the model for these groups of American Indians could be problematic as sample sizes could be too small.

5.3.1 General Algorithm for Assignment to Minority Estimation Groups

Next we describe the general algorithm for assigning persons. The general method is to assign a person based upon local information, but failing this, to assign based on the next level of geography. In this case, the smallest geography will be defined to be the jurisdictional level.

- If a person belongs to two or more language minority groups, then assign the person to the one of these groups with the largest "alone" population within that jurisdiction provided that this population is at least ten persons.
- If an assignment cannot be made or the above results in a tie at the jurisdiction level, then assign the person to the one of these groups with the largest "alone" population within that state provided that this population is at least ten persons.
- If an assignment cannot be made or the above results in a tie at the state level, then assign the person to the one of these groups with the largest "alone" population at the national level provided that this population is at least ten persons.
- If the three steps above cannot resolve the underlying assignment, then an assignment is made to the one of these groups with the lowest code number from Table 3.

Note that no language minority persons can belong to group #74.

All assignments of persons to minority estimation groups are based on counts from the 2010 Census.

5.4 Effective Sample Size Computation

As described in Section 4.1.2, we carry out the effective sample size computation as described by (10). This computation was carried out for every minority estimation group as defined in Table 3 at a national level. Variances were computed under the replication weight methodology as stated in U.S. Census Bureau [2009]. Variances were not computed for estimated proportions of zero or one.

6 Partitioning of Jurisdictions into Classes

The purpose of forming classes of jurisdictions is to take advantage of the similarity amongst jurisdictions so that more precise estimates of the model parameters can be formed. This section describes **(1)** the cases for which classes of jurisdictions are created within the minority estimation groups and **(2)** the procedure for constructing classes of jurisdictions for those cases.

Given the available ACS data, classes of jurisdictions were constructed for modeling U.S. citizenship and modeling limited-English proficiency given U.S. citizenship by the minority estimation groups of Hispanics (Table 3, Codes 1-6) and Asians (Table 3, Codes 7-22). In other words, separate classes were formed for **(1)** Hispanic U.S. citizenship and limited-English proficiency models and **(2)** Asian U.S. citizenship and limited-English proficiency. Classes were not formed for any model of illiteracy, nor were they formed for any models of American Indian/Alaska Native (Table 3, Codes 23-73) or persons who do not belong to a language minority group (Table 3, Code 74).

For those cases where classes were formed, all jurisdictions in the nation were separated into ten classes in a manner similar to the stratification construction of Dalenius and Hodges [1959]. Let y be a variable related to the variable of interest. In Dalenius and Hodges, bins of the ordered values of y (e.g. 0-5, 5-10, 10-15, etc.) are constructed. Within each bin the frequency, F, of sample units is counted and then used to form strata according to the cumulative square-root of F for each bin and all preceding bins.

We use a method similar to Dalenius and Hodges to create classes of similar jurisdictions. Jurisdictions relating to minority estimation group parameters are partitioned based on a

correlated proxy and in relation to a measure which gives an indication of size. Let us call this proxy for each jurisdiction (which depends upon the minority estimation group and the model of interest) y_j^*. Note that not all jurisdictions have sufficient ACS data to form the proxy y_j^*.

Similar to the method of Dalenius and Hodges which deals with frequency, the 2010 Census value relating to the total number of persons is used. For each jurisdiction within a minority estimation group, F_j is defined as the total number of 2010 Census persons who are either any Hispanic or any Asian within that jurisdiction, depending on whether the minority estimation group is a Hispanic or Asian group.

With a proxy y_j^* and count value F_j in hand, the cumulative-square root rule is used to form classes. Similar to the ordering in Dalenius and Hodges the jurisdictions are ordered by y_j^* from lowest to highest.

For the j^{th} jurisdiction, compute $\sqrt{F_j}$. Then define T as

$$T = \sum_j \sqrt{F_j}. \tag{17}$$

So, in order to partition into K classes, the cumulative total

$$C_j = \sum_{b=1}^{j} \sqrt{F_b} \tag{18}$$

for the j^{th} jurisdiction is found. Then all jurisdictions j for which $C_j \leq T/K$ are assigned to the first class. Next, all jurisdictions j for which $T/K < C_j \leq 2T/K$ are assigned to the second class, and so on, until all jurisdictions are distributed among the K classes.

Finally, for those few jurisdictions which lacked sufficient ACS data to develop regression estimates, the jurisdictions are assigned to the K classes as follows: assign the jurisdiction to the class which occurs most commonly in the state for the minority estimation group and model of interest. This only occurs in jurisdictions which have 25 or fewer persons in the 2010 Census population count.

For minority estimation groups, we shall provide a correlated proxy y^* based upon a regression model. This proxy will depend upon the modeled parameter. For models of U.S. citizenship,

Table 8: Independent Variables Used in the Partition Model.

Variable Name	Variable Description	Source
pund18	Percent persons under 18 years	2010 Census
p1854	Percent persons 18 to 54 years of age	2010 Census
p5564	Percent persons 55 to 64 years of age	2010 Census
p65	Percent persons 65 years of age or older	2010 Census
paian	Percent AIAN	2010 Census
pasian	Percent Asian	2010 Census
phisp	Percent Hispanic	2010 Census
pmale	Percent male	2010 Census
aveage	Average age of all persons	2010 Census
hhsize	Average household size	2010 Census
lhshld	Number of households	2010 Census
phh18	Percent households with one or more persons under 18 years	2010 Census
avhh60	Average number of persons 60 years or older per household	2010 Census
avhh65	Average number of persons 65 years or older per household	2010 Census
psingle	Percent household persons living in a single-unit home	2010 Census
pmulti	Percent household persons living in a multi-unit home	2010 Census
pown	Percent household persons living in an owned home	2010 Census
t(group)	Total for the race group being modeled	2010 Census
aper_FB	Percent foreign born	2005-09 ACS
aper_FBV	Percent foreign born that are voting eligible	2005-09 ACS
ed1rate	Percent persons completed less than 5th grade	2005-09 ACS
ed2rate	Percent persons completed at least 5th grade but no HS dipolma	2005-09 ACS
ed3rate	Percent persons with at least a HS diploma but no Bachelor's degree	2005-09 ACS
ed4rate	Percent persons with a Bachelor's degree	2005-09 ACS
aper_pov1	Percent persons with income under 50% of the poverty threshold	2005-09 ACS
aper_pov2	Percent persons with income 50-99% of the poverty threshold	2005-09 ACS
aper_pov3	Percent persons with income 100-124% of the poverty threshold	2005-09 ACS
aper_pov4	Percent persons with income 125-149% of the poverty threshold	2005-09 ACS
aper_pov5	Percent persons with income 150-184% of the poverty threshold	2005-09 ACS
aper_pov6	Percent persons with income 185-199% of the poverty threshold	2005-09 ACS
aper_pov7	Percent persons with income over 200% of the poverty threshold	2005-09 ACS

we shall find the jurisdiction's ACS estimate for citizenship amongst all voting age Hispanic or Asian persons. We will then use the variables of Table 8 and form a regression model computed from SAS PROC GLMSELECT using the model selection criterion, BIC. The model variables are selected using a stepwise selection admitting second-order and cross-product terms (provided the first-order terms are already in the model) starting from a model containing only an intercept term. Similarly, for models of limited-English proficiency given U.S. citizenship, we shall find the jurisdiction's ACS estimate for limited-English proficiency amongst all voting age U.S. citizen Hispanic or Asian persons. Model selection proceeds in an identical manner. All dependent variables are transformed using an arcsine-square root transformation. The independent variable (Table 8) for the number of households is modeled on a logarithmic scale and the independent variables relating to foreign born persons and poverty status are transformed using the arcsine-square root transformation. A prediction for each jurisdiction is obtained from the models. Note that if dependent variables are not available for a jurisdiction then they are not placed within the regression model, but so long as all the independent variables are not missing, we can form a prediction. If independent variables are missing, then the jurisdiction is assigned to a class by a secondary procedure.

Table 9 gives the variables which enter the regression model. We use x to indicate that the variable is present in the model either as a single term or as a cross-product, and we use xx to indicate that the variable is present in the model as a quadratic term.

As a result of the predictions from the developed model, Table 10 gives the number of jurisdictions belonging to each class identified by A-J. In each case, these classes are sorted from the lowest rates of U.S. citizenship and limited-English proficiency to the largest rates. Note that classes with higher concentrations of small areas implies that the number of voting age persons who are Hispanic or Asian is smaller than those with lower concentrations. Also note that the classes themselves are only labels and are only important within each model. The 706 Asian CIT jurisdictions and the 469 Hispanic CIT jurisdictions in Class A may be similar to each other in that they have low rates of citizenship but there is no direct linkage across models belonging to the same class label.

Table 11 gives an accounting of the resulting number of models by applying classes on the minority estimation groups. To understand this table, consider that Hispanics have 6 minority estimation

Table 9: Stepwise Model Selection Results

Variable	Asian CIT	Hispanic CIT	Asian LEP	Hispanic LEP
pund18				x
p1854	xx	x		
p5564	x	x		x
p65		x	xx	x
paian		x	x	x
pasian		x	xx	x
phisp	xx	xx	x	xx
pmale	x			xx
aveage		x		
hhsize	xx	x		x
lhshld	x	xx	xx	xx
phh18				xx
avhh60			xx	
avhh65		x		x
psingle	xx		x	x
pmulti		x	x	x
pown	x	x	x	x
t(group)	x	x	xx	x
aper_FB	xx	xx		x
aper_FBV	xx	xx	xx	xx
ed1rate			xx	
ed2rate		xx	xx	x
ed3rate	xx	xx	x	x
ed4rate		xx		
aper_pov1	x	x	x	x
aper_pov2			x	
aper_pov3				
aper_pov4	x			
aper_pov5		x		
aper_pov6				
aper_pov7		xx	x	x

Table 10: Number Jurisdictions by Class Assignment and Model

Class	Asian CIT	Hispanic CIT	Asian LEP	Hispanic LEP
A	706	469	3,866	3,899
B	406	330	1,366	1,280
C	464	307	785	746
D	486	271	562	510
E	488	351	385	358
F	600	414	256	293
G	639	592	182	221
H	693	752	158	197
I	1,089	1,095	126	180
J	2,321	3,311	206	208

Table 11: Model by Class by Parameter Counts

	Models			Count
	U.S. Citizenship	Limited-English	Illiteracy	
Hispanic	6×10 models	6×10 models	6×1 model	126 models
Asian	16×10 models	16×10 models	16×1 model	336 models
AIAN	51×1 model	51×1 model	51×1 model	153 models
Other	1×1 model	1×1 model	1×1 model	3 models
			Total:	618 models

groups, Asians have 16 minority estimation groups, AIANs have 51 minority estimation groups, and all other persons have one group. Since we only form classes for models on U.S. Citizenship and limited-English proficiency for Hispanic and Asian groups then each class will form an independent modeling case. For instance, Asian Citizenship parameters are estimated amongst 16 minority estimation groups and 10 classes for each minority estimation group which means that there are 160 distinct small area models constructed for estimating the rates of Asian citizenship. Likewise, if we were to consider AIAN citizenship, we have 51 minority estimation groups, which have only one class apiece. This means that there are 51 distinct small area models being used for estimation.

7 Specific Estimation of the Model Forms

Apart from the class partitions, estimates were computed using the R-2.11.0 software package with the `rbeta` function copied from R-2.12.2. The empirical Bayes estimates were performed using the function `optim`.

The full Bayesian portion was generated using an ARMS sampling procedure from the `HI` library using a function called `arms`. The `arms` function only takes arguments on a closed interval, and as a result, the density of M_{dh} was transformed to a $[0, 1]$ interval by use of the function $\frac{x}{1+x}$. Further, this was narrowed by only computing on intervals for which the density had a non-zero value as computed in R.

The variance estimates were produced by creating 1,000 burn-in iterations of the MCMC procedure followed by 50,000 realizations at the various levels of geography of interest on each of six compute nodes of a Linux cluster, each cluster running an independent R job. A total of 300,000 realizations were computed which were then combined to provide the variance estimates.

Various portions of the computational operation were made easier by use of SQLite, the `multicore` library within R, and the usage of a multiple node computer. Additionally, a specialized tallying function was constructed solely for the purpose of this project.

Tabulations for counts and rates were made at the national level, state level, jurisdiction level, and for AIA/ANA reservations. The empirical Bayes estimates were computed using decimal figures which were rounded using the `round` function with the precision set to 0. In other words, estimates were rounded after all summations were taken and rounded to the nearest whole number. These whole number figures were used in making determinations under Section 203.

7.1 Estimation of the National Illiteracy Rate

As we recall, in Section 2 we described how jurisdictions would qualify for language assistance. One of the quantities mentioned was $\frac{N_{\cdot 1 \cdot 1}}{N_{\cdot 1 \cdot \cdot}}$, the national illiteracy rate amongst voting age citizens (as found in equations (5) and (7)). Our model form cannot give an estimate for the numerator of this fraction. As a result it was decided that this ratio would be estimated by the standard weight-based estimator which is 1.16% (.0116). Modeling this quantity was avoided so as not to further complicate the model, as the national level estimate for the voting age citizen illiteracy rate should be sufficiently precise.

8 Concluding Remarks

This technical report has presented the model and estimation methods used to provide the point estimates and standard errors for ratios and totals of the voting age citizenship, limited-English proficiency, and illiteracy characteristics within the US population in 2010 as required under Section 203 of the Voting Rights Act. Further discussion and evaluations of modeling decisions and estimator forms will be provided in future technical reports and Census Bureau documents.

References

M. Beaghen. "Requirements for Tabulations from the Hundred Percent Detail File for the 2011 Section 203 Determinations Using American Community Survey Data (ACS11-R-7)". 2011 American Community Survey Memorandum Series, June 2011. Memorandum to D. Stoudt from D. C. Whitford, Washington, DC: U.S. Census Bureau.

B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, Boca Raton, Florida, 2nd edition, 2000.

T. Dalenius and J. L. Hodges. Minimum variance stratification. *Journal of the American Statistical Association*, 54:88–101, 1959.

M. H. De Groot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, New York City, 1970.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–533, 1996.

W. R. Gilks, N. G. Best, and K. C. C. Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics).*, 44(4):455–472, 1995.

U.S. Census Bureau. *Design and Methodology: American Community Survey*. U.S. Government Printing Office, Washington, DC, 2009.

U.S. Census Bureau. American Community Survey Puerto Rico Community Survey 2009 Subject Definitions, 2011. Obtained from `http://www.census.gov/acs/www/Downloads/data_documentation/SubjectDefinitions/2009_ACSSubjectDefinitions.pdf` on December 9, 2011.

Appendix

A Details of Variance Estimation

A.1 Measuring Uncertainty Using a Full Bayesian Method

As outlined in Sections 4.1.3, 4.2, and 4.3, the empirical Bayes component estimates were derived and used to construct the point estimates needed to make the Section 203 determinations. Empirical Bayes estimates were used because we knew we could obtain a solution from deterministic methods, unlike fully Bayes estimates based on Markov Chain Monte Carlo methods which have some degree of simulation error. However, since empirical Bayes variance estimates tend to underestimate the true variation, Bayesian methods with nearly non-informative priors for the hyper-parameters are adopted since this provides more conservative estimates of variation. The steps used to sample $NCIT_j$, $NLEP_{gj}$, and $NILL_{gj}$ from their posterior predictive distributions follow. First, the steps used to sample $NCIT_j$, $NLEP_{gj}$, and $NILL_{gj}$ from their conditional distributions is provided in Section A.1.2. Next, Section A.1.3 describes how samples of the hyper-parameters $\mu_{dh_C}^{(C)}$, $\mu_{dh_L}^{(L)}$, $\mu_{dh_I}^{(I)}$, $M_{dh_C}^{(C)}$, $M_{dh_L}^{(L)}$, and $M_{dh_I}^{(I)}$ are taken from their posterior distributions. In Section A.1.1 these parts are combined in order to detail the procedure for drawing realizations from the full posterior predictive distribution. Finally, Section A.2 shows how variance estimates are calculated.

A.1.1 Sampling from the Full Posterior Predictive Distribution

Based on the hierarchical model used, the full posterior predictive distribution can be expressed as the product of the three conditional distributions:

1) P($NCIT_j$, $NLEP_{gj}$, and $NILL_{gj}$ for all g, j | $p_{dj}^{(C)}$, $p_{dj}^{(L)}$, and $p_{dj}^{(I)}$ for all d, j)

2) P($p_{dj}^{(C)}$, $p_{dj}^{(L)}$, and $p_{dj}^{(I)}$ for all d, j | ACS^*, $\mu_{dh_C}^{(C)}$, $\mu_{dh_L}^{(L)}$, $\mu_{dh_I}^{(I)}$, $M_{dh_C}^{(C)}$, $M_{dh_L}^{(L)}$, $M_{dh_I}^{(I)}$ for all d, h)

3) P($\mu_{dh_C}^{(C)}$, $\mu_{dh_L}^{(L)}$, $\mu_{dh_I}^{(I)}$, $M_{dh_C}^{(C)}$, $M_{dh_L}^{(L)}$, $M_{dh_I}^{(I)}$ for all d, h | ACS^*)

where "ACS^*" denotes the adjusted counts: $\{m_{dj}^{*(C)}, m_{dj}^{*(L)}, m_{dj}^{*(I)}\}$. The procedures for sampling from the first distribution is given in Section A.1.2. The second distribution is a product of Beta distributions (see e.g. De Groot [1970], Section 9.8), and sampling from distribution is

straightforward. The procedure for sampling from the last distribution is given in Section A.1.3.

A.1.2 The Posterior Predictive Distributions of $NCIT_j$, $NLEP_{gj}$, and $NILL_{gj}$

As part of the posterior-predictive distribution, we obtain the conditional predictive distribution at the individual level as a sample of i.i.d. Bernoulli random variables with mean specified by (8). Consequently, we can take any combination $(\underline{\nu}, j)$ as a series of Binomial distribution draws.

For each $c_{\underline{\nu}j} \neq 0$, we draw:

$$\begin{aligned} N_{\underline{\nu}j}^{(C)} &\sim \text{Binomial}(c_{\underline{\nu}j}, p_{d(\underline{\nu},j)j}^{(C)}), \\ &\quad \text{otherwise, set } N_{\underline{\nu}j}^{(C)} = 0. \end{aligned} \tag{19}$$

For each $N_{\underline{\nu}j}^{(C)} \neq 0$, we draw:

$$\begin{aligned} N_{\underline{\nu}j}^{(L)} &\sim \text{Binomial}(N_{\underline{\nu}j}^{(C)}, p_{d(\underline{\nu},j)j}^{(L)}), \\ &\quad \text{otherwise, set } N_{\underline{\nu}j}^{(L)} = 0. \end{aligned} \tag{20}$$

For each $N_{\underline{\nu}j}^{(L)} \neq 0$, we draw:

$$\begin{aligned} N_{\underline{\nu}j}^{(I)} &\sim \text{Binomial}(N_{\underline{\nu}j}^{(L)}, p_{d(\underline{\nu},j)j}^{(I)}), \\ &\quad \text{otherwise, set } N_{\underline{\nu}j}^{(I)} = 0. \end{aligned} \tag{21}$$

The counts of the people defined by their U.S. citizenship status, LEP status, and illiteracy status, as estimated in (16), can be specified for each:

$$NCIT_j = \sum_{\underline{\nu}} N_{\underline{\nu}j}^{(C)}, \tag{22}$$

$$NLEP_{gj} = \sum_{\underline{\nu}:\nu_g=1} N_{\underline{\nu}j}^{(L)}, \tag{22'}$$

and

$$NILL_{gj} = \sum_{\underline{\nu}:\nu_g=1} N_{\underline{\nu}j}^{(I)} \tag{22''}$$

Analogously, when the denominator is positive, the finite population terms needed to make the determinations can be constructed as follows:

$$\begin{array}{ll} \text{LEP count:} & LEP_{gj} = NLEP_{gj}, \\ \text{LEP rate:} & RLEP_{gj} = \frac{NLEP_{gj}}{NCIT_j} \\ \text{Illiteracy rate:} & RILL_{gj} = \frac{NILL_{gj}}{NLEP_{gj}}. \end{array} \tag{23}$$

When the denominator is zero, the corresponding value is not used and left as undefined.

A.1.3 Posterior Distributions of $\boldsymbol{\mu^{(C)}_{dh_C}}$, $\boldsymbol{\mu^{(L)}_{dh_L}}$, $\boldsymbol{\mu^{(I)}_{dh_I}}$, $\boldsymbol{M^{(C)}_{dh_C}}$, $\boldsymbol{M^{(L)}_{dh_L}}$, and $\boldsymbol{M^{(I)}_{dh_I}}$

The likelihood solely for $\mu^{(C)}_{dh_C}$, $\mu^{(L)}_{dh_L}$, $\mu^{(I)}_{dh_I}$, $M^{(C)}_{dh_C}$, $M^{(L)}_{dh_L}$, and $M^{(I)}_{dh_I}$ can be obtained by combining (11) with (12′) and integrating with respect to $p^{(C)}_{dh_C}$, $p^{(L)}_{dh_L}$, and $p^{(I)}_{dh_I}$. The resulting distribution is a product of Beta-Binomial distributions. For the location parameters $\mu^{(C)}_{dh_C}$, $\mu^{(L)}_{dh_L}$, and $\mu^{(I)}_{dh_I}$, an independent uniform prior was used so that any value between zero and one is equally likely before any data is observed. The scale parameters, $M^{(C)}_{dh_C}$, $M^{(L)}_{dh_L}$, and $M^{(I)}_{dh_I}$, can be any positive value and, hence, a proper uniform distribution cannot be defined. However, independent half-Cauchy distributions were used because this type of distribution is considered to be a good choice for a non-informative distribution since the distribution is extremely flat (i.e., heavy-tailed) and a likelihood, if at all informative, will negate its minimal effects. See Gelman [1996] for further justifications of the use of a half-Cauchy as a "non-informative" prior.

Formally, for each minority estimation group d and each minority estimation group defined class of jurisdictions h (Section 6) the prior distributions are:

$$\begin{array}{rcll} p(\mu^{(C)}_{dh_C}) & = & 1, & 0 \leq \mu^{(C)}_{dh_C} \leq 1 \\ p(\mu^{(L)}_{dh_L}) & = & 1, & 0 \leq \mu^{(L)}_{dh_L} \leq 1 \\ p(\mu^{(I)}_{dh_I}) & = & 1, & 0 \leq \mu^{(I)}_{dh_I} \leq 1 \\ p(M^{(C)}_{dh_C}) & = & \frac{2}{\pi[1+(M^{(C)}_{dh_C})^2]}, & 0 \leq M^{(C)}_{dh_C} \\ p(M^{(L)}_{dh_L}) & = & \frac{2}{\pi[1+(M^{(L)}_{dh_L})^2]}, & 0 \leq M^{(L)}_{dh_L} \\ p(M^{(I)}_{dh_I}) & = & \frac{2}{\pi[1+(M^{(I)}_{dh_I})^2]}, & 0 \leq M^{(I)}_{dh_I}, \end{array} \tag{24}$$

all independently distributed. The full conditional posterior of each parameter, $\mu^{(C)}_{dh_C}$, $\mu^{(L)}_{dh_L}$, $\mu^{(I)}_{dh_I}$, $M^{(C)}_{dh_C}$, $M^{(L)}_{dh_L}$, and $M^{(I)}_{dh_I}$ is sampled from a "Gibbs Sampler" using the Adaptive Rejection Metropolis Sampler (ARMS) of Gilks et al. [1995].

A.2 Estimating Variance

The full Bayesian paradigm is used to provide an estimate of error for the empirical Bayes estimates. We apply this approach as estimates of variances from empirical Bayes procedures tend be underestimates of the true variance and relying on standard Bayesian procedures with vague priors should yield a more conservative value for the estimate of variance. Further, as empirical Bayes estimates were used as the point estimates, we employ a mean squared error calculation from the posterior predictive distributions of (22), (22′), and (22″) in place of the variance. Using Markov Chain Monte Carlo draws from the posterior distribution outlined in this section, a Monte Carlo estimate of the posterior mean squared error of $\widehat{NLEP}_{gj}$ in relation to $NLEP_{gj}$ (22′) is

$$\widehat{\text{MSE}}(\widehat{NLEP}_{gj}) = E[(NLEP_{gj} - \widehat{NLEP}_{gj})^2 | ACS^*]. \tag{25}$$

Computation for the posterior mean squared error of ratio forms $RLEP_{gj}$ and $RILL_{gj}$ are handled in a different way. Instead of using a form where we are simulating ratios (and thus, the numerator and denominator) from the posterior and then disregarding all cases for which the denominator is zero, we instead calculate an approximate MSE. As a result we can estimate the MSE of $RLEP_{gj}$ in the following manner:

Define $R = \widehat{RLEP}_{gj} = \frac{\widehat{NLEP}_{gj}}{\widehat{NCIT}_j}$. Using the Taylor series approximation about $\widehat{NLEP}_{gj}$ and $\widehat{NCIT}_j$ we obtain

$$\frac{NLEP_{gj}}{NCIT_j} \approx R + \frac{NLEP_{gj} - R * NCIT_j}{\widehat{NCIT}_j}.$$

The mean squared error is then approximated by

$$\widehat{\text{MSE}}(\widehat{RLEP}_{gj}) \approx \text{E}\left[\left(\frac{NLEP_{gj} - R * NCIT_j}{\widehat{NCIT}_j}\right)^2 \middle| ACS^*\right]. \tag{26}$$

Analogously, we can estimate the MSE of $RILL_{gj}$ in the following manner:

Define $R' = \widehat{RILL}_{gj} = \frac{\widehat{NILL}_{gj}}{\widehat{NLEP}_{gj}}$. Using the Taylor series approximation about $\widehat{NILL}_{gj}$ and $\widehat{NLEP}_{gj}$ we obtain

$$\frac{NILL_{gj}}{NLEP_{gj}} \approx R' + \frac{NILL_{gj} - R * NLEP_{gj}}{\widehat{NLEP}_{gj}}.$$

The mean squared error is then approximated by

$$\widehat{\mathrm{MSE}}(\widehat{RILL}_{gj}) \approx \mathrm{E}\left[\left(\frac{NILL_{gj} - R' * NLEP_{gj}}{\widehat{NLEP}_{gj}}\right)^2 \middle| ACS^*\right]. \tag{27}$$

Monte Carlo estimates for these measures are then created by the MCMC sampling scheme and these figures are reported as our estimates of variance.