Model-Aided Sampling: An Empirical Review

Marcus Berzofsky[1], Susan McRitchie[2], Mark Brendle[3]
[1]RTI International, 3040 Cornwallis Rd, RTP, NC 27709
[2]RTI International, 3040 Cornwallis Rd, RTP, NC 27709
[3]National Center for O*NET Development, 700 Wade Avenue, Raleigh, NC 27611

Abstract
Model-aided sampling (MAS) is a hybrid sampling approach that combines probability based sampling with a representative sampling paradigm. MAS is ideally suited for simultaneously sampling multiple target populations and optimizing the sample yield across each population; thereby, reducing the burden to the public while minimizing potential bias in the estimates. The O*NET Data Collection Program (O*NET), a large nationally representative establishment survey of occupations, tested MAS through a simulation study and presented its finding at ICES-III. O*NET simultaneously collects data on over 900 occupations each of equal importance to the study's objectives. The simulation study indicated that MAS would not substantively bias estimates for an occupation while reducing the level of burden to the study. Since 2007, O*NET has sampled and published estimates for over 300 occupations using the MAS paradigm. This paper presents the results of an empirical study evaluating the effectiveness of MAS to reduce burden to the public without introducing bias to the estimates by comparing the estimates solely obtained through large sampling theory and estimates obtained using the MAS paradigm.

Key Words: Model-aided sampling, O*NET, occupational survey, sample design, quota sampling, burden to public, estimate bias

1. Introduction

1.1 Overview of the O*NET Data Collection Effort

The O*NET Data Collection Effort (O*NET) is a large, nationally representative establishment survey of job incumbents from over 900 occupations. Sponsored by the U.S. Department of Labor and conducted by the O*NET Center and RTI International, for each occupation of interest, O*NET seeks to obtain information on the work context, work activities, knowledge needed to perform the occupation, tasks involved in performing the occupation, education, and work styles for the occupation. This survey data combined with analyst ratings is used to provide information to persons interested in a particular occupation as well as employers wanting to know what characteristics they should look for in candidates for a particular occupation. Occupation information is available to the public on the website www.onetcenter.org. Because information on each occupation is of equal importance (i.e., there are over 900 target populations) the study design must ensure that the statistical precision of each occupation meets the same minimum requirements.

O*NET has a continuous data collection sample design based on a wave design whereby multiple samples for a set of occupations are selected. Each wave contains approximately 50 occupations. Sub-waves, for a particular set of occupations, are separated by 7 – 9 months. Occupations that have met the requirements for being completed are not included in future sub-waves (Section 1.2 describes the requirements for an occupation being deemed complete). Within each wave, the sample design is a two stage design. In the first stage, establishments are selected. In the second stage, incumbents from the occupations of interest are selected.

1.2 What is Model-Aided Sampling

While this basic design has been in place since data collection began in 2001, it has been continuously improved to better ensure equal precision across all occupations. As described in Berzofsky et. al. (2007), one of the major challenges for O*NET is that some occupations (e.g., Secretaries, Secondary Teachers) are easy to identify in the population while other occupations (e.g., Bridge and Lock Tenders) are difficult to find. This leads to a greater number of completed surveys in some occupations while other occupations struggle to meet to minimum requirement. This costs the study money (i.e., spending time sending out surveys for occupations where they are not needed) and uses the limited amount of burden to the public (the U.S. Office of Budget and Management allots a fixed amount of time that each study sponsored by the Federal government can spend surveying the pubic as an attempt to limit the overall burden on the public) allotted to O*NET on incumbents in occupations where more interviews than needed have been conducted.

To minimize this issue, O*NET developed model-aided sampling (MAS; Berzofsky, et. al. 2008). The precision targets on O*NET require at least 15 questionnaires completed in each domain type (e.g., work activities, work context). MAS refined these requirements by incorporating sample distribution requirements across three frame characteristics: size of the establishment, region establishment is in, and industry of the establishment. Once one of the distributional requirements was met, MAS allowed data collection to be stopped for occupation among establishments in that frame characteristic (e.g., if enough

questionnaires for an occupation have been obtained in the Northeast than no additional establishments in the Northeast will be asked about that occupation). Therefore, unlike a traditional sampling paradigm (i.e., a paradigm based on large sampling theory that uses survey weights to make inference to the population of interest), by ensuring the distribution of sampled incumbents fits a pre-defined model for the occupation, MAS allows data collection to be halted for an occupation prior to all sampled establishments being worked in the field.

1.3 Prior Simulation Study

By relying on a model, in addition to sampling weights, rather than solely the sampling weights from a traditional paradigm, MAS has the potential to introduce bias in its estimates if the model is misspecified. To assess if this would occur on O*NET, prior to implementing MAS, a simulation study was conducted to assess the level of bias induced by a MAS design and the amount of burden to the public that would be saved had MAS been used rather than a traditional sampling paradigm. The details of the simulation study can be found in Berzofsky, et. al. (2006). The simulation study found that 99.5% of estimates were not substantively different (i.e., their estimate did not differ by +/- 1 point for 5-point items and +/- 1.5 points for 7-point items) under MAS compared to the estimates sampled by a traditional paradigm. Furthermore, the burden on the public would decrease 58.5% under MAS because number of completed interviews for easy to find occupations would be greatly reduced.

1.4 Goals of the Study

Based on the results of the simulation study, RTI implemented MAS in 2008. Prior to this 867 occupations had been fielded using the traditional paradigm. In the 4 years since the implementation of MAS, estimates for 332 occupations that were sampled using MAS, and were previously fielded using a traditional paradigm, have been analyzed and published. The purpose of this study is to empirically assess how MAS performs compared to a traditional sampling paradigm. In doing so, this paper seeks to answer two questions:

1. Do estimates produced under MAS substantively differ from estimates produced under a traditional sampling paradigm?
2. How much, if any, burden to the public is saved by using a MAS design?

2. Methods

2.1 Selection and Description of Occupations in Analysis

The 332 occupations that were selected for analysis were those that had been analyzed and published under both the traditional sampling paradigm and MAS. These occupations had similar distributions compared to all 818 occupations that had been published under the traditional paradigm according wage, size of the company employing the occupation and the job zone (amount of preparation needed to perform the job). These distributions are shown in tables 1, 2, and 3.

Table 1: Occupations by Average Wage

Average Wage	Number Analyzed	Percent	All Occupations	Percent
0 - 29,999	51	15.36%	143	17.48%
30,000 - 44,999	77	23.19%	271	33.13%
45,000 - 59,999	75	22.59%	167	20.42%
60,000 - 74,999	54	16.27%	99	12.10%
75,000-89,999	33	9.94%	61	7.46%
90,000 - 104,999	19	5.72%	40	4.89%
105,000 - 119,999	6	1.81%	14	1.71%
120,000 plus	17	5.12%	23	2.81%
Total	332		818	

Table 2: Occupations by Establishment Size

Average Number of Employees	Number Analyzed	Percent	All Occupations	Percent
0 - 49,999	100	30.12%	422	51.59%
50,000 - 99,999	61	18.37%	121	14.79%
100,000 - 249,999	80	24.10%	135	16.50%
250,000 - 499,999	40	12.05%	59	7.21%
500,000 - 749,999	21	6.33%	22	2.69%
750,000 - 999,999	10	3.01%	14	1.71%
1,000,000 plus	20	6.02%	45	5.50%
Total	332		818	

Table 3: Occupations by Job Zone

Job Zone	Number Analyzed	Percent	All Occupations	Percent
1 - Little or No Preparation Needed	16	4.82%	53	6.14%
2 - Some Preparation Needed	74	22.29%	261	30.24%
3 - Medium Preparation Needed	99	29.82%	241	27.93%
4 - Considerable Preparation Needed	71	21.39%	179	20.74%
5 - Extensive Preparation Needed	72	21.69%	129	14.95%
	332		863	

2.2 Analysis

2.2.1Assessing Bias

In order to determine if there were substantive differences between the estimates produced under MAS and the traditional paradigm we created confidence bands for the 5-point items and the 7-point items. The thresholds for determining substantive differences in the simulation study were also used in this study. For the 5-point estimates we classified an estimate as substantially different if there was at least a 1.0 point difference between the MAS estimate and the estimate under the traditional paradigm. The threshold for the 7-point estimates was +/- 1.5 points.

2.2.2 Assessing Burden

In order to determine if burden was reduced, we compared the number of questionnaires sent to potential respondents under the traditional paradigm to MAS. In addition we compared the distribution of questionnaires shipped for all 332 occupations.

3. Results

3.1 Comparison of Estimates

3.1.1 Five-Point Estimates

For the questions related to importance of work activities, we found that 95% of the estimates were not substantially different (Figure 1). Similarly 97% of the estimates for importance items related to knowledge were not substantially different (Figure 2) and 96% of the estimates related to work context were not substantially different (Figure 3).

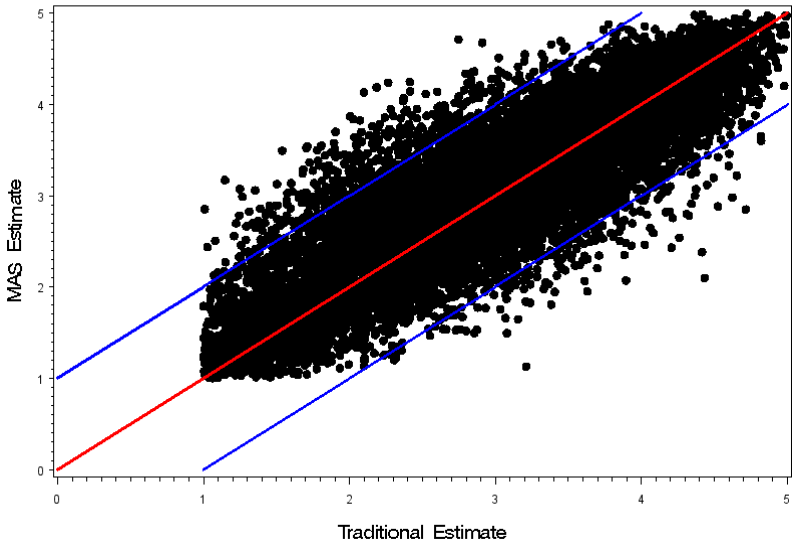

Figure 1: Work Activities—Importance (5-point items)

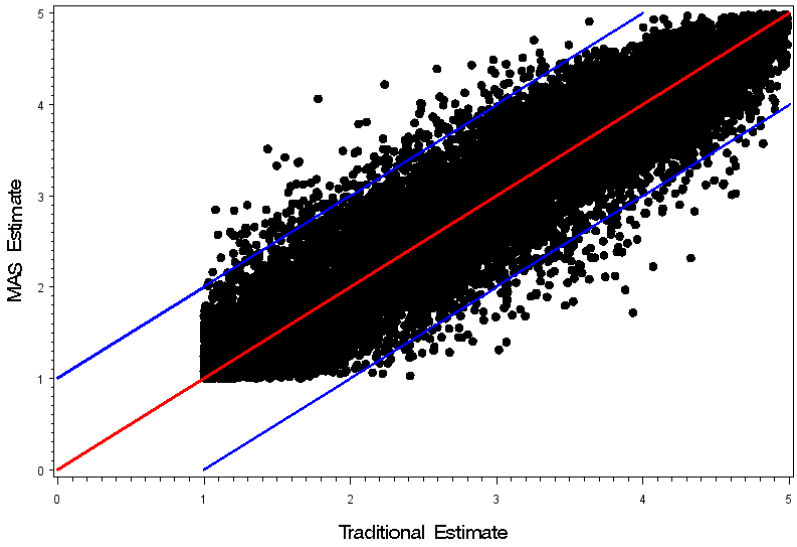

Figure 2: Knowledge--Importance (5-point items)

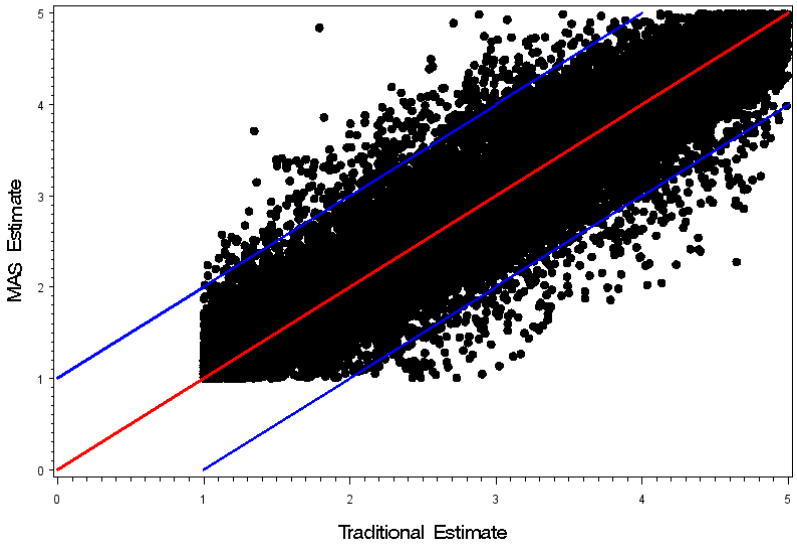

Figure 3: Work Context (5-point items)

3.1.2 Seven-Point Estimates

For the level questions related to work activities, we found that 94% of the estimates were not substantially different (Figure 4). Similarly 97% of the estimates for the level items related to knowledge were not substantially different (Figure 5).

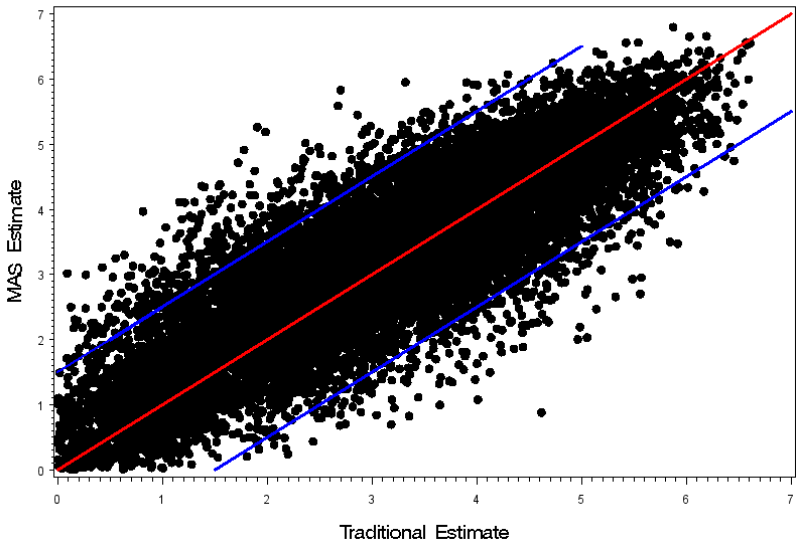

Figure 4: Work Activities--Level (7-point items)

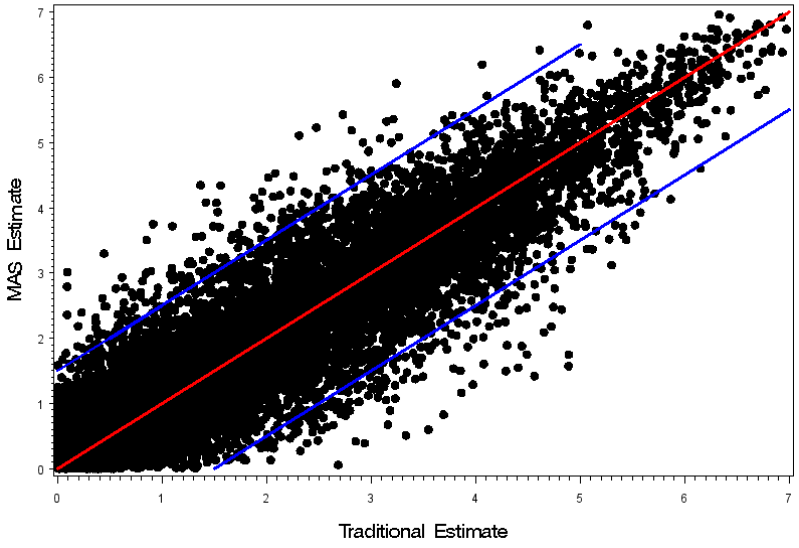

Figure 5: Knowledge--Level (7-point items)

3.2 Comparison of Burden

In order to determine whether there were any differences in the burden of fielding occupations between the traditional paradigm and MAS, we first compared the number of questionnaires shipped to establishments. As shown in Figure 6, we found that fewer questionnaires were shipped for 197 of the 332 occupations under MAS (59.3%). There were three occupations for which over 850 questionnaires were shipped under the traditional paradigm. These were secretaries except legal, medical and executive; food preparation workers; and network and computer systems administrators. Under MAS we shipped X, Y, and Z respectively.

We also compared the distribution of questionnaires shipped for all 332 occupations under both the traditional paradigm and MAS (Figure 7). Under the traditional paradigm the mean number of questionnaires shipped for an occupation was 210.5 (median=157) with a minimum of 69 and a maximum of 1,070. Under MAS the mean number of questionnaires shipped for an occupation was 152.6 (median=149) with a minimum of 64 and a maximum of 345. The standard deviation between the mean number of questionnaires shipped per occupation was 157.3 for the traditional paradigm and 41.9 for MAS. The difference between the two means was 57.9 questionnaires, which was statistically significant ($p < .0001$).

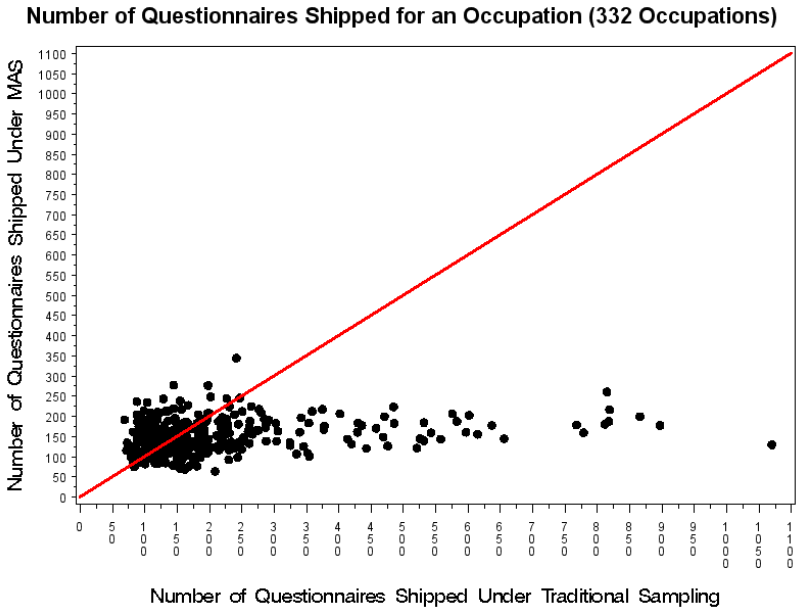

Figure 6: Number of Questionnaires Shipped for an Occupation (332 Occupations)

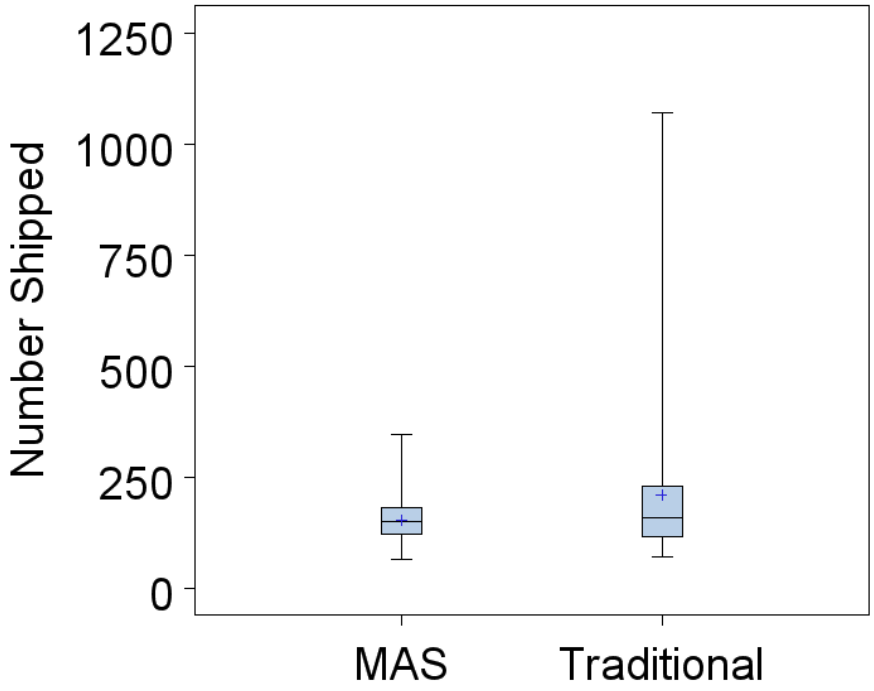

Figure 7: Distribution of Shipped Questionnaires

As another measure of burden, we compared the distribution of received questionnaires (Figure 8). Under the traditional paradigm a median of 100 questionnaires per occupation were received with a minimum of 55 and a maximum of 640. Under MAS we received a median of 98 questionnaires with a minimum of 52 and a maximum of 247. We sent out 11,500 fewer questionnaires under MAS than the traditional paradigm (34,137 vs 45,643). This reduction resulted in an estimated savings of 5,753 hours in employee burden.

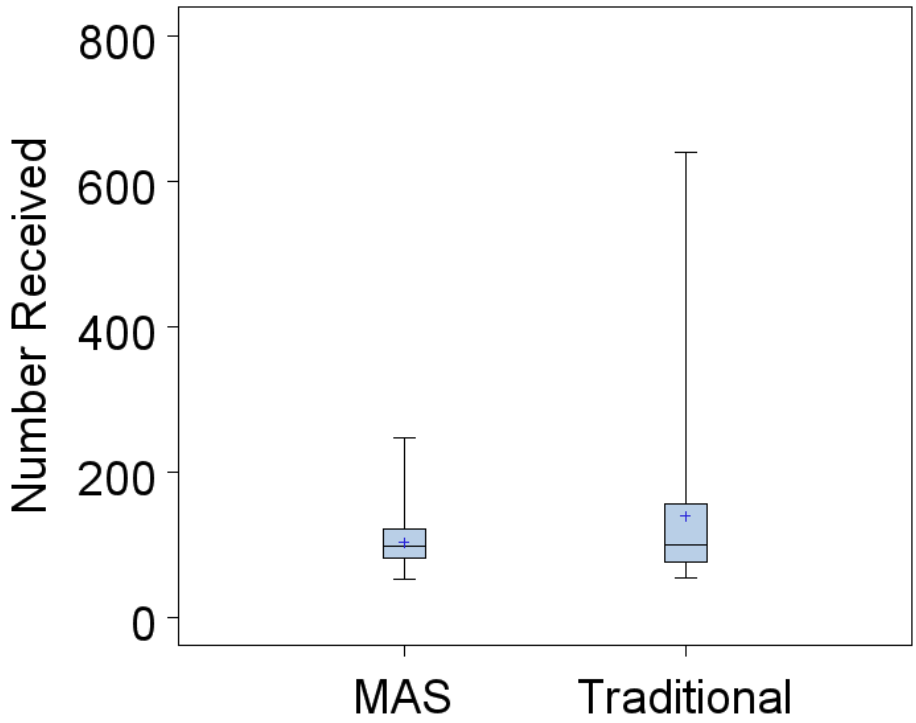

Figure 8: Distribution of Received Questionnaires

4. Discussion

4.1 Comparison of Empirical Results to Simulation

The simulation study that was conducted prior to the implementation of MAS predicted that approximately 99.5% of all estimates (both 5-point and 7-point) would be within the substantive threshold. The simulation also predicted a 58.5% reduction in the number employee burden hours. The comparison of the empirical results to date and the predicted simulation results are shown in Table 4.

Table 4: Empirical Results Compared to the Simulation

	Simulation	Empirical Results
5-Point Questions: Percentage of Estimates Within the 1 Point Confidence Band	99.5%	~ 95%
7-Point Questions: Percentage of Estimates Within the 1.5 points Confidence Band	99.5%	~ 94%
Change in Employee Burden	- 58.5%	-26.2%

4.2 Discussion

It does not appear that the introduction of MAS has introduced an appreciable amount of substantive bias in the estimates. At least 95% of the 5-point estimates were within +/- 1.0 point; and at least 94% of the 7-point estimates were within +/- 1.5 points. The

simulation conducted prior to the introduction of MAS predicted an even higher percentage of the estimates to be within the substantive threshold, however, the empirical results are still indicate that the integrity of the estimates are preserved under MAS.

MAS has reduced the burden to both establishments and employees and this reduction is a result of reducing the variability in the number of questions sent and received. This cost savings as measured in burden hours to the employee is approximately 5,753 hours.

References

Berzofsky, M. E., Welch, B., McRitchie, S., & Williams, R. (2007, June). Improving efficiency in a complex establishment survey design: The O*NET data collection program. In *Proceedings of the International Conference on Establishment Surveys (ICES-III)*. http://www.amstat.org/meetings/ices/2007/index.cfm?fuseaction=presentations_list

Berzofsky, M., Welch, B., Williams, R., and Biemer, P. (2008). Using a Model-Aided Sampling Paradigm Instead of a Traditional Sampling Paradigm in aNationally Representative Establishment Survey. RTI Press publication No. MR-0004-0802. Research Triangle Park, NC: RTI International. Retrieved [date] from http://www.rti.org/rtipress.