Composite Measure of Size Evaluation and Primary Sampling Unit Formation for NHTSA's Redesign of the National Automotive Sampling System

William Cecere[1], Rui Jiao[1], Martha Rozsi[1]
Jacqueline Severynse[1], Sharon Lohr[1], James Green[1]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850

Abstract

One of the primary objectives of the revised National Automotive Sampling System (NASS) sample design is to update the previous NASS' primary sampling unit (PSU) sample. For probability proportional to size sampling, finding a composite measure of size (MOS) that is closely related to the multiple outcome variables of interest will reduce the variability of the estimates. In order to achieve this, external information from multiple sources was considered for both the redesign of NASS' Crash Report Sampling System (CRSS) and the Crash Investigation Sampling System (CISS) modules. The external information was used to develop and evaluate multiple composite measures of size against key outcome variables for each module. A MOS was then selected on the basis of correlation with outcome variables and the anticipated variance. The MOS for the secondary sampling units based on obtained crash counts is also presented. The selected MOS was also used to define a minimum MOS for CRSS PSU formation, while a different minimum PSU MOS variable was used for the CISS module.

Key Words: Primary sampling units, measure of size, probability proportional to size sampling, anticipated variance

1. Introduction

The National Automotive Sampling System (NASS), established in the 1970s, has been an integral part of the National Highway Traffic Safety Administration's (NHTSA's) efforts to fulfill its mission of providing nationally representative estimates and data about vehicles, crashes and injuries. The current NASS comprises two systems: the General Estimates System (GES) and the Crashworthiness Data System (CDS). Each system is based on a nationally representative probability sample of crashes selected from police accident reports (PARs). GES data, abstracted directly from the sampled PARs, focus on the larger overall crash picture, and are used to identify traffic safety problem areas, to estimate how many crashes of different types occur and to examine trends over time. These estimates provide a basis for regulatory and consumer initiatives (NHTSA, 2014). CDS data focus on passenger vehicle crashes and are used to evaluate the effectiveness of safety standards and to investigate injury mechanisms that could be affected by improvements in vehicle design. The CDS sample is smaller than the GES sample; for each PAR selected for the CDS, trained crash investigators visit the crash site, photograph and measure the vehicles' crash damage, interview victims, and review medical records.

Since its inception, NASS has proven to be a reliable resource for NHTSA and the broader motor vehicle safety research community (NHTSA, 2013).Improvements to automotive safety, and changes in population size and vehicle travel patterns, however, have resulted in changes in the locations and distributions of crashes of different types. The NASS system is being redesigned to better support NHTSA's and stakeholders' needs for information about overall crash estimates, crashworthiness, and crash avoidance topics. The redesigned GES will be called the Crash Report Sampling System (CRSS) module, to reflect that it will obtain information directly from PARs. The redesigned CDS will be called the Crash Investigation Sampling System (CISS) module, and data for this module will continue to be collected by trained crash investigators.

This paper focuses on the redesign of both modules of NASS at the Primary Sampling Unit (PSU) and Secondary Sampling Unit (SSU) levels. The three stages of the NASS design, which are shared by both modules, are displayed in Table 1. The first stage is composed of counties or groups of counties. For the redesign of NASS, these PSUs are stratified by region, urbanicity, crash types, and mileage counts. The second stage consists of Police Jurisdictions (PJs), which are police agencies that investigate motor vehicle crashes and submitted them to the state. Although a PJ can cross county boundaries, the crashes from a PJ can be attributed to an individual county. These PJs are stratified explicitly by a measure of size and implicitly by urbanicity defined at a local level. The third and final stage consists of PARs. Strata at this stage are defined as analytic domains. NHTSA has specified new analytic domains for both modules.

Table 1: Structure of the NASS three stage design

Stage	*Sampling Unit*	*Stratification*
1	County or group of counties	Region, Urbanicity, Crash types, Mileage counts
2	Police Jurisdiction	Measure of size (explicit), Urbanicity (implicit)
3	Police Accident Report (PAR)	Analytic domains

2. New Analytic Objectives and External Information

2.1 Modules

2.1.1 New Crash Report Sampling System (CRSS) Module

The purpose of the CRSS module (formerly the GES module) is to provide annual, nationally representative estimates of the number, types and characteristics of police-reported motor vehicle crashes from an on-going, PAR-based study. Estimates are to be obtained by vehicle type, injury type, crash severity, and vehicle model year. The target population is all motor vehicle crashes for which there is a PAR. This does not include all motor vehicle crashes, but it is thought that the majority of crashes that do not result in a PAR are minor, involving property damage only.

The new CRSS was required to be designed in a way that was best suited to maximize its own efficiency --- that is, designed independently of other modules. This is a departure from the previous design, in which the CDS PSUs were a proper subset of the GES PSUs. Other designs were also considered, in which CISS PSUs were nested within CRSS PSUs

or the surveys would be designed with maximum overlap of PSUs using methods such as those described in Keyfitz (1951) and Ohlsson (1998).

The new CRSS also comes with a new set of Analytic Domains, or PAR strata, as shown in Table 2. These PAR strata, and their target sample allocations, are defined so that the sample will contain adequate sample sizes of rarer crashes such as those involving fatalities or severe injuries. Crashes involving late model year vehicles are also to be oversampled. The strata are structured in a hierarchical fashion such that each stratum does not contain crashes included in previous strata.

Table 2: CRSS Analytic Domains – PAR Strata

Stratum	*Target Sample Allocation %*	*Population %*
1. Crashes involving a killed or injured non-motorist	9%	2.2%
2. Crashes not in Stratum 1 involving a killed or injured motorcycle or moped rider	6%	1.4%
3. Crashes not in Stratum 1 or 2 involving a killed or incapacitated late model year vehicle occupant	4%	0.42%
4. Crashes not in Stratum 1-3 involving a killed or incapacitated old model year vehicle occupant	7%	1.6%
5. Crashes not in Stratum 1-4 involving an injured late model year vehicle occupant	14%	6.2%
6. Crashes not in Stratum 1-5 involving a medium or heavy truck or bus	6%	5.7%
7. Crashes not in Stratum 1-6 involving an injured old model year vehicle occupant	12%	15.0%
8. Crashes not in Stratum 1-7 involving no injury to a late model year vehicle occupant and no person in the crash is killed or injured	22%	28.4%
9. Crashes not falling in previous strata	20%	39.0%

Note: A late model year vehicle is defined as one manufactured in the previous four years and an old model year vehicle is one more than four years old.

The CRSS PAR stratum target sample allocation presented in Table 2 differs from a proportional allocation, reflecting NHTSA's analytic objectives for CRSS. Under these circumstances, the design effect due to differential weighting or weighting effect is customarily used to evaluate the effect of the deviation from proportional allocation on the variance and thus precision of overall estimates. Using equation 4.1 of (Kish, 1992), the design effect due to differential weighting was calculated for the new CRSS to be 1.41, a substantial improvement to the weighting effect of 1.73 with the old CRSS (GES). Note that the design effect due to differential weighting only affects estimates that are based on the entire sample. Estimates that are calculated using crashes within a single PAR stratum, such as characteristics of crashes involving a killed or injured non-motorist, will not have variance inflation from disproportional weighting.

2.1.2 New Crash Investigation Sampling System (CISS) Module

The CISS module (formerly the CDS) gathers accurate, detailed, nationally representative information on passenger vehicle crashes. To obtain more details on passenger vehicle crashes, researchers carry out crash site inspections, vehicle inspections, in-person interviews, and gather medical records. This new CISS will be independent of CRSS and will oversample severe crashes and crashes involving late model year passenger vehicles. As with the CRSS, a new set of PAR strata was specified by NHTSA with an even greater use of disproportionate sampling for CISS to obtain a sufficient number of rare, severe, injury crashes.

The new CISS PAR strata shown in Table 3 have the same hierarchical structure as the CRSS PAR strata. Using the same approach done for CRSS to measure the effects of differential weighting, the weighting effect was calculated for the new CISS to be 1.95. This was an increase from the previous CDS of 1.80 but is due mainly to the increase in oversampling of newer vehicles and severe crashes.

Table 3: CISS Analytic Domains – PAR Strata

Stratum	*Target Sample Allocation %*	*Population %*
1. Crashes involving a killed passenger vehicle occupant.	5%	0.51%
2. Crashes not in Stratum 1 involving a recent model year passenger vehicle in which an occupant is incapacitated	10%	0.93%
3. Crashes not in Stratum 1 or 2 involving a recent model year passenger vehicle in which an occupant is injured (including crashes in which injury severity is unknown)	20%	8.71%
4. Crashes not in Stratum 1-3 involving a recent model year passenger vehicle in which no occupants are injured	15%	17.48%
5. Crashes not in Stratum 1-4 involving a mid-model year passenger vehicle in which an occupant is incapacitated	6%	1.28%
6. Crashes not in Stratum 1-5 involving a mid-model year passenger vehicle in which an occupant is possibly injured	12%	11.31%
7. Crashes not in Stratum 1-6 involving a mid-model year passenger vehicle in which no occupants are injured	10%	22.5%
8. Crashes not in Stratum 1-7 involving an old model year passenger vehicle in which an occupant is incapacitated	6%	1.55%
9. Crashes not in Stratum 1-8 involving an old model year passenger vehicle in which an occupant is possibly injured	10%	11.87%
10. Crashes not in Stratum 1-9 in which no occupants are injured	6%	23.87%

Note: For the model year ages, recent indicates a vehicle manufactured in the previous four years, mid indicates five to nine years old, and old indicates ten years or older.

2.2 Auxiliary Information

Due to the new analytic objectives for the CRSS and CISS, new PSU frames were required. As stated in Table 1, PSUs are counties or groups of counties, and a new measure of size was needed to use in Probability Proportional to Size (PPS) sampling and frame construction. An ideal measure of size would be based on population crash counts for each PAR strata for every PSU. Since this information is only available for previously sampled PSUs and not the entire frame, appropriate quality sources of information were needed to construct a measure of size.

Table 4 displays external sources of information available for this study. The Fatality Analysis Reporting System (FARS) data are a comprehensive census of fatal crashes (NHTSA FARS, 2014). Multiple years of data were needed to increase the stability of the crash counts, and to reduce the number of counties with zero fatal crashes. From the American Community Survey, information representing overall population, pedestrians, and cyclists were extracted. The State Data Systems (SDSs) provided rich information regarding the number of crashes with different injury severities (NHTSA SDS, 2014). The SDS data, however, were only available for 33 states: imputation models were developed to impute the estimates for the remaining states.

NHTSA purchased data from the R.L. Polk Company that contained information on vehicle registrations. This included vehicles miles driven by vehicle type and model year as well as vehicle counts and proportions. Data made available by Highway Loss Data Institute (HLDI) consist of crashes reported to insurance companies and provide information on the claims. However, since HLDI records list the crash in the county where the vehicle is registered as opposed to where the crash occurred, this information was used strictly as outcome measures and not considered as a MOS component.

After the PSU frames were formed and PSUs were sampled, NHTSA regional offices compiled crash information from the PJs within the sampled CRSS and CISS PSUs into the PJ list frame. Six types of crash counts were given representing fatal, injury, pedestrian, motorcycle, commercial motor vehicle, and total crashes.

Table 4: Available sources of information

Name	*Source*
Fatality Analysis Reporting System (FARS)	NHTSA
American Community Survey (ACS)	Census Bureau
State Data System (SDS)	NHTSA
POLK	R.L. Polk Company
Highway Loss Data Institute (HLDI)	HLDI
PJ List Frame (for sampled PSUs only)	NHTSA

3. Measure of Size Analysis and PSU Formation

3.1 Composite Measure of Size

A number of candidates were available for use as a measure of size (MOS). ACS population estimates could be used as a MOS since we expect the number of crashes and miles driven to be correlated with county and PSU populations. Alternatively, counts of fatal crashes, available from FARS, could be used as a measure of size since they too are expected to be positively correlated with numbers of other types of crashes. Used by themselves, however, counts of fatal crashes are unstable for small counties even when several years of data are aggregated.

For this redesign, several composite MOSs were considered because of the desire to ensure that crashes in some of the PAR strata are oversampled. Composite MOSs were evaluated as a way of balancing features of individual variable information with PAR strata (Folsom, 1987). Using a MOS of this form gave a roughly self-weighted sample for multiple domains. This allowed us to map relevant information for each PAR strata and create a MOS using the formula

$$MOS_i = \sum_{k=1}^{9} \frac{n_k}{n} \frac{N_{ik}}{N_k}$$

where

N_{ik} = the number of crashes in PAR stratum k and in PSU i

N_k = the number of crashes in PAR stratum k in the population

n_k = the desired sample size of crashes in PAR stratum k

n = the desired total sample size of PARs

The MOS in the above equation reflects the desire to obtain crashes of all types in the sample, but to oversample the more severe crashes. The values of N_{ik} are unknown, but were estimated using information from the data sources listed in Table 4. Thus, N_{ik} could be estimated for each PSU using information for the five most recent years of FARS data, along with information on the proportions of vehicle registrations in each PSU that were from late model year vehicles. Because the composite measure of size involves the proportions N_{ik}/N_k, estimates could be drawn from different sources, and have different scales, for the different PAR strata. Since there was no reliable estimate of the number of crashes in which motorcyclists were injured, for example, the number of miles traveled by motorcyclists, available from the POLK data, was used to estimate the values of N_{i2}/N_2.

Use of a composite measure of size had another advantage. Some of the small PSUs could have a value of zero for one of the estimated components in the MOS: for example, a PSU might have no fatal crashes in the previous five years. Every PSU had positive values for some of the PAR strata, however, so that every PSU had positive composite MOS.

For the CRSS MOS, ACS population was examined along with three composite MOS candidates using POLK, FARS, and SDS data to match to appropriate PAR strata. Outcome variables representing fatal, injury, and property damage only crashes were used as a basis for evaluation of the MOS candidates.

For the CISS MOS, ACS population was examined along with four composite MOS candidates. Similar outcome variables were used with the addition of HLDI insurance claim variables. As would be expected, all of the MOS variables were highly correlated with each other and with the outcome variables. In addition to evaluating these candidate MOS variables on correlation with outcomes and data quality, we examined anticipated variance of outcomes using each MOS. This could be done for the CISS due to the frame construction not depending on the MOS variable as explained in Section 3.2.

Since each MOS is scaled such that it sums to one, we calculated the within-stratum with-replacement selection probability for a specific MOS as

$$P_{hi} = MOS_i/(\text{sum of } MOS_i \text{ for all PSUs in stratum } h)$$

Let Y_{hi} denote the population total of the outcome variable in PSU i of stratum h, let Y_h denote the population total summed over all PSUs in the stratum, and let n_h denote the number of PSUs to be sampled from stratum h. Then the anticipated variance for the outcome variable, per sampled first stage unit, is

$$AV(MOS) = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{N} P_{hi} \left(\frac{Y_{hi}}{P_{hi}} - Y_h \right)^2 .$$

This formula for the anticipated variance is sensitive to small changes in the probabilities because of the factor that is squared. Thus, looking at a variety of outcomes can "smooth" some of that sensitivity to small deviations in the probabilities.

The outcome measures are on different scales, so the relative variance for each outcome and MOS is calculated as

$$\frac{\text{AV(MOS) for that MOS and outcome}}{\text{Minimum value of AV(MOS) for that outcome, among the MOS's considered}} .$$

The chosen MOS variable had the lowest average anticipated variance for all outcomes. This was a composite MOS containing FARS data representing the fatality stratum while using POLK to calculate the proportion of vehicles in age categories and multiplying them by the number of related injury crashes as estimated by the SDS.

As mentioned in Section 2.2, NHTSA compiled a frame of crash counts for all PJs in sampled CRSS and CISS PSUs. Regression models were developed for each module to apportion these six categories into counts representing each of the PAR strata, as described in Jiao et al. (2014). Using the same approach as for the PSUs to form a composite MOS, a PJ composite MOS was constructed for each module using the modeled counts for each PJ and the same sampling fractions from Tables 2 and 3.

3.2 Minimum Measures of Size

Due to analytic objectives needing to be met for a PSU sampling frame, a minimum PSU MOS is needed as the first step in forming PSUs. Using a unit such as a single county does not yield enough crashes to meet needed objectives for CRSS or CISS. Forming all PSUs larger than the minimum MOS ensures that the PSUs selected for the sample will have a sufficient number of crashes in the different PAR strata to support the research objectives of the CRSS and CISS.

For the CRSS module, the minimum MOS is calculated such that a self-weighting sample within case or PAR strata can be selected across the sampled PSUs. Green et al (2002) give general guidelines for forming PSUs so that they meet minimum MOS criteria. The minimum MOS requirement for PSUs is determined by solving the following equation:

$$Min(MOS_i) \geq f_{max} \frac{\sum_{i=1}^{N_h} MOS_i}{n},$$

where

$Min(MOS_i)$	=	minimum MOS for all PSUs;
f_{max}	=	largest overall sampling rate;
MOS_i	=	MOS for PSU i; and
n	=	number of PSUs to be selected.

For some CRSS PAR strata, the f_{max} was too high to keep the minimum MOS small enough to meet operational constraints. Therefore, the next largest overall sampling rate was used. This provided an approximate but not exact self-weighting sample for these PAR strata whereas it will be exact for the others.

For the CISS module, PSUs were formed with the goal of having a high probability of obtaining at least 5 fatal crashes involving a passenger vehicle in each PSU each year. The minimum of 5 fatal crashes involving a passenger vehicle was chosen since it is roughly 5% (CISS PAR stratum 1) of 104 total crashes annually which is the expected crash investigation workload for a CISS PSU with one researcher. The following reasoning was used to arrive at the probability. Assume that, for a given PSU, the number of crashes of a certain type (denoted by random variable X) follows a Poisson distribution with mean λ. Then, the probability that there are at least k crashes of that type in the PSU is

$$P(X \geq k) = 1 - \sum_{x=0}^{k-1} e^{-\lambda} \frac{\lambda^x}{x!}$$

Solving this equation we find that a λ value of 8 gives a probability of 0.9 of having at least 5 fatal crashes.

3.3 Forming the PSU Frame

In order to draw a PPS sample of PSUs, a frame of CRSS and CISS PSUs had to be created first. The PSUs were formed according to the following criteria:

- PSUs were formed as counties or groups of adjacent counties
- PSUs were required to achieve a minimum (with few exceptions) PSU MOS
- PSUs respected region, state and urbanicity status
- Outlying areas of AK and HI were excluded

For CISS PSUs, the requirement of respecting state boundaries was relaxed. Due to the requirements of researchers being able to travel to the scene of a crash, additional distance constraints were made. They were that an urban PSU be no more than 65 miles and a rural PSU be no more than 135 miles from end-to-end. An urban PSU contains one or more metropolitan areas with more than 250,000 people.

To form PSUs, the Westat proprietary software WESPSU was used (Green 2002). The software uses an optimization approach that minimizes travel costs by minimizing end-to-end PSU distance, subject to a maximum distance constraint while meeting the county contiguity, minimum MOS, region and urbanicity constraints listed above.

Figures 2 and 3 show the PSU frames of the west region for CRSS and CISS respectively. The CRSS PSUs averaged approximately six counties per PSU and the CISS averaged less than three. The smaller PSUs for the CISS frame are suitable for the researcher travel time requirements.

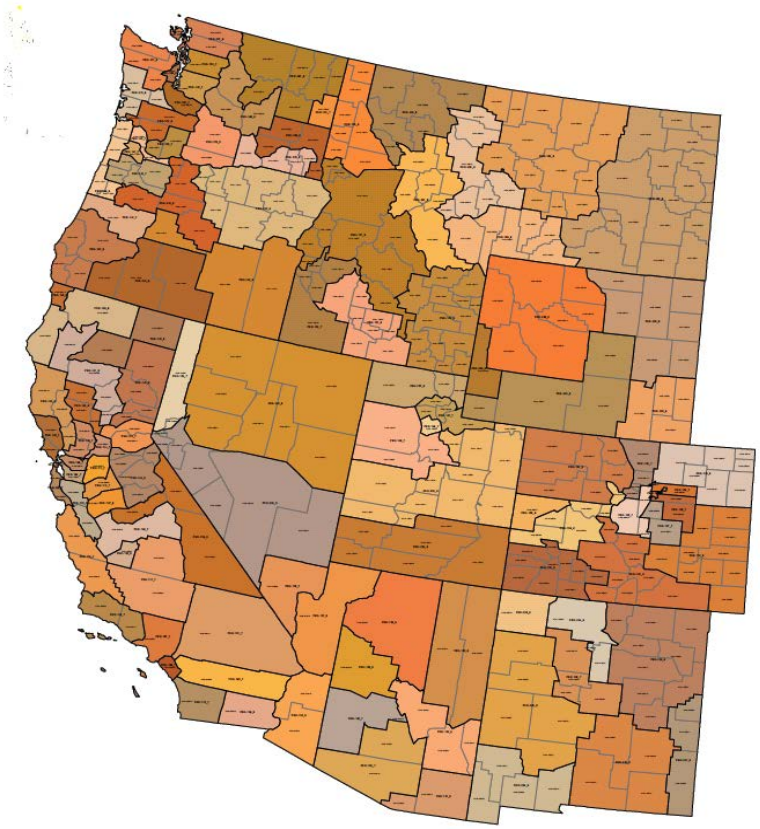

Figure 2: CRSS PSU frame for the west region

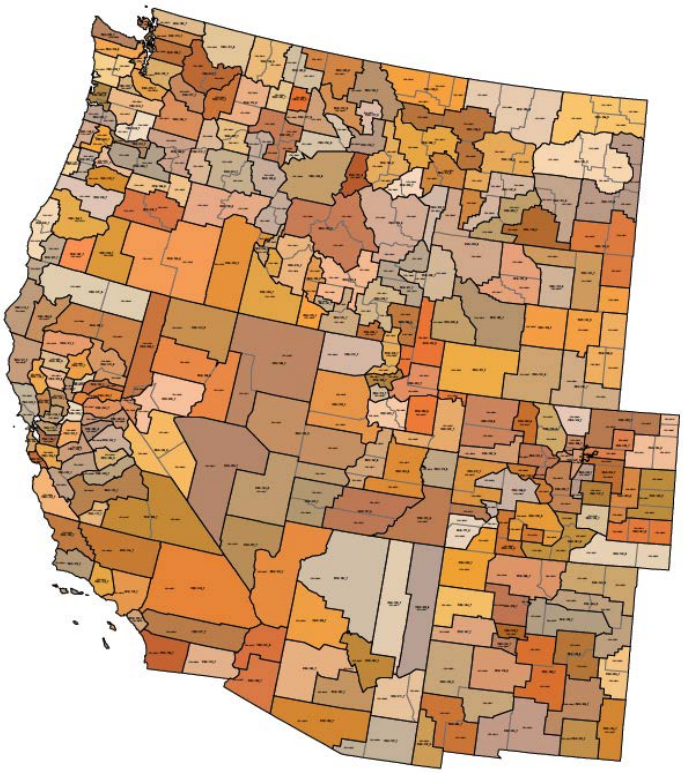

Figure 3: CISS PSU frame for the west region

4. Discussion

The redesign of NASS involved new analytic domains for both the CRSS and CISS modules. Using our available external data sources, we were able to represent the populations of the PAR strata for both frames. Appropriate composite measures of size were formed and evaluated based on correlation with outcomes of interest, data quality, and anticipated variance across the outcomes. These criteria were used to select a MOS for each module.

A minimum measure of size was constructed based on sampling rates for the CRSS module while likelihood to achieve the desired number of fatal crash investigations was used for the CISS module. Finally, PSU frames were formed for each module using a set of constraints and a PJ-level MOS was constructed.

The current GES design does not take advantage of electronic Police Accident Reports (ePARs) that are currently collected by some states, and NHTSA requested an initial design for the CRSS that does not use ePARs. One advantage of having independent designs for the CRSS and CISS is that the CRSS design can be easily modified to take advantage of future ePAR availability, should that be desired. Making use of ePARs in the future has the potential to lower the cost of data collection in the CRSS, particularly if a uniform ePAR form were to be adopted and data could be transmitted electronically. Since the CRSS respects state boundaries, states with electronic data collection could be sampled with less cost, and the resources from the survey could be diverted toward improving precision in the remaining states.

References

Folsom, R.E., Potter, F.J., and Williams, S.R. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 792-796.

Green, J.L., Chowdhury, S., and Krenzke, T. (2002). Developing Primary Sampling Unit (PSU) Formation Software. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, 1239-1243*

Jiao, R., Sugawara, Y., Rozsi, M., Lohr, S., Green, J. and Cecere, W. (2014). Estimating Population and Design Parameters for NHTSA's New National Automotive Sampling System. In press, *Proceedings of the American Statistical Association Section on Survey Research Methods.*

Kish, L. (1992), Weighting for Unequal P_i, *Journal of Official Statistics*, 8, 183-200.

Keyfitz, N. (1951). Sampling with probabilities proportionate to size: Adjustment for changes in probabilities. *Journal of the American Statistical Association*, 46, 105-109.

National Highway Traffic Safety Administration (NHTSA, 2013). *National Automotive Sampling System -- Crashworthiness Data System: 2012 Analytical User's Manual.* Washington, DC: US Department of Transportation. Last accessed 9/19/2014 from http://www-nrd.nhtsa.dot.gov/Pubs/811830.pdf.

National Highway Traffic Safety Administration (NHTSA, 2014). NASS General Estimates System. Last accessed 09/19/2014 from http://www.nhtsa.gov/Data/National+Automotive+Sampling+System+(NASS)/NASS+General+Estimates+System.

National Highway Traffic Safety Administration (NHTSA FARS, 2014). *Passenger Vehicle Occupant Fatalities: The Decline for Six Years in a Row from 2005 to 2011*. Washington, DC: US Department of Transportation. Last accessed 9/19/2014 from http://www-nrd.nhtsa.dot.gov/Pubs/812034.pdf.

National Highway Traffic Safety Administration (NHTSA SDS, 2014). *State Data System Crash Data Report: 2000-2009*. Washington, DC: US Department of Transportation. Last accessed 9/19/2014 from http://www-nrd.nhtsa.dot.gov/Pubs/812052.pdf.

Ohlsson, E. (1998). Sequential Poisson Sampling. *Journal of Official Statistics*, 14, 149-162.