# EXHIBIT B

ORACLE USA, INC., ET AL

v.

SAP AG, ET AL

CASE No. 07-CV-01658

# EXPERT REPORT OF BRUCE D. SPENCER

MARCH 17, 2010

BRUCE D. SPENCER, Ph.D.

Subject to Protective Order
Confidential Information

## 1. Executive Summary

1.1. This is a rebuttal to the expert report of Dr. Levy: *Expert Report of Daniel S. Levy, Ph.D.,* dated November 16, 2009, including his corrected reports dated February 5, 2010 and February 12, 2010. Dr. Levy attempted to accomplish three things. First, he attempted to design a "statistically valid sample" of TN's PeopleSoft HRMS payroll tax and regulatory Updates or Fixes. Second, he attempted to select samples of fixes – one sample of Critical Support Fixes and one sample of Retrofit Fixes – according to his design specifications. Third, he attempted to correctly formulate statistics to calculate and report, and he attempted to calculate them correctly. Dr. Levy concludes his report with the statements that

> "In this report, I have applied standard statistical theory to the question at hand. I have discussed the reasons that sampling is appropriate in this particular setting. I have presented my results above. These results are based on standard statistical formulas that are used in sampling situations." (p. 36)

1.2. As discussed below in detail, his application of standard statistical theory was questionable in some cases and simply wrong in others. Dr. Levy chose an inefficient sample design and inefficient estimators. He chose sample sizes that led to large levels of sampling variability by his own formulas. He made serious mistakes in selection of samples, he did not adequately document how he selected his most important samples, and he made serious mistakes in choosing which "standard statistical formulas" to use. When Dr. Levy presented his estimates of averages, totals, and ratios based on his sample, he presented estimates that did not take into account the level of sampling error.

1.3. Having made the decision to sample, and thus to introduce sampling error, Dr. Levy reports estimates of population averages, totals, and ratios without adjusting for sampling error. For example, he reports "90% confidence intervals" that typically are of the form "estimate plus or minus margin of error". His estimates typically are the middle of his confidence interval. If the margin of error were to be twice as large, then based on how he presented his findings in his report, it appears that Dr. Levy's estimate would be unchanged – i.e., it would still be the center of the confidence interval. Furthermore, he does not acknowledge the potential for measurement error in the underlying data on which his estimates are based.

1.4. Dr. Levy did not adequately justify the appropriateness for sampling. Not even a rudimentary cost-benefit analysis was reported. Based on Plaintiffs' alleged damages, the financial stakes in this case are quite high, in the billions of dollars. In that context and depending on how Plaintiffs intend to use Dr. Levy's estimates, a modest sampling error can translate to a very large financial consequence. Yet, Dr. Levy considered only the person-hours associated with the data collection in analyzing why a sample should be done, and his criteria underlying his choice of sample sizes were also formulated without regard to the financial stakes associated with Plaintiffs' use of his estimates from the sample. Dr. Levy sought sample sizes such that when the 90% confidence interval was of the form "estimate plus or minus margin of error", his margin of error would not exceed 25% of the estimate. That is a large margin of error when the financial stakes are high, and by his own accounting that margin of error was exceeded a number of times by large amounts.

1.5. Dr. Levy chose to use simple random sampling with replacement, which allows for duplication of observations, instead of the more common simple random sampling without replacement. Having chosen to use simple random sampling with replacement, Dr. Levy chose to estimate the population average by the sample average, counting duplicates in the sample multiple times. This is not good statistical practice, and his estimators of averages and totals are described in the statistical literature as inadmissible.

1.6. Dr. Levy did not keep a record of the process or processes by which he selected the Retrofit and Critical Support samples, only the outcomes of the selection process. This lack of documentation makes it difficult to confirm that the samples were randomly selected as Dr. Levy claims. Statistical analysis shows that his Critical Support sample was far from what would tend to arise from sampling from the Critical Support population if the sample was in fact selected with simple random sampling as he claims it was.

1.7. Dr. Levy made a host of errors related to standard errors and confidence intervals. He used the wrong formula for estimating standard errors for his estimator, he offered the wrong interpretation of confidence intervals, he failed to check the validity of his use of the normal approximation to develop his confidence intervals, and he used the wrong mathematical distribution to form his confidence intervals. Although he claims that his confidence intervals have 90% coverage rates, or non-coverage rates of 10%, analysis shows that for some measures the non-coverage rates are many times greater than claimed.

1.8. In an attempt to improve some of his confidence intervals, which were on their face problematic, Dr. Levy used a method called bootstrapping. Bootstrapping involves taking the original sample and selecting constant size subsamples from it 10,000 times independently and analyzing the results. Dr. Levy's method for selecting bootstrap samples appears to be seriously flawed in at least three ways. One member of his original sample had no chance of appearing in the 10,000 samples. The subsamples were not constant size. The subsamples were not selected independently.

1.9. Dr. Levy attempted to use a method known as the bootstrap to improve his confidence intervals. As Dr. Levy intended to use it, the method involves taking the original sample as if it were the population, sampling repeatedly and independently from it according to the original sample design to generate a large number of estimates, and then constructing an interval from the observed distribution of the estimates. Unlike his original sample selection, Dr. Levy did provide documentation of the process by which he selected the bootstrap samples. A number of errors are apparent, including failure to make his samples independent of each other, failure to generate constant size samples, and failure to see that the samples were selected according to his original sample design.

1.10. In light of the foregoing, Dr. Levy has failed to provide adequate support to establish the reliability of his estimates.

## 2. Introduction – Background and Qualifications

2.1. I have a Ph.D. in Statistics from Yale University. I am a Professor and am currently the Chair of Northwestern University's Department of Statistics, of which I have been a member since I helped start it in 1987. I have chaired the Department for a total of 17 years. In addition to teaching numerous graduate classes related to statistical sampling, I have been engaged in academic research and written articles relating to sampling theories, techniques, and methods. I have served on numerous advisory committees and consultancies to Federal government agencies on matters related to statistical sampling, such as the Department of Justice, Department of Education, Census Bureau, Department of Energy, and the Government Accountability Office (GAO), and also to high level non-government groups such as the National Academy of Sciences, National Academy of Education, and National Institute of Statistical Sciences. I also conducted extensive research into the cost-benefit analysis of statistical data collections (including samples), as evidenced by my 1980 book, *Benefit-Cost Analysis of Data Used to Allocate Funds*, and seven papers and book chapters on the topic.[1] Last fall, I was invited by the Census Bureau to participate in a meeting of experts to help the Census Bureau develop plans for the 2020 census, and during that meeting the Director of the Census Bureau asked me to talk about cost-benefit analysis.

2.2. I have also been involved in several large scale public studies utilizing various methods and techniques of statistical sampling including the National Educational Longitudinal Study of the High School Class of 1988 (NELS:88), the National Assessment of Educational Progress, and the National Immunization Survey. I have carried out private sector work on samples for companies including Amgen and

[1] The book is item B3 in my CV and the articles and chapters are items A8, A11, A15, A20, A24, A34, and A51 in my CV. The CV appears in Appendix 2.

BlueCross BlueShield Association. As part of these projects, I was intimately involved in developing, drawing, testing, and analyzing samples to ensure valid results.

2.3. Based on these experiences, I have an extensive understanding of statistical sampling theories, methodologies, and techniques as well as practical experience with designing, drawing, testing, and analyzing samples. I also have extensive understanding of cost-benefit analysis as applied to sample designs. My CV is provided in Appendix 2.

2.4. My billing rate for this matter is $650 per hour. My compensation is not contingent on my opinions or the outcome of the proceedings.

## 3. Assignment

3.1. I was asked to review and evaluate the *Expert Report of Daniel S. Levy, Ph.D.,* dated November 16, 2009 ("Levy"), including his corrected reports dated February 5, 2010 ("Levy-R1) and February 12, 2010 ("Levy-R2") and the AACG Expert Report Errata Dated February 5, 2010 ("Levy-Errata-1") and February 12, 2010 ("Levy-Errata-2"). I was asked to review and analyze materials from *Analysis of SAP TN's Collection and Use of Oracle Software and Related Materials*, dated November 16, 2009, produced by Mandiant Corporation, including corrections and revisions produced on January 22, 2010 and February 12, 2010 ("Mandiant"), to the extent that it relates to Dr. Levy's work. Dr. Levy's work is completely dependent on information he obtained from Mandiant. Thus, errors in relevant parts of Mandiant's data will cause inaccuracies in Levy's statistics. I was provided with certain underlying supporting materials as well, as described in Appendix 1. I was also asked to review and analyze materials from FT Works, Paul Meyer, Paul C. Pinto, and Douglas Gary Lichtman for context and to the extent that they relate to Dr. Levy's work. These also are listed in Appendix 1.

3.2. This report and all of the facts and opinions contained herein are based on information gleaned from the written documents provided to me. I reserve the right to modify this report and any of my findings and opinions based on additional information that becomes available to Defendants, e.g., from depositions, and I reserve the right to respond to any additional or amended reports and/or supporting material in this matter that become available.

3.3. Due to preplanned travel out of the country between March 3 and March 26, 2010 that was scheduled based on the original expert rebuttal deadlines in the case, I am unavailable to substantively review any additional supplements or additions or revisions to Oracle's expert reports until after March 26, 2010.

## 4. Materials Reviewed

4.1. I read the Levy report and the Mandiant report, including the appendixes to those reports. I also reviewed the materials from FT Works, Paul Meyer, Paul C. Pinto, and Douglas Gary Lichtman for context and to the extent that they relate to Dr. Levy's work as listed in Appendix 1. For the Mandiant report and related materials, I relied only on the latest versions of the files that were provided to me as of March 2, 2010. The list of materials is subject to all clawed-back and replaced files as of March 2, 2010.

4.2. A list of materials I received for review is provided in Appendix 1.

## 5. Apparent Purpose of Levy's Work

5.1. Dr. Levy attempted to accomplish three things. First, he attempted to design a "statistically valid sample"[2] of TN's PeopleSoft HRMS payroll tax and regulatory Updates from lists of Updates provided to him by Mr. Mandia. There is some ambiguity about whether Dr. Levy attempted to sample Updates or Fixes. In the same paragraph in his report he states that he was "retained by counsel for the Plaintiffs . . . to design a

---

[2] Levy states that he was "retained by counsel for the Plaintiffs . . . to design a statistically valid sample of [ ] TN's PeopleSoft HRMS payroll tax and regulatory Updates that can be used to scientifically estimate the number of Fixes delivered to customers by [ ] TN that infringe Oracle copyrights or otherwise resulted from impermissible cross-use of Oracle's software." (Levy, 7)

statistically valid sample of [ ] TN's PeopleSoft HRMS payroll tax and regulatory *Updates*" (emphasis added) and "My role in this engagement is to generate a random sample of *Fixes* to be reviewed and to calculate population and sample statistics for a number of measures. . ." (emphasis added).[3] Levy makes the following distinction: "From my conversations with Mr. Mandia, I understand that an Update is a group of Fixes delivered together in a single deliverable, either by Oracle or SAP TN."[4]

5.2. Second, he attempted to select samples of fixes – one sample of Critical Support Fixes and one sample of Retrofit Fixes – according to his design specifications. As demonstrated below, there are significant issues associated with how he selected his samples.

5.3. Third, he attempted to correctly formulate statistics to calculate and report and he attempted to calculate them correctly. As demonstrated below, there are errors in his statistical analyses.

5.4. Dr. Levy concludes his report with the statements that

> "In this report, I have applied standard statistical theory to the question at hand. I have discussed the reasons that sampling is appropriate in this particular setting. I have presented my results above. These results are based on standard statistical formulas that are used in sampling situations."[5]

As discussed below, his application of standard statistical theory was questionable in some cases and simply wrong in others. He made serious mistakes in selection of samples, he did not adequately document how he selected his most important samples,

[3] Levy, 7.

[4] Levy, 8, note 27.

[5] Levy, 36.

and he made serious mistakes in choosing which "standard statistical formulas" to use. Specifically, Dr. Levy:

1. chose a sample design that was simple random sampling with replacement instead of simple random sampling without replacement or stratified simple random sampling, which would have yielded more accuracy for the same sample size (see paragraphs 6.2, 7.3);
2. failed to adequately document the process by which he selected his samples (paragraph 6.9);
3. selected a Critical Support sample that is extreme compared to more than the great majority of possible simple random samples that could have been selected (paragraphs 6.12-6.14);
4. used the wrong estimator for his sample – his estimates of population averages and totals are inadmissible (paragraphs 6.17-6.22);
5. used the wrong formula for estimating standard errors for his estimator (paragraph 6.24);
6. offered the wrong interpretation of confidence intervals (paragraph 6.25);
7. failed to check whether the assumptions underlying his use of the normal approximation were valid (paragraphs 6.32-6.33);
8. used a mathematical distribution to form his confidence intervals that would be wrong even if the assumptions underlying his use of the normal approximation were valid (paragraphs 6.27-6.32); and
9. generated his samples incorrectly when he computed confidence intervals by "sampl[ing] repeatedly from the data to estimate the upper and lower bounds"[6] (paragraphs 6.37-6.44).

In addition to presenting his "results", Dr. Levy presents his opinions. The latter are discussed in section 8, below.

---

[6] Levy, 40.

## 6. Tasks that Dr. Levy Apparently Performed

6.1. As far as I can reconstruct from Dr. Levy's report, he received a list of Critical Support Fixes and Retrofit Fixes from Mr. Mandia. The list was separated into two lists, one for Critical Support Fixes and the other Retrofit Fixes. The lists are in the spreadsheet "Fix_IDs_For_AA_with Sampling Order.xlsx", tabs "Critical" and "Retrofit" respectively, and the spreadsheet is contained in the folder ORCLX-AACG-000005. There are 1386 Critical Support Fixes in the list and 223 Retrofit Fixes.

6.2. Dr. Levy chose to select samples using a method known as simple random sampling.[7] Moreover, Dr. Levy selected the sample using "with replacement" sampling rather than "without replacement" sampling. In sampling "with replacement", the same item can be selected multiple times for the sample. A sample selected without replacement is better. One gets more information from a sample that is selected "without replacement" because every item in the sample is distinct and there is no duplication. For example, if a sample of size 223 from the Retrofit population had been selected without replacement, there would be zero sampling error because every item in the population would be in the sample. That would typically not be the case if the sample were selected with replacement, because some items would likely be omitted entirely while others would appear multiple times. The formulas for the precision of estimates from simple random samples show higher levels of precision when the sample is selected without replacement than when it is selected with replacement.[8]

6.3. The only justification Dr. Levy offers for use of sampling is that

[7] Dr. Levy's report only refers to the design as random sampling, but some of the formulas he uses (sometimes incorrectly) are only applicable to simple random sampling. Furthermore, the method he used to select the sample – if carried out correctly – would yield a simple random sample selected with replacement. Simple random sampling is a particular form of random sampling.

[8] William G. Cochran (1977) *Sampling Techniques. 3rd Edition.* New York: Wiley ("Cochran"), p. 30; Carl-Erik Särndal, Bengt Swensson, and Jan Wretman (1992) *Model Assisted Survey Sampling*, New York: Springer ("Särndal et al."), p. 73.

> "In conversations with me, Mr. Mandia thought that it would require thousands of hours of time by highly trained computer forensic staff to capture data for some groups of measures across the entire population of Fixes."[9]

Based on Plaintiffs' alleged damages, the financial stakes in this case are quite high, in the billions of dollars according to Plaintiffs' expert Paul K. Meyer.[10] By carrying out a sample rather than a census of populations of Updates/Fixes, Plaintiffs are accepting some sampling error. As discussed in paragraph 6.49 below, Dr. Levy's own estimates show quite large levels of sampling error, sometimes well exceeding 50% of the estimated total in the population. The consequences of sampling error could be large, depending on the factual context in which Dr. Levy's estimates will be used in this case. For example, when compared with the significant damages Plaintiffs seek in this case, the cost savings from sampling may be modest in comparison. Dr. Levy's report contains no cost-benefit analysis related to Plaintiffs' decision to analyze a sample of the population rather than the entire population. In the context of the damages Plaintiffs seek in this case, I believe that a cost-benefit analysis should have been done. A cost-benefit analysis might have led to much-increased sample sizes, which if large enough it would then lead to the practical decision of dispensing with sampling altogether.

6.4. At some point, either Dr. Levy, Mr. Mandia, or someone else made a decision that "Fixes with a status of "Cancelled," Research Only," or "0". . . were uninformative for the purposes of the measures of interest."[11] Excluding these Fixes from the populations reduces their sizes from 1386 to 973 (Critical) and from 223 to 212 (Retrofit).

---

[9] Levy R2 at 9, 13, notes 29 and 33.

[10] Expert Report of Paul K. Meyer, Navigant Consulting, Inc., November 16, 2009, revised December 2, 2009, e.g., page 87. Also Supplemental Expert Report of Paul K. Meyer, TM Financial Forensics, LLC. February 23, 2010.

[11] Levy 15, note 36.

6.5. Dr. Levy conducted analyses to determine what sample sizes to use. He settled on sample sizes of 46 for the Retrofit Fixes and 238 for the Critical Support Fixes. His analysis to choose the sample sizes is revealing of flaws in logic and statistical analysis. He states that:

> "The sample size was determined based on examination of the data available for two measures of interest for which Mr. Mandia was able to collect data for the entire population of Fixes. Mr. Mandia provided this data to me as ORCLX-MAN-000060, the Excel workbook containing the results of his findings for the two measures of interest across the entire population of Fixes in the Retrofit and Critical Support populations. Additionally, there was discussion of the characteristics of one of the measures for which it was extremely costly to gather data; this measure was *the number of Environments used in the development or testing of the Fix, as identified in the development, test or other documentation.*" [emphasis in original]
>
> ". . . Measure 116, the number of Environments used in the development or testing of the Fix, as identified in development, test, and other documentation, was the basis for determining the sample size. The assumed averages and standard deviations used to calculate the sample size were based on simulated data for measure 116, where it was assumed that measure 116 should be zero whenever measure 104 is zero and that the distribution of the non-zero values measure 116 would be similar to that of measure 115."
>
> ". . . I was asked by counsel to calculate samples sizes based on a 90% confidence level and 50% precision range for this measure, which yielded a sample size of 46 for Retrofit and 238 for Critical Support Fixes."[12]

[12] Levy, 15-16.

The "precision range" is defined by Dr. Levy as the difference between the upper endpoint of his 90% confidence interval minus the lower end of the interval, expressed as a percentage of his estimate.[13] A small precision range is desirable, and a large precision range is not desirable. Dr. Levy did not make it completely clear how he constructed his confidence intervals for his sample size analysis, but, for the method that it appears that he used, the precision range will be proportional to the coefficient of variation, which is the ratio of the standard deviation to the average.[14] He based his analysis on measure 115, but other measures have higher coefficients of variation and had he based his analysis on those, his sample size specifications would have been different. As shown in paragraph 6.49 below, some of the precision ranges associated with the confidence intervals produced by Dr. Levy are around 100% or larger, well in excess of the 50% precision range he was asked to achieve.

6.6. Dr. Levy's statement that he "was asked by counsel to calculate samples sizes based on a 90% confidence level and 50% precision range for this measure"[15] makes clear that counsel told him to use the combination of "a 90% confidence level and 50% precision

---

[13] Levy-R2, 12 defines precision range as follows: "In order to design the sample size, she decides to use a 90% confidence level and a 20% precision range; that is, she wants to be able to say that if she sampled from this population repeatedly, 90% of the time the true number of computers would be within plus or minus 10% of her result."

[14] Dr. Levy does not appear to have reported the formulas that he used to calculate precision ranges. The worksheet showing his calculations shows only the numerical values resulting from his formulas rather than the formulas themselves; worksheet "Sample Size 116" in spreadsheet "Sample_Selection.xlsx" in ORCLX-AACG-000003. If the confidence intervals are based on the normal approximation, the lower endpoint of the interval equals the sample average minus a factor times the estimate of standard error and the upper endpoint equals the sample average plus a factor times the estimate of standard error, so the difference between the two endpoints equals 2 times the product of the factor times the standard error. If the formula for standard error does not contain a finite population correction, the estimate of standard error equals the standard deviation divided by the square root of the sample size, and in that case the precision range is equal to the product of 2 times the factor times the coefficient of variation times the reciprocal of the square root of the sample size. This implies that the precision range is proportional to the coefficient of variation. Dr. Levy did not make clear whether he used the finite population correction (defined in Levy, page 38) in his determination of sample size. If he did not, the formula is slightly more complicated, but the general pattern is similar.

[15] Levy, 16.

range". This is quite a large range of uncertainty to deliberately aim for in a case where Plaintiffs allegations place large sums of money at stake, and as shown in paragraph 6.49 below even the meager standard Plaintiffs' counsel told Dr. Levy to use was often not achieved.

6.7. Because the methodological choice to base the sample sizes on certain properties of confidence intervals was inherently statistical, I infer that Dr. Levy made that choice rather than counsel. That choice is questionable, however, in light of Cochran's view that

> "A more logical approach to the determination of sample size can sometimes be developed when a practical decision is to be made from the results of the sample."[16]

Cochran made that statement as the lead sentence in the section of his book that discusses the use of decision theory or cost-benefit analysis to determine sample sizes.[17] The difference between the cost-benefit analysis approach and Dr. Levy's is that the former takes into account the practical consequences, including financial consequences, of sampling error in the estimates based on the sample.

6.8. Dr. Levy used Excel software to generate the samples. With respect to the Retrofit Fixes, the sample was selected from the full list of 223 Retrofit Fixes (including those with a status of "Cancelled," Research Only," or "0"). Each Retrofit Fix in that list was given an item number from 1 up through 223. Then Dr. Levy used Microsoft Excel software to choose a whole number from 1 to 223.[18] The first number shown in his list is 8. That means that the Retrofit Fix with item number 8 was selected first. Then he repeated the process 223 times. The first 46 item numbers (including duplicates) would

---

[16] Cochran, 83.

[17] Cochran, 83-85.

[18] In essence, one does this by having Excel generate a random number between 0 and 1 (but not including 1), multiplying the number by 223, and then rounding up to the nearest whole number.

comprise the sample of Retrofit Fixes. If an item number chosen for the sample corresponded to a Fix with a status of "Cancelled," Research Only," or "0" then it was skipped, and the next item number generated would be used instead. The treatment for the Critical Support Fixes was similar, except that instead of using the Retrofit population size and sample size, 223 and 46, the Critical Support population size of 1386 and sample size of 238 were used.

6.9. In my more than 25 years of statistical consulting experience in drawing samples, I have consistently observed the generally accepted practice of fully documenting the sample selection process. That documentation includes provision of sufficient information such that the sample selection process can be duplicated exactly by anyone else. There are serious questions about whether the samples put forth by Dr. Levy are properly generated random samples obtained from Excel. It is a significant problem that Dr. Levy did not produce a copy of a program that can be run to duplicate his sample selection process. Plaintiffs produced item numbers only – the results of the supposed valid random sampling – but not a proper record or an audit trail of the process. Apparently Dr. Levy did not maintain one.[19] When a sample is selected by a qualified statistician for use in litigation or in other important situations, it is standard practice for the statistician to select the sample in such a way that the sample selection can be reproduced. This is simple to do, and indeed Dr. Levy did follow this practice in a later analytic step that involved "10,000 repeated draws from the sample".[20] That he neglected to maintain a record of the process for the main task with which he was charged – selecting the Critical Support and Retrofit samples – is stunning. The entire set of data with which he is working is derived from the auditing of computer files. Given that Dr. Levy apparently worked closely with a "computer forensics expert" (Mr. Kevin Mandia),

---

[19] Appendix 3 contains a copy of email correspondence between JonesDay and Plaintiffs' counsel on this point, which is Bates-labeled SAP-SPE-000001 through SAP-SPE-000006.

[20] Levy, 40. His method for generating the 10,000 draws has its own significant problems, as discussed below (paragraphs 6.34 – 6.44). However, it is because Dr. Levy's latter selection method – unlike his method for selecting the Retrofit and Critical Support samples – is reproducible that its flaws are so readily apparent.

it is difficult to comprehend why he would select his samples in such a way that the process was hidden and could not be reproduced.

6.10. Dr. Levy received data from Mr. Mandia on a number of measures that Mandiant had determined for all of the fixes in the population, Critical Support or Retrofit as the case may be. Those data permit checks to be conducted to see how likely it is that a sample such as Dr. Levy's would be selected with the procedure he claims to have used. As will be shown (paragraphs 6.12 - 6.16), Dr. Levy's sample is discrepant on a number of dimensions. I instructed Dr. Walter Vandaele and his colleagues to run a simulation 10,000 times to replicate the sampling procedure that Dr. Levy claims to have used.[21] Fewer than 1 in 10 of the 10,000 samples generated in the simulation were as discrepant as Dr. Levy's Critical Support sample. I did not find the same degree of discrepancy for the Retrofit sample. The Retrofit sample was smaller than the Critical Support sample, with size 46 compared to 238, and it is harder to detect deviations from chance for a small sample.

6.11. A simple example demonstrates both (i) how one can test whether a sample was selected with simple random sampling and also (ii) why it is harder to show deviations from chance for a smaller sample. Consider a population of size N that consists of equal numbers of red and white items. Suppose that someone picks a sample of size n that contains only red items and no white items and claims that the sample was selected as a simple random sample with replacement. The chances of getting all red items depends on the sample size. For a simple random sample drawn with replacement from the population of equal numbers of red and white items described above, if the sample consists of just 2 items then the probability of getting all red items is 25%. The probability of getting all red items is 12.5% if the sample is of size 3, and the probability of getting all red items is 6.25% if the sample size is 4. So, for a sample of size 4 from this population we could say that the chances of getting such an extreme sample – all red

[21] Appendix 4 contains a copy of the written instructions I gave to Dr. Vandaele and his colleagues at LECG, which is Bates-labeled SAP-SPE-000007 through SAP-SPE-000009. In some cases I supplemented the written instructions with oral instructions.

items – is small enough to raise serious doubts about how the sample was really selected, because such extreme samples are not likely to arise if the sample were chosen as claimed. For smaller size samples it would be harder to support an assertion that the sample was not selected as claimed, since samples as extreme as the one obtained could occur more than 10% of the time.

6.12. The fact that 26 measures are known for all items in the population allows us to check how plausible it is that the sample really was selected with [simple] random sampling as Dr. Levy says it was. For the 26 numerical measures that Mandiant determined for the whole population and for the Critical Support sample, I compared the sample average with the population average.[22] In order to put the differences between the sample average and the population average on the same scale for each measure, I divided each difference by the standard error for the sample average.[23] The resulting statistics are sometimes called standard scores, z-scores, or z statistics. My calculation is a standard practice in the field of statistics. The z statistic is interpreted as a measure of the actual discrepancy between a given sample average and the population average.

6.13. If the sample were selected as Dr. Levy claims, and if the values of the measure in the population were distributed according to a bell-shaped curve, or what is called the

---

[22] The numerical measures available for every item in the population are those numbered 101, 104-113, 115, 118, 121, 122, 125, 127, 130, 133, 135-138, and 142-144. The values for these measures assigned by Mandiant for all items in the population are shown in the worksheet "Population" in the spreadsheet "Fix Measures to AACG.xls", which is contained in ORCLX-MAN-000061. The items selected in the Critical Support sample and in the Retrofit sample are identified, respectively, in the worksheets "Critical_Support_Sample" and "Retrofit_Sample" in that spreadsheet. This analysis does not assume that the data values in the spreadsheet are true or correct. The analysis is concerned with discrepancy between the sample and the population. Probability theory tells us that the discrepancies should not be larger than what chance dictates, whether or not the underlying data values are true.

[23] The sample average will vary from one sample to another, depending on which sample was selected. The dispersion of the sample average across possible simple random samples selected with replacement can be calculated because the data are known for the full population, and the dispersion is quantified by what is called the standard error. See paragraph 6.23 for further discussion.

"normal distribution" or the "Gaussian distribution", the z statistic for a single sample would have a 90% chance of falling between −1.644853627 and +1.644853627.[24] For purposes of this analysis, I will call a z statistic "discrepant" if it falls outside the range from −1.644853627 to +1.644853627.[25] I use so many decimal places here to avoid ambiguity.

6.14. For Dr. Levy's Critical Support sample, 8 of the 26 numerical measures Mandiant determined for the full population had a discrepant z statistic. To see whether such a large number of discrepancies is consistent with what one gets from actual simple random sampling from the population, I directed Dr. Walter Vandaele and his colleagues to select a simple random sample of size 238 with replacement from the Critical Support population and to record how many measures had discrepant z statistics. I had them do this not only once, but 10,000 times, such that the samples were selected independently of each other. I refer to this as a simulation. For each sample, the number of measures whose z statistic was discrepant was counted. The average number of measures per sample with a discrepant z statistic was 2.46, compared to 8 in Dr. Levy's sample. Only

[24] John A. Rice (2007) *Mathematical Statistics and Data Analysis. 3rd Edition.*, Belmont, California: Duxbury Press, Table 2C on p. A7 with supporting discussion on pp. 59-60 shows the result to about 2 decimal places after interpolation. For more decimal places one can use Excel with the cell entries "=NORMINV(0.05,0,1)" and "=NORMINV(0.95,0,1)". Paragraph 6.30 shows that when the sample really is selected with simple random sampling with replacement, the z statistics for the 26 measures known for the entire population do have an approximately 90% chance of falling in the range from −1.644853627 to +1.644853627

[25] I picked this range from −1.644853627 to +1.644853627 to be consistent with the upper and lower bounds as used by Dr. Levy in his analysis. He chose his upper and lower bounds in an attempt to develop 90% confidence intervals. For example, his SAS® program includes the statements "avg_lb_90 = avg - (1.644853627*stderr_avg);" and "avg_ub_90 = avg + (1.644853627*stderr_avg);" to have lower bounds (lb) and upper bounds (ub) computed; see "analysis.sas" in the folder "SAS Code", which is contained in the folder ORCLX-AACG-000004, which is contained in the folder ORCLX-AACG-003. In my analysis, I do not need to make a theoretical assumption about the shape of the distribution because I know the full population for each of the 26 measures I examine.

927 out of the 10,000 samples had 8 or more measures with a discrepant z statistic.[26] That is, fewer than 10% of the samples had as many or more discrepant z statistics as Dr. Levy's sample. Put another way, if one picked a simple random sample of size 238 with replacement from the population of Critical Support Updates/Fixes, the chances of getting as many discrepant z statistics as Dr. Levy did is less than 10%. (As will be discussed below in paragraph 6.26, Dr. Levy is using a threshold of 10% error as a standard.)

6.15. Dr. Levy's Retrofit sample had only 1 measure with a discrepant z statistic, and the chances of getting 1 or more measures with a discrepant z statistic are large, at about 58%.[27]

6.16. To recapitulate, Dr. Levy chose to use simple random sampling with replacement, which allows for duplication of observations, instead of the more common simple random sampling without replacement. He did not keep a record of the process or processes by which he selected the Retrofit and Critical Support samples, but only the outcomes of the selection process. This lack of documentation makes it difficult to confirm that the samples were randomly selected as Dr. Levy claims. Statistical analysis shows that his Critical Support sample was far from what would tend to arise from sampling from the Critical Support population if the sample was in fact selected with simple random sampling as he claims it was.

6.17. Faced with a sample that he had selected with replacement, Dr. Levy had a choice between two different ways of analyzing it. He could have thrown out the duplicates, so that each item in the sample was in the sample only once. Or, he could have kept the duplicates in, so that some items counted multiple times in the sample and others counted

---

[26] Worksheet "Exhibit A" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

[27] There were 5842 out of the 10,000 samples with 1 or more discrepant z statistics, as shown in Worksheet "Exhibit B" in "All Exhibits.xls" located in Appendix 5 and Bates-labeled SAP-SPE-000010. That is 58%.

only once. That Dr. Levy's choice is inferior has been known for more than 50 years.[28] In the statistical literature, Dr. Levy's estimator – the average for all items in the sample, no matter how many times they are duplicated – is classified as "inadmissible".[29] [30]

6.18. For measures that were based in whole on samples, Dr. Levy used SAS® software to estimate averages for the sample. The problem was not with the SAS® software, but rather with the way he used it.

6.19. To estimate the total for the population, he would multiply the sample average by the number of Fixes in the corresponding population, namely 973 (Critical Support) and 212 (Retrofit). Because his estimate of the total is simply the population size times Dr. Levy's estimate of the average, and his estimate of the average is inadmissible, his estimate of the total is similarly inadmissible. To estimate a ratio of two population averages, Dr. Levy uses the ratio of his estimates of the population averages. Consequently, this too does not represent good practice.

6.20. Some measures, such as 117, 120, 128, 129, and 134, were based in part on the sample and in part on the population. As these estimates involve Dr. Levy's inadmissible estimates of population averages, these too are inadmissible.

6.21. Dr. Levy overstates the properties of his estimators. He claims

---

[28] Cochran, 30.

[29] John A. Rice (1995) *Mathematical Statistics and Data Analysis. 2nd Edition*, Belmont, California: Duxbury Press, p. 585 says "An estimate is said to be admissible if it is not strictly dominated by an other estimate." An estimate is strictly dominated if its accuracy is never better and is sometimes worse than that of another estimator.

[30] J. N. K. Rao. (1966) "On the Comparison of Sampling with and without Replacement", *Review of the International Statistical Institute*, Vol. 34, No. 2, pp. 125-138, states on pp. 126-127 that "If a simple random sample of fixed size $n$ is drawn with replacement from a population of size $N$, the usual estimator of the population mean $\bar{Y}$ based on all the units in the sample is $\bar{y}$ [the average of the $n$ values]. Basu (1958) and Des Raj and Khamis (1958) have shown that $\bar{y}$ is inadmissible by proving that the estimator [that is the average of] the . . . distinct units in the sample is uniformly more efficient than $\bar{y}$."

> "The estimates of the population total, the population mean, and their respective standard errors are **scientifically valid estimates of the true values** in the population when the sample is generated randomly." [emphasis added]

and he cites the sampling text Levy and Lemeshow to buttress his claim.[31] However, the cited passage from Levy and Lemeshow states that "estimates resulting from a simple random sample are unbiased estimates of the corresponding population parameters",[32] and Levy and Lemeshow also state "the estimated total in simple random sampling appears to be an unbiased estimate of the true population total."[33] There is a subtle, but important, distinction to make here. When Levy and Lemeshow discuss the "corresponding population parameters" they are referring to the average, total, etc. that would be calculated if the sample consisted of the entire population. For example, suppose that a sample survey was conducted to estimate average income in a population, and everyone in the population would always understate their actual income. No matter how good the sampling method, the estimate from the sample would tend to be too small relative to the true value. Several pages prior to the passage cited by Dr. Levy, Levy and Lemeshow discussed this measurement issue relative to the validity and reliability of findings based on samples,

> "The *reliability* of an estimated population characteristic refers to how reproducible the estimator is over repetitions of the process yielding the estimator. **If we assume that there is no measurement error in the survey, then the reliability of an estimator can be stated in terms of its sampling variance or, equivalently, its standard error.** The smaller the

---

[31] Levy, 21, note 39 gives the citation source "Paul S. Levy and Stanley Lemeshow, Sampling of Populations, Fourth Edition (New York: Wiley & Sons, 2008), 55-58." I refer to this source as "Levy and Lemeshow".

[32] Levy and Lemeshow, 55.

[33] Levy and Lemeshow, 56.

standard error of an estimator, the greater is its reliability. [emphasis in boldface added]

"The *validity* of an estimated population characteristic refers to how the mean of the estimator over repetitions of the process yielding the estimate, differs from the true value of the parameter being estimated. Again, **if we assume that there is no measurement error, the validity of an estimator can be evaluated by examining the bias of the estimator.** The smaller the bias, the greater is the validity."[34] [emphasis in boldface added]

Similarly, Cochran notes that

"Even with estimators that are unbiased in probability sampling, errors of measurement and nonresponse may produce biases in the numbers that we are able to compute from the data."[35]

In Dr. Levy's study, measurement error arises when the numbers that Mandiant provides him are imperfect. It appears that Dr. Levy is making the assumption that there is no measurement error, but he does not state this critically important assumption nor does he offer any justification for making such an assumption. Dr. Levy's report does not mention any evaluation he conducted to verify that there was no measurement error. Thus, even if he could demonstrate that his samples were generated randomly (and as discussed in paragraphs 6.9-6.16 he has not as of this date been able to do so) he could not justify a claim that his estimates of the population total, the population mean, etc. are scientifically valid estimates of the true values in the population. Dr. Levy cannot claim validity simply by assuming it.

[34] Levy and Lemeshow, 35.

[35] Cochran, 12.

6.22. To recapitulate, having chosen to use simple random sampling with replacement, which allows some members of the population being sampled to appear multiple times in the sample although they appear only once in the population, Dr. Levy chose to estimate the population average by the sample average, counting duplicates in the sample multiple times. This is not good statistical practice, and his estimator is described in the statistical literature as inadmissible. Estimators that incorporate his inadmissible estimate of the average are flawed as well. All of his estimates depend on the underlying data being correct, and he has not indicated that he has made any attempt to verify that. His claims that his estimates are scientifically valid estimates of the true values in the population are unjustified.

6.23. In addition to estimates of averages, ratios, and totals, Dr. Levy calculated and reported standard errors, lower bounds, and upper bounds. To understand these terms, we note that the expected value of a statistic calculated from a sample is defined as the average value of the statistic that would occur if the sample were selected repeatedly and independently. The standard error is a widely used measure of the typical size of the error in an estimate based on a sample. It is defined as the square root of the expected squared difference between the statistic and its expected value.

6.24. Although formulas for estimates of the standard error are available for the statistics that Dr. Levy reported, Dr. Levy used the wrong formulas. The standard error for a statistic depends, among other things, on the sampling design used to get data for the statistic. If one uses simple random sampling with replacement, as Dr. Levy did, there is a formula for computing the usual estimate of standard error.[36] If one uses simple

[36] The formula for the square of the estimate of standard error of Dr. Levy's estimate of the population total is shown in Särndal et al., page 73, equation (3.3.24). The formula for the estimate of standard error of Dr. Levy's estimate of the population total is simply the square root of the expression in equation (3.3.24). The formula for the standard error of Dr. Levy's estimate of the population average is simply the ratio of the square root of the expression in equation (3.3.24) to the population size. In most cases, Dr. Levy indicated which formulas he used in his report, but in some cases additional details on the formulas are contained within the SAS® program Dr. Levy used to calculate his results. I asked Dr. Vandaele and his colleagues to transcribe some of the formulas used in Dr.

random sampling without replacement, there is a different formula for estimating the standard error. This formula yields smaller estimates of standard error, because it is based on the assumption that a better sampling design has been used. It is highly inappropriate to use the latter formula for estimating standard error when the sampling design is simple sampling with replacement – yet that is just what Dr. Levy did to estimate standard errors in his November 16, 2009 report.[37] Dr. Levy later corrected the error, without explanation, and issued a revised report. [38]

6.25. Dr. Levy computed and reported "90% confidence intervals" for a number of the population values that he used the sample to estimate. For example, he reports that a

> "90% confidence interval . . . [for the number of] customers [that] received each Retrofit Fix as a First or Identified Deliverable . . . ranges from 5.94 to 8.31. This means that in repeated samples the true value in the population has a 90% chance of falling within the confidence intervals constructed in this way from the sample."[39]

Dr. Levy is incorrect. First, his confidence interval "from 5.94 to 8.31" is constructed incorrectly and relies on assumptions that he does not test and which are questionable in

---

Levy's SAS® programs, and the resulting document is located in Appendix 7 and has Bates numbers SAP-SPE-000032 through SAP-SPE-000038.

[37] Levy, 37, third line from bottom of the page; Levy 38, third line from top; Levy 40, second line.

[38] Levy-R2 (see also Levy-Errata-2) corrects this error, and uses instead the correct formula for estimating the standard error for the estimators used by Dr. Levy if the sampling design is simple random with replacement. The impact of this change in formula is to increase Dr. Levy's estimates of standard error, to decrease his lower bounds, to increase his upper bounds, and to widen his confidence intervals for population averages and totals, with the exception of Measure 124 for Critical Support. For this measure, his revised estimate of standard error would cause the lower bound that he calculated with his "normal approximation" to be negative (Levy-R2, 40), so he used bootstrap methodology to develop his upper and lower bounds and confidence intervals and he chose not to report his estimate of the standard error.

[39] Levy, 3, note 7. Similar claims are made in Levy 13 and 24 (twice).

light of the analysis discussed below in paragraphs 6.32 - 6.34.[40] Second, he confuses the concept of "true value" with "population value". As discussed in paragraph 6.21, Dr. Levy has provided no support for making any claims about truth. If he had selected and analyzed his sample appropriately, he would have grounds for making claims about what his sample would have yielded if it had included 100% of the items in the population, or what is called the "population value". But, to the extent that data provided to Dr. Levy contain errors, the population value is not the true value. There is no supportable basis for Dr. Levy to assert that the sample data values are correct because he received the data from Mr. Mandia, and there is no evidence that he verified that Mr. Mandia made absolutely no errors in his analysis.[41]

6.26. In plain terms, a 90% confidence interval is a range of numbers calculated from the sample that have the property that the population value being estimated will fall within the range 90% of the time if the sample is repeated independently a very large number of times. There is a tradeoff in confidence intervals between the length of the confidence interval and the coverage probability. Dr. Levy (and Plaintiffs' counsel, as discussed in paragraph 6.6) chose 90% for the coverage probability. Had he (and Plaintiffs' counsel) chosen 95%, the intervals would have a greater probability of covering the true value, which is good, but the intervals would also be wider, which is undesirable. Dr. Levy and Plaintiffs' counsel chose to accept a 10% probability for his confidence intervals failing to include the population value.

6.27. Dr. Levy made a number of mistakes in his constructions of confidence intervals for population averages and population totals. The confidence intervals he constructed for population averages and population totals have the form: estimate of population

[40] Although changing 5.94 to 5.79 and 8.31 to 8.46 as in Levy-Errata-2, so that "the 90% confidence interval ranges from 5.79 to 8.46" (Levy-R2, 3, note 7) is an improvement, this revised confidence interval is still incorrectly constructed. This point is further discussed in paragraphs 6.27 - 6.36.

[41] I see nothing in Dr. Levy's report showing that he validated or checked the input data he received from Mr. Mandia. Changes in those inputs will in all likelihood cause changes in the statistics calculated and reported by Dr. Levy. Errors in those inputs will tend to cause errors in Levy's estimates.

quantity plus or minus the product of the estimate of standard error times a factor that is often referred to as a critical value.[42] He refers to confidence intervals of this form as being "based on the normal approximation."[43] Using intervals of this form for population averages and totals is standard in sample surveys and it is widespread as well for ratios. Dr. Levy was right to avoid using the normal approximation for intervals when they would exceed the limits of plausibility, such as proportions being negative or exceeding 100%. However, in his application of the normal approximation for developing confidence intervals he made mistakes regarding each of the three components.

6.28. First, as discussed in paragraphs 6.17-6.20, his estimates of averages and totals were inadmissible, and his estimates of ratios of population averages or totals were based on ratios of inadmissible estimates.

6.29. Second, as discussed in paragraph 6.24, he used the wrong estimates of standard error.[44]

6.30. I directed that an analysis be run to see how well Dr. Levy's confidence intervals would perform if they could be based on the true standard error instead of an estimate of the standard error.[45] Specifically, I directed Dr. Vandaele and his colleagues to run analyses to check on the validity of the confidence intervals. They then performed the analysis for the 26 numerical measures that were available for all fixes in the

---

[42] That is, calculate the estimate of the population quantity and subtract the product of the standard error and the factor to get the lower bound, and then take the estimate and the add the product of the standard error and the factor to get the upper bound.

[43] Levy, 40; Levy-R2, 40. The phrase "normal approximation" refers to the mathematical approximation of a distribution by the Gaussian or normal distribution. The adjective "normal" refers to the distribution and not to the action of using the approximation.

[44] Levy, 37 and Levy-R1, 37 for example, show the incorrect formula, which was used in a very large number of calculations. Only in his second revised report, Levy-R2, was this particular mistake corrected. The errata (Levy-Errata2) is 8 pages long, and as far as I can tell, the corrections solely involve this formula. I counted more than 250 changes to numerical results and 6 changes to formulas and text.

[45] For the purposes of this analysis, Mandiant's measures are taken to be correct.

population.[46] For these measures the average in the population is known, and so sampling is not required. The standard error of the sample average is known as well. However, to test how well Dr. Levy's confidence interval method and other confidence interval methods worked, I directed Dr. Vandaele and colleagues to select a simple random sample with replacement with the same sample size as Dr. Levy, and I directed that they check to see whether the confidence interval included the population average. I had them repeat this procedure 10,000 times for each population, with independent samples selected each time. At my direction, Dr. Vandaele and colleagues constructed confidence intervals as Dr. Levy did, except that the known standard error was used in place of the estimated standard error, and they tested the performance of the intervals. If the theoretical assumptions held perfectly, the confidence intervals should cover the population average about 9,000 times out of 10,000 (or 90%). For the Retrofit sample, the coverage rates ranged from 89.7% to 92.5%,[47] and, for the Critical Support sample, the coverage rates ranged from 89.9% to 91.5%.[48] Thus, when the confidence intervals were based on the true standard error, their performance was fairly consistent with what theory predicts.

6.31. A third mistake made by Dr. Levy was that the critical value that he used to multiply the estimate of standard error by is wrong. The critical value that he chose is based on the normal or Gaussian distribution. This choice would be appropriate if the assumptions behind the normal approximation hold and he actually knew the standard error. If the assumptions behind the normal approximation hold, then because he does not know the actual standard error but only has an estimate of it, he should base his critical value on the "t distribution" or "Student's t distribution".[49] To use the normal

[46] The numerical measures available for every item in the population are those numbered 101, 104-113, 115, 118, 121, 122, 125, 127, 130, 133, 135-138, and 142-144.

[47] Worksheet "Exhibit C" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

[48] Worksheet "Exhibit D" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

[49] The t distribution is often called the "Student's t" distribution because William Sealy Gosset, the author of the 1908 paper analyzing the distribution, wrote the paper under

distribution when one should instead use the t distribution is a mistake that college freshmen and even high school statistics students are taught not to make.[50] The impact of this mistake on the confidence intervals is not as large as the impact of using the incorrect estimate of standard error (paragraph 6.24), although fixing the mistake will improve the coverage rates, as shown below in paragraph 6.34 and Tables 1 and 2.

6.32. Yet, even the assumption that the sampling distribution is the t distribution can be a bad assumption if it is not appropriate for the population being sampled. Statisticians are taught to check their assumptions and not rely on an assumption just because it is commonly used. For example, Cochran's book contains a discussion on "The Validity of the Normal Approximation"[51] in which he points out

> "Failure of the normal approximation occurs mostly when the population contains some extreme individuals who dominate the sample average when they are present."[52]

> "There is no safe general rule as to how large n [the sample size] must be for use of the normal approximation in computing confidence limits. For populations in which the principal deviation from normality consists of marked positive skewness, a crude rule that I have occasionally found

---

used the pseudonym "Student" to hide the fact that he was employed by Guinness Brewery, which did not want its employees publishing papers.

[50] For example, the introductory statistics courses at the Department of Statistics at Northwestern University teach this. Even high school statistics courses teach students to account for uncertainty in the estimate of standard error by using the t distribution instead of what Dr. Levy did, which was to use the normal or Gaussian distribution. The Advanced Placement curriculum for statistics, which can be found at

http://apcentral.collegeboard.com/apc/public/repository/ap08_statistics_coursedesc.pdf,

includes the t distribution (or Student's t distribution, as it is commonly called) in section III.D.7 and it includes confidence intervals for the population average (or mean) in section IV.A.6 of AP Statistics Curriculum.

[51] Cochran, 39-44.

[52] Cochran, 44.

useful is [the sample size should exceed 25 times the square of Fisher's measure of skewness]." [53]

As discussed in paragraph 6.5, I see no evidence that Dr. Levy heeded such cautions in this analysis or even tried to apply Cochran's "crude rule" in determining what size sample to use or deciding whether to rely on the normal approximation. I directed Dr. Vandaele and his colleagues to use SAS® to compute the measure of skewness for each measure in each population as well as the sample, and some of the skewness measures are large.[54] For example, measures 104, 110, and 135 in the Critical Support population had skewness coefficients of 8.7, 10.4, and 10.7, respectively, in the population. Their corresponding skewness coefficients for the sample did not fully reveal the extent of skewness in the population – the sample skewness coefficients were 4.4, 1.7, and 5.6, respectively. Other measures in the Critical Support sample that were not available for the population had large sample skewness coefficients – measure 124 had a skewness of 12.3 and measure 140 had a skewness of 15.4.[55]

6.33 The discussion in paragraphs 6.25- 6.32 show a number of theoretical problems with Dr. Levy's confidence intervals. To check whether the problems are likely to impact his analyses, I directed Dr. Vandaele and his colleagues to run analyses to check on the validity of the confidence intervals. I had them carry out the simulation, as described in paragraph 6.30, and check how often the confidence intervals constructed by Dr. Levy's method covered the population averages. The results are shown in the Tables 1 and 2 for the Retrofit and Critical Support populations, respectively. The measures in each table have been sorted in order of the coverage rates when the confidence intervals are based on the Student's t distribution.

[53] Cochran, 42.

[54] Worksheets "Exhibit E", "Exhibit F", "Exhibit G", "Exhibit H" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

[55] Worksheets "Exhibit E" and "Exhibit F" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

6.34 According to Dr. Levy's claims for his procedure, the confidence intervals should cover the population average about 9,000 times out of 10,000 (or 90%). However, for the Retrofit sample, not a single coverage rate reached 90%. The worst coverage rate was about 60% (measure 121), three additional coverage rates were under 80% (at 77.8% for measure 110, 78.0% for measure 107, and 79.9% for measure 113), 10 rates were between 80% and 85%, and 12 rates were greater than 85% and under 90%.[56] The results in Tables 1 and 2 show that the confidence intervals, whether constructed by Dr. Levy's method or based on a Student's t distribution, can perform far worse than Dr. Levy claims. That is, the confidence intervals are supposed to have only 10% non-coverage rates but for the Retrofit population the non-coverage rate rose as high as 40% and for the Critical Support population it rose to 18%. For both populations, basing the confidence intervals on the Student's t distribution, which is the standard approach, gives a slight improvement on the method that Dr. Levy employed as evidence by the coverage probabilities tending to move closer to 90%. For example, when the confidence intervals were based on the Student's t distribution instead of the normal distribution as Dr. Levy did, their average coverage probability increased from 83.8% to 84.4% for the Retrofit and increased from 88.1% to 88.2% for the Critical Support. However, using the Student's t distribution does not solve the larger problem, which is that for some measures his confidence intervals can have coverage rates that are much poorer than what Dr. Levy claims. This point applies with particular force to his confidence intervals for the measures available only for his sample, because their coverage rates could be much less than the 90% that Dr. Levy claims.

[56] Worksheets "Exhibit I" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

| Measure | Coverage Rate (%) of Confidence Interval as Constructed by Dr. Levy | Coverage Rate (%) of Confidence Interval Based on Student's t Distribution |
|---|---|---|
| 121 | 59.9 | 60.4 |
| 110 | 77.8 | 78.3 |
| 107 | 78.0 | 78.6 |
| 113 | 79.9 | 80.0 |
| 109 | 81.5 | 82.2 |
| 135 | 81.8 | 82.3 |
| 108 | 81.9 | 82.5 |
| 137 | 82.0 | 82.6 |
| 104 | 82.2 | 82.7 |
| 111 | 82.2 | 82.8 |
| 112 | 82.6 | 83.1 |
| 106 | 82.9 | 83.5 |
| 133 | 83.1 | 83.6 |
| 122 | 84.3 | 84.9 |
| 143 | 86.8 | 87.2 |
| 105 | 87.7 | 88.3 |
| 127 | 87.9 | 88.5 |
| 118 | 87.9 | 88.6 |
| 125 | 88.1 | 88.7 |
| 136 | 88.1 | 88.8 |
| 142 | 88.4 | 89.0 |
| 144 | 89.0 | 89.0 |
| 115 | 88.2 | 89.1 |
| 101 | 88.9 | 89.2 |
| 130 | 89.0 | 89.7 |
| 138 | 89.4 | 90.0 |
| average | 83.8 | 84.4 |

Table 1. Confidence intervals' coverage rates for averages for 26 measures in the Retrofit population[57]

[57] Worksheets "Exhibit I", "Exhibit J", Exhibit K" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

| Measure | Coverage Rate (%) of Confidence Interval as Constructed by Dr. Levy | Coverage Rate (%) of Confidence Interval Based on Student's t Distribution |
|---|---|---|
| 135 | 82.2 | 82.3 |
| 137 | 83.4 | 83.5 |
| 104 | 84.8 | 84.9 |
| 110 | 85.3 | 85.4 |
| 133 | 85.9 | 86.1 |
| 112 | 86.6 | 86.7 |
| 113 | 87.3 | 87.4 |
| 111 | 87.6 | 87.7 |
| 107 | 87.9 | 88.0 |
| 109 | 88.0 | 88.1 |
| 115 | 88.0 | 88.1 |
| 108 | 88.0 | 88.1 |
| 106 | 88.0 | 88.2 |
| 143 | 89.3 | 89.4 |
| 142 | 89.2 | 89.4 |
| 105 | 89.2 | 89.4 |
| 101 | 89.4 | 89.5 |
| 136 | 89.8 | 89.9 |
| 122 | 89.8 | 89.9 |
| 138 | 89.9 | 90.0 |
| 144 | 89.9 | 90.1 |
| 130 | 89.9 | 90.1 |
| 121 | 90.2 | 90.3 |
| 118 | 90.3 | 90.4 |
| 125 | 90.4 | 90.5 |
| 127 | 90.4 | 90.6 |
| Average | 88.1 | 88.2 |

Table 2. Confidence intervals' coverage rates for averages for 26 measures in Critical Support population[58]

[58] Worksheets "Exhibit L", "Exhibit M", Exhibit N" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

6.35. To recapitulate, Dr. Levy made a host of errors related to standard errors and confidence intervals. He used the wrong formula for estimating standard errors for his estimator, he offered the wrong interpretation of confidence intervals, he failed to check the validity of his use of the normal approximation to develop his confidence intervals, he used the wrong mathematical distribution to form his confidence intervals. Although he claims that his confidence intervals have 90% coverage rates, or non-coverage rates of 10%, analysis shows that for some measures the non-coverage rates are many times greater than claimed.

6.36. Dr. Levy found that some of his confidence intervals based on the normal approximation gave ranges that extended beyond the range of possibility. It is well established and, Dr. Levy recognizes, that counts cannot be negative and percentages must lie between 0% to 100%. Dr. Levy states in his report that

> "For measure 140 for the Critical Support, of the 238 observations in the sample, . . . [T]he formula for the confidence interval based on the normal approximation would yield a lower bound of less than zero. . . For measure 132, which is the ratio of measure 129 to measure 120, the standard formula for the confidence interval would yield an upper bound of greater than 100%."[59]

Given this problem, Dr. Levy used an analytic method known as the bootstrap to construct alternative confidence intervals. He stated that

> "To address these issues that can occur at the boundaries (i.e. close to 0 in the case of a count variable such as measure 140 or beyond 0 or 1 for a proportion measure such as measure 132), I sample repeatedly from the data to estimate the upper and lower bounds. This method involves

[59] Levy, 40. Levy-R2, page 40 notes in addition that "[f]or measure 124 for Critical Support, the formula for the confidence interval based on the normal approximation yields a lower bound of less than zero."

drawing repeatedly and with replacement from the sample to calculate a statistic of interest in each of the samples. The resulting distribution of values of that statistic allows the estimation of the upper and lower bounds. My methodology involves 10,000 repeated draws from the sample for each of these measures."

Dr. Levy cites the book by Efron and Tibshirani[60] to provide justification of his method.

6.37. Dr. Levy's report fails to mention that the bootstrap methodology he uses requires that he select the 10,000 samples independently. Efron and Tibshirani make this point clearly.

> "The bootstrap algorithm begins by generating a large number of independent bootstrap samples. . ."[61]

> "The bootstrap algorithm works by drawing many independent bootstrap samples. . ."[62]

6.38. Dr. Levy's 10,000 bootstrap samples were not selected independently. It is easy to show this. I directed Dr. Vandaele and his colleagues to rerun the SAS® program Dr. Levy used for his bootstrap analysis but to write out intermediate output so that I could analyze it.[63] Each of his samples consisted of a simple random sample of size 238 selected with replacement from his original sample of 238 Critical Support Updates/Fixes. Table 3, below, shows the first six and the last six selections from the first ten of his bootstrap samples. Each selection should have an equal chance of being

---

[60] Bradley Efron and Robert J. Tibshirani (1993) *An Introduction to the Bootstrap.* New York: Chapman & Hall ("Efron and Tibshirani"), 168-176.

[61] Efron and Tibshirani, 13.

[62] Efron and Tibshirani, 46.

[63] The SAS® program is "repeat_sample.sas" in the folder "SAS Code", which is contained in the folder ORCLX-AACG-000004, which is contained in the folder ORCLX-AACG-003.

any number from 1 through 238, and there should be no patterns observable in the data aside from what one might obtain by chance.

| | Replication | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Selection** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| 1 | 52 | 96 | 140 | 184 | 228 | 34 | 78 | 122 | 166 | 210 |
| 2 | 115 | 108 | 101 | 94 | 87 | 80 | 73 | 65 | 58 | 51 |
| 3 | 147 | 4 | 99 | 194 | 52 | 147 | 4 | 99 | 194 | 51 |
| 4 | 62 | 124 | 186 | 10 | 71 | 133 | 195 | 19 | 80 | 142 |
| 5 | 233 | 214 | 196 | 177 | 158 | 140 | 121 | 102 | 84 | 65 |
| 6 | 234 | 226 | 219 | 212 | 205 | 197 | 190 | 183 | 175 | 168 |
| 233 | 18 | 171 | 86 | 1 | 155 | 70 | 223 | 138 | 53 | 207 |
| 234 | 179 | 69 | 196 | 86 | 213 | 103 | 230 | 120 | 9 | 137 |
| 235 | 192 | 105 | 18 | 169 | 82 | 233 | 146 | 59 | 210 | 123 |
| 236 | 135 | 209 | 44 | 117 | 190 | 25 | 98 | 171 | 6 | 79 |
| 237 | 104 | 156 | 209 | 24 | 76 | 129 | 182 | 234 | 49 | 102 |
| 238 | 14 | 147 | 42 | 174 | 69 | 202 | 96 | 229 | 123 | 18 |

Table 3. The first six and last six items selected by Dr. Levy in his first 10 bootstrap replication samples.[64]

6.39. The first row in Table 3 shows the items that Dr. Levy selected first in each of the 10 replications. In the first replication, the first item selected was number 52, in the second replication the item number was 44 higher, at 96; in the third replication the item number was again 44 higher, at 140; in the fourth replication the item number was again 44 higher, at 184; in the fifth replication the item number was again 44 higher, at 228. In the sixth sample, the item number was 44 higher again, provided that once we reach 228 we start counting all over again. So, 228 + 44 = 272; subtract 238 to get 34.[65] To get the

[64] Worksheet "Exhibit O" in spreadsheet "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

[65] This kind of arithmetic is called arithmetic modulo 238. We are familiar with arithmetic modulo 12, which we use in computing times of day if we do not pay attention to a.m. or p.m. For example, the time that 8 hours later than 10:00 is 6:00, which we calculate by adding 8 and 10 to get 18 and then subtract 12 to get 6. In Dr. Levy's sample, we add 228 plus 44 to get 272, and then we subtract 238 to get 34. Thus, in arithmetic modulo 238, the number 238 is the constant such that once we reach it we start counting over again from 0, just like 12 is the constant for the 12-hour clock.

first selections in replications 7, 8, 9, and 10 we similarly keep adding 44 to the previous item selection number.

6.40. Looking at the second row in Table 3 we see a similar pattern, with the second selection in each of replications 2 through 10 being obtained by subtracting 7 from the item number selected in the previous replication – except that the item number for replication 8 is obtained by subtracting 8 from the selection in replication 7, i.e., 65 = 73 – 8. The difference does not reflect a break in pattern, but represents rounding, as described in the next paragraph.

6.41. Dr. Levy determined his item selections by first generating "random" numbers between 0 and 1 (but not including 1), then multiplying his random numbers by 238, and then dropping any decimal parts. For example, the first random number generated was 0.2211; multiplying by 238 yields 52.6 = 0.2211×238, and dropping the decimal part yields 52. This process yields numbers between 0 and 237; this point is further discussed in paragraph 6.43. The patterns in the underlying "random" numbers are exact. Table 4 shows the random numbers underlying the item numbers in Table 3.

| | Replication | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Selection | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 0.2211 | 0.4060 | 0.5910 | 0.7760 | 0.9609 | 0.1459 | 0.3309 | 0.5158 | 0.7008 | 0.8857 |
| 2 | 0.4862 | 0.4563 | 0.4264 | 0.3965 | 0.3666 | 0.3367 | 0.3068 | 0.2768 | 0.2469 | 0.2170 |
| 3 | 0.6192 | 0.0190 | 0.4189 | 0.8187 | 0.2185 | 0.6183 | 0.0182 | 0.4180 | 0.8178 | 0.2176 |
| 4 | 0.2646 | 0.5240 | 0.7834 | 0.0428 | 0.3022 | 0.5616 | 0.8210 | 0.0804 | 0.3398 | 0.5992 |
| 5 | 0.9804 | 0.9020 | 0.8236 | 0.7452 | 0.6668 | 0.5884 | 0.5100 | 0.4316 | 0.3532 | 0.2748 |
| 6 | 0.9845 | 0.9537 | 0.9230 | 0.8923 | 0.8616 | 0.8308 | 0.8001 | 0.7694 | 0.7387 | 0.7079 |
| 233 | 0.0767 | 0.7204 | 0.3641 | 0.0078 | 0.6515 | 0.2952 | 0.9389 | 0.5826 | 0.2263 | 0.8700 |
| 234 | 0.7562 | 0.2919 | 0.8275 | 0.3632 | 0.8989 | 0.4346 | 0.9702 | 0.5059 | 0.0416 | 0.5772 |
| 235 | 0.8094 | 0.4436 | 0.0778 | 0.7120 | 0.3462 | 0.9804 | 0.6146 | 0.2488 | 0.8830 | 0.5172 |
| 236 | 0.5711 | 0.8782 | 0.1852 | 0.4923 | 0.7994 | 0.1065 | 0.4136 | 0.7207 | 0.0278 | 0.3348 |
| 237 | 0.4377 | 0.6590 | 0.8802 | 0.1015 | 0.3227 | 0.5440 | 0.7653 | 0.9865 | 0.2078 | 0.4290 |
| 238 | 0.0620 | 0.6194 | 0.1767 | 0.7341 | 0.2915 | 0.8489 | 0.4062 | 0.9636 | 0.5210 | 0.0784 |

Table 4. The random numbers underlying the first six and last six items selected by Dr. Levy in his first 10 bootstrap replication samples.[66]

In the first row of Table 4, the random number for each replicate from 2 through 10 is obtained by adding 0.184963 to the number for the previous replicate (allowing for rounding error in the table), but if the sum is 1 or greater then 1 is subtracted. That is, that the random number for replication 5 is 0.9609, and adding 0.184963 to 0.9609 yields 1.1459; this exceeds 1, so 1 is subtracted to yield 0.1459 = 1.1459 – 1.[67] The same process may be applied to every row to yield the "random number" for every replication after the first.

6.42. Thus, Dr. Levy's bootstrap samples are flawed in that they are not independent as required by the bootstrap methodology.

6.43. It appears to me based on the information that has been produced that Dr. Levy's bootstrap sampling was flawed in an additional way as well. The 10,000 samples he generates for his bootstrap analysis should all have the same number of selections

[66] Worksheet "Exhibit P" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010. To enhance readability, the numbers in Table 4 are shown to only 4 decimal places. The first 6 numbers in the first row are, to 6 decimal places, 0.221077, 0.406040, 0.591002, 0.775965, 0.960927, 0.145890.

[67] This is arithmetic modulo 1.

(counting duplicates), specifically 238. However, his samples often have fewer than 238 selections. I asked Dr. Vandaele and his colleagues to tabulate the sample sizes for the first 20 of the 10,000 samples generated by Dr. Levy. The tabulations show that only 7 had 238 selections and the other 13 had fewer than 238 selections, and in fact one of the samples had only 233 selections.[68] [69] The reason for this is clear. As noted in paragraph 6.41, he generated item numbers to include in the bootstrap sample by multiplying his random number between 0 and 1 by 238, yielding numbers between 0 and 237. However, the items in Dr. Levy's original sample are numbered from 1 to 238 not from 0 to 237. It appears based on the information provided that sometimes his procedure would indicate that item 0 should be selected for the bootstrap sample – there is no item 0, however, and in those cases his sample would simply be smaller than 238, with the shortage equal to the number of times item number 0 was designated for selection. I directed Dr. Vandaele and his colleagues to tabulate how many of the 10,000 samples each of the 238 items was selected into. Item 0 was selected into about 63% of the samples, and, logically, because there is no item 0 it follows that 63% of the samples were of size smaller than 238 items (including duplicates).[70]

6.44. Understanding the method by which he generates his 10,000 samples for his bootstrap analysis leads us to see another flaw in his procedure as well. As noted in paragraph 6.43, he generated item numbers to include in bootstrap sample by multiplying his random number between 0 and 1 by 238, yielding numbers between 0 and 237. This suggests that item 238 would not be selected for any of the 10,000 samples. As one

[68] Worksheet "Exhibit Q" in spreadsheet "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

[69] It is my understanding that Plaintiffs produced additional log files from Dr. Levy, bates labeled ORCLX-AACG-000008 through ORCLX-AACG-000012, after the close of business central standard time on March 2, 2010. As noted in Section 3.3, due to my pre-planned travel out of the country, I have not been able to receive or review these log files. I was provided an excerpt of the repeat_sample.log file from ORCLX-AACG000011 created by Dr. Vandaele and his colleagues which is located in Appendix 6 and Bates-labeled SAP-SPE-000011 through SAP-SPE-000031. My review of this information suggests that the log file data further supports the position stated above.

[70] Worksheet "Exhibit R" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

would expect, not one of the 10,000 samples generated by Dr. Levy includes item number 238.[71]

6.45. In summary, Dr. Levy's method for selecting bootstrap samples appears to be seriously flawed in three ways. First, each bootstrap sample was supposed to be a simple random sample with replacement from his original sample of 238 items, but as discussed in paragraph 6.43 one item (item 238) had no chance of being selected. Second, each bootstrap sample was supposed to have 238 items, but as discussed in paragraph 6.43 some of the samples were smaller than that. Third, his bootstrap samples were supposed to be selected independently of each other, but as discussed in paragraphs 6.36 through 6.42, Dr. Levy's bootstrap samples were not selected independently, as required by the bootstrap methodology.

6.46. These flaws could be detected because Plaintiffs provided the computer programs used to select the samples, and the materials provided were sufficient for Defendants to at least replicate the sampling procedures and understand how they were carried out.

6.47. Returning to the matter of the generation of confidence intervals using the bootstrap, the bootstrap is not a panacea. For example, the text by Efron and Tibshirani on which Dr. Levy relies for his bootstrap methodology[72] contains an empirical example where the bootstrap method was tested against a known population. In the example, the bootstrap method favored by Dr. Levy was used to generate a 95% confidence interval, which implies that only 5% of the confidence intervals should fail to include the population value. However, the failure rate for the intervals was twice that – 10% of the intervals did not cover the population value.[73]

---

[71] Worksheet "Exhibit R" in "All Exhibits.xls", located in Appendix 5 and Bates-labeled SAP-SPE-000010.

[72] Levy, 40, footnote 47.

[73] Efron and Tibshirani, 175, Table 13.3.

6.48. To recapitulate, Dr. Levy attempted to use a method known as the bootstrap to improve his confidence intervals. As Dr. Levy intended to use it, the method involves taking the original sample as if it were the population, sampling repeatedly and independently from it according to the original sample design to generate a large number of estimates, and then constructing an interval from the observed distribution of the estimates. Dr. Levy did provide documentation of the process by which he selected those samples, and a number of errors are apparent, including failure to make his samples independent of each other, failure to generate constant size samples, and failure to see that the samples were selected according to his original sample design.

6.49. Even with all the flaws in his constructions of the confidence intervals, the levels of sampling error that he reports can be quite large. Dr. Levy's own estimates of the level of sampling error are quite high for some measures. He reports 90% confidence intervals for estimates that he calculates from his sample and he defines the "precision range" as the difference between the upper end of his interval minus the lower end of his interval, expressed as a percentage of his estimate.[74] A small precision range is desirable, and a large precision range is not desirable. For a number of measures, as shown in Table 5, the precision range as calculated from the data in his report is about 100% or more.

---

[74] Levy-R2, 12 defines precision range as follows: "In order to design the sample size, she decides to use a 90% confidence level and a 20% precision range; that is, she wants to be able to say that if she sampled from this population repeatedly, 90% of the time the true number of computers would be within plus or minus 10% of her result."

| | Dr. Levy's Estimates | | | | | | Source in Levy-R2 | |
|---|---|---|---|---|---|---|---|---|
| Measure | Estimate | Lower Bound | Upper Bound | Lower Bound (% of Estimate) | Upper Bound (% of Estimate) | Precision Range | Table | Page Number |
| | Retrofit Population | | | | | | | |
| 114 | 1631 | 699 | 2564 | 43% | 157% | 114% | 10A | 28 |
| 119 | 604 | 321 | 887 | 53% | 147% | 94% | 12A | 31 |
| 139 | 903 | 368 | 1439 | 41% | 159% | 119% | 9A | 26 |
| 140 | 396 | 203 | 590 | 51% | 149% | 98% | 9A | 26 |
| 141 | 1300 | 586 | 2013 | 45% | 155% | 110% | 9A | 26 |
| | Critical Support Population | | | | | | | |
| Measure | | | | | | | | |
| 114 | 65 | 23 | 107 | 35% | 165% | 129% | 10B | 29 |
| 119 | 16 | 3 | 30 | 19% | 188% | 169% | 12B | 31 |
| 124 | 319 | 49 | 683 | 15% | 214% | 199% | 13B | 32 |
| 139 | 37 | 11 | 63 | 30% | 170% | 141% | 9B | 26 |
| 140 | 8 | 0 | 25 | 0% | 313% | 313% | 9B | 26 |
| 141 | 45 | 16 | 74 | 36% | 164% | 129% | 9B | 26 |

Table 5. Precision range and other measures of sampling error for selected measures in Dr. Levy's analysis.[75]

## 7. Dr. Levy's Omissions

7.1. Dr. Levy's work is marred by a number of omissions. One omission, which limits his ability to make generalizations from his analysis, is that he did not attempt to design a sample from Updates or Fixes other than PeopleSoft HRMS payroll tax and regulatory Updates or Fixes. This reflects an apparently conscious choice that puts limits on the inferences that can be drawn from his analysis – no statistical inferences about Updates or Fixes other than PeopleSoft HRMS payroll tax and regulatory Updates or Fixes can be supported by his analysis.

[75] In Table 5, the estimates, lower bound, and upper bound were taken from Dr. Levy's report in the locations shown at the right of the table. To calculate the lower bound and upper bound as a percentage of the estimate, I divided the lower and upper bounds reported by Dr. Levy by the estimate reported by Dr. Levy and I expressed the ratio as a percentage. Consistent with Dr. Levy's specifications, the precision ranges are equal to the upper bound minus the lower bound expressed as percentages.

7.2. I see nothing in Dr. Levy's report showing that he validated or checked the quality of the data that Mr. Mandia provided to him for use in his work. Dr. Levy relied completely on Mr. Mandia's data in all phases of his work. Any flaws in the measures given by Mr. Mandia to Dr. Levy will transmit through Dr. Levy's analysis and can affect his results.

7.3. Dr. Levy apparently did not consider better sample designs than simple random sampling, let alone simple random sampling with replacement. For example, because Dr. Levy had information for a large number (26) of measures for every item in the population, he could have stratified items according to those measures and selected a stratified sample. Cochran says quite clearly that "[i]f intelligently used, stratification nearly always results in a smaller variance for the estimated mean or total than is given by a comparable simple random sample."[76] It is hard to believe that Dr. Levy would not be aware of the possibility that one can reduce the standard error one achieves from a given size sample by choosing a better design, but he does attempt to contradict Cochran when he makes the incorrect claim that

> "The statistician has no influence over the standard error found in a sample of a given size. The standard error of the sample is simply a characteristic of the sample she has drawn that reflects the variability in the sample of a given size."[77]

7.4. A most significant omission by Dr. Levy was his failure to keep track of exactly how he selected his samples of Critical Support and Retrofit Updates/Fixes. Plaintiffs provided the outcomes of his sample selection – the item numbers designated for inclusion in the sample – but they did not provide documentation of the process by which the samples were selected. The information does not allow us to duplicate the way the

[76] Cochran, p. 99. The variance is the square of the standard error, so reducing one is equivalent to reducing the other.

[77] Levy 13, note 32.

item numbers were chosen for the sample. There is a substantial variety of flaws in Dr. Levy's selection of samples for his bootstrap analysis. In addition, his Critical Support sample is far from what would tend to arise from sampling from the Critical Support population if the sample was in fact selected with simple random sampling as he claims it was. Together, these issues, coupled with the fact that Dr. Levy's report and supporting materials do not adequately document that his sample was actually selected with simple random sampling as he claims, raise significant questions about the validity of Dr. Levy's sample selection procedures for his samples of Critical Support and Retrofit Updates/Fixes.

## 8. Dr. Levy's Opinions

8.1. Dr. Levy does not always stick to reporting the facts of his analysis, but he also offers opinions based on his analysis, particularly, but not exclusively, in his Executive Summary.[78] For example, he writes

> "Mr. Mandia also gathered information to understand Contamination[22] of Objects and Environments. One measure to capture this information, which was counted for the entire population of Fixes for Retrofit and for Critical Support, is the percentage of instances in which customers received a First Deliverable that was contaminated based on Mr. Mandia's analysis of Objects. For Retrofit Fixes, 89.75% of the time that customers received a First Deliverable, that deliverable was contaminated based on Mr. Mandia's analysis of Objects found in Delivered Updates and Fixes. This same measure for Critical Support Fixes shows that 93.72% of the time, customers received a First Deliverable that was contaminated based on Mr. Mandia's analysis of Objects found in Delivered Updates and fixes."[79]

He also writes

[78] Levy-R2, 1-6.

[79] Levy-R2, 5. His footnote 22 reports his understanding of "Contamination".

> "Measures 128, 129, 131 and 132 are reported with a 90% confidence interval [in the tables]. Measure 131 shows that in instances in which customers received a first deliverable Retrofit Fix, 83.92% of the First Deliverables were contaminated based on Object analysis. This same measure for the Critical Support Fix population is 99.12%."[80]

These are misleading statements about percentages. Dr. Levy offers them as unconditional truth, but in fact he is ignoring sampling error and he is ignoring any potential for measurement error by Mandiant. The fact that his percentages are based on a sample, which is subject to sampling error, implies that he is almost surely incorrect to some extent in his claims. That is, the percentages he estimates to be 89.75% and 93.72% are not exactly as he estimates – there will be some error in his estimates.

8.2. In other places Dr. Levy is more conscientious about saying that his numbers are estimates based on a sample. But, he still does not make allowance for sampling error. In many cases he offers the confidence interval in a footnote, and although the width of Dr. Levy's confidence interval represents an allowance for sampling error, the widths of those intervals do not affect what he reports as his definitive estimates or opinions. Thus, Dr. Levy's estimates and opinions do not take sampling error into account. His estimates and opinions are for the most part the midpoints of his confidence intervals.[81]

8.3. The decision to sample and how to sample was made unilaterally by Plaintiffs. I am informed by Defendants' counsel that Plaintiffs have the ultimate burden of proof on their claims in this case. Because the burden of proof is on the Plaintiffs, and Plaintiffs decided to introduce sampling error, Dr. Levy should construct his estimates so that Defendants are not penalized by Plaintiffs' decision to introduce sampling error. As

---

[80] Levy-R2, 33.

[81] For most measures his estimates equal the midpoints of his confidence intervals. The exceptions are the measures whose confidence intervals he constructs using bootstrap methods.

noted above, although the width of Dr. Levy's confidence intervals represents an allowance for sampling error, Dr. Levy's estimates do not take sampling error into account. Dr. Levy's approach shifts some of the burden from sampling error onto Defendants because, other things being equal, the smaller the sample size the greater the chance for a large error unfavorable to the Defendants.

## 9. List of Appendixes

1. List of reports and supporting materials reviewed
2. Curriculum vitae
3. Correspondence related to documentation of Dr. Levy's sampling process, which has Bates numbers SAP-SPE-000001 through SAP-SPE-000006
4. Written instructions to LECG, which has Bates number SAP-SPE-000007 through SAP-SPE-000009
5. LECG Exhibits, which has Bates number SAP-SPE-000010
6. Excerpt from repeat_sample.log file from LECG, which has Bates number SAP-SPE-000011 through SAP-SPE-000031
7. Some formulas used in Dr. Levy's SAS® program, as transcribed by LECG, which has Bates numbers SAP-SPE-000032 through SAP-SPE-000038

# Appendix 1

| Description | Bates Nos. |
|---|---|
| **Expert Report of Daniel S. Levy (AACG) & Supporting Materials** | |
| · AACG Report Final.pdf | None |
| · 2010-02-05 Corr. Levy Report.pdf<br>· 2010-02-05 Corr. Levy Report Errata.pdf | None |
| · AACG Report 2010-02-12.pdf | None |
| · AACG Report REDLINE 2010-02-12.pdf | None |
| · AACG Report Errata 2010-02-12.pdf | None |
| · 2010-02-12 Polito letter to McDonell re expert reports and production.pdf | None |
| · ORCLX-AACG001 | ORCL-AACG-000001-<br>ORCL-AACG000006 |
| ·ORCLX-AACG002 | ORCLX-AACG-000001-<br>ORCLX-AACG-000002 |
| · AACG001_Index.xls | None |
| · ORCLX-MAN016 - contains corrected ORCLX-MAN-000205 | ORCLX-MAN-000205 |
| · ORCLX-AACG003 | ORCLX-AACG-000001-<br>ORCLX-AACG-000002<br>ORCLX-AACG-000004 |
| · ORCLX-AACG004 | ORCLX-AACG-000007 |
| | |
| **Expert Report of Kevin Mandia (Mandiant) & Supporting Materials** | |
| · Mandiant Report.pdf<br>· Mandiant Report Appendices.pdf | None |
| · Mandiant_Report_Corrected_CLEAN_01-22-10.pdf | None |
| · Mandiant_Report_Corrected_REDLINE_01-22-10.pdf | |
| · Mandiant_Appendices_Corrected_CLEAN_01-22-10.pdf | |
| · Mandiant_Appendices_Corrected_REDLINE_01-22-10.pdf | |
| · Mandiant_Errata_01-22-10.pdf | |
| · Mandia Supplemental Report 2010-02-12.pdf | None |
| · Mandian Supplemental Report REDLINE 2010-02-12.pdf | None |
| · Mandian Supplemental Appendices 2010-02-12.pdf | None |
| · Mandian Supplemenal Appendices REDLINE 2010-02-12.pdf | None |
| · ORCLX-MAN001 | ORCLX-MAN-000001-139, 141<br>143, 145-150, 152-154, 158-<br>159, 161, 162-216 |
| · ORCLX-MAN002 | ORCLX-MAN-000218-299 |
| · ORCLX-MAN003 | ORCLX-MAN-000300-309 |
| · ORCLX-MAN004 | ORCLX-MAN-000134 |
| · ORCLX-MAN005 | ORCLX-MAN-000140<br>ORCLX-MAN-000144<br>ORCLX-MAN-000151<br>ORCLX-MAN-000160 |
| · ORCLX-MAN006 | ORCLX-MAN-000310-316 |
| · ORCLX-MAN007 | ORCLX-MAN-000317 |
| · ORCLX-MAN008 | ORCLX-MAN-000318-319 |
| · ORCLX-MAN009 Cover.pdf | |
| · ORCLX-MAN010 | ORCLX-MAN-000322-323 |
| · ORCLX-MAN011 | ORCLX-MAN-000324-326 |
| · ORCLX-MAN012 | ORCLX-MAN-000327-329 |
| · ORCLX-MAN013 = Mandiant_Index.xls | None |
| · ORCLX-MAN014 | ORCLX-MAN-000330-341 |

| Description | Bates Nos. |
|---|---|
| · ORCLX-MAN015 | ORCLX-MAN-000103-104, 107, 109, 116, 137, 142, 208, 220, 222, 224-226, 229, 231, 236, 239, 242, 246, 253-254, 259, 264-265, 267-268, 271, 274-275, 293 |
| · ORCLX-MAN016 | ORCLX-MAN-000205 |
| · ORCLX-MAN-000217 [image only].pdf | |
| · ORCLX-MAN017 | ORCLX-MAN-000342-382 |
| | |
| **Expert Report of Paul Meyer (Navigant) & Supporting Materials** | |
| (1) PKM Expert Report_111709.pdf;<br>(2) 20091116 POS.pdf;<br>(3) Expert Report of Paul Meyer - Schedules (all);<br>(4) Expert Report of Paul Meyer - Schedules2 (all) | None |
| (1) REVISED Oracle Report_120209.pdf;<br>(2) PKM Errata Letter 12.4.09.PDF;<br>(3) Proof_Updated Expert Report.pdf;<br>(4) Updated Schedules_2009-12-04 (all) | None |
| (1) Attachment 1 - Meyer CV.pdf;<br>(2) Attachment 2 - Meyer 10 Year Testimony List.pdf;<br>(3) Attachment 3 - Information Considered.pdf | |
| Schedules to Report - ALL including updates & natives | None |
| (1) ORCLX-NAV001<br>(2) ORCLX-NAV002<br>(3) ORCLX-NAV003_2009-12-04 | ORCLX-NAV-000001-<br>ORCLX-NAV-000069 |
| 2009-12-147 Jindal ltr fwd Meyer Report Attachs. 1-3 and natives.PDF | None |
| Navigant_Index.xls | None |
| SUPPLEMENTAL Oracle Report 022310 Highlighted.pdf | None |
| | |
| **Expert Report of Paul C. Pinto** | |
| · Paul C Pinto Report.pdf | |
| · Appendix A to Expert Report of Paul Pinto.pdf;<br>· Appendix B to Expert Report of Paul Pinto.pdf;<br>· Appendix C to Expert Report of Paul Pinto.pdf | |
| ORCLX-PIN01 | ORCLX-PIN-000001-<br>ORCLX-PIN-000085 |
| ORCLX-PIN01.xls | None |
| | |
| **Expert Report of FTWorks & Supporting Materials** | |
| · FTWorks_Expert_Report_11_16_09.pdf | None |
| · ORCLX-FT001 | ORCLX-FT-000001-<br>ORCLX-FT-000081 |
| FT_Works_Index.xls | None |
| | |
| **Expert Report of Douglas Lichtman & Supporting Materials** | |
| · Lichtman Report.pdf | None |
| (1) Lichtman Ex 1.pdf<br>(2) Lichtman Ex 2.pdf<br>(3) Lichtman Ex 3.pdf | None |
| (1) ORCLX-DL001<br>(2)ORCLX-DL002_2009-12-04 | ORCLX-DL-000001-<br>ORCLX-DL-000002 |
| ORCLX-DL001-Index.xls | None |
| ORCLX-DL002-Index.xls | None |

# Appendix 2

CURRICULUM VITAE

# BRUCE DAVID SPENCER


Northwestern University

Department of Statistics
Northwestern University
2006 Sheridan Road
Evanston, Illinois 60208-4070
(847) 491-5810
(847) 491-4939 *fax*
bspencer@northwestern.edu

Spencer Statistics, Inc.

Spencer Statistics, Inc.
1404 Asbury Avenue
Evanston, IL 60201-4109
bruce.spencer@sbcglobal.net


Citizenship: United States

## EDUCATION

| | |
|---|---|
| Ph.D. | Yale University, 1979, Statistics |
| M.S. | Florida State University, 1974, Statistics |
| B.S. | Cornell University, 1973, Biometry |

## PROFESSIONAL EXPERIENCE

1980 - present Northwestern University. Professor (1992 - present), Associate Professor (1986-92), Assistant Professor (1980-86)

- Department of Statistics (member 1987 - present; chair 1988 - 1999, 2000 - 2001, 2007 - present)
- School of Education and Social Policy (member 1980 - 1989, by courtesy 1989 - present)
- Institute for Policy Research (Faculty Fellow, 1984 - 1985, 2000 - present)
- Northwestern Institute for Complex Systems (Member, 2004 - present)

1987 - present Spencer Statistics, Inc. President

1982 - 1994 National Opinion Research Center (NORC), University of Chicago. Senior Research Statistician (1992-1994), Director of Methodology Research Center (1985-1992), Sampling Statistician (1982-1985)

1984 Visiting Assistant Professor, Department of Statistics and Center for Economic Policy Research, Stanford University

1978 - 1980 Senior Staff Officer and Study Director, Committee on National Statistics, National Academy of Sciences

## PROFESSIONAL ACTIVITIES

### Selected Advisory Committees and Boards

National Academy of Sciences Panel to Review the Programs of the Bureau of Justice Statistics (2006-2009)
National Institute of Statistical Sciences, Expert Task for Reporting of Effect Sizes from NCES Surveys (2006-2007)
National Academy of Sciences Panel on Bureau of Transportation Statistics International Trade Traffic Study (2003-2004)
National Academy of Sciences Panel on Formula Allocations (2001-02)
U.S. Bureau of Census, Census Committee on Statistical Policy and Sampling and Estimation Steering Committee (1996-1999)
U.S. Steering Committee, Third International Mathematics and Science Study (1991-)
National Academy of Sciences Mathematical Sciences Assessment Panel (1991-3)
American Statistical Association Committee on Energy Statistics (1990-2)
American Statistical Association JES Management Committee (1989-95)
Institute of Mathematical Statistics Outreach Committee (1989-1991)
American Statistical Association Section on Survey Research, Executive Committee (Publications Officer) (1989-91)
American Statistical Association Journal of Educational Statistics Management Committee (1989-91)
National Research Council Panel on Statistical Issues in AIDS Research (1988-89)
National Research Council/Institute of Medicine Ad Hoc Advisory Panel on Substance Abuse (1988)
American Statistical Association Technical Panel on the Census Undercount (1982-83)
Bureau of Census, Steering Committee for Conference on Census Undercount (1979-80)

### Awards

Senior Fellow, National Institute of Statistical Sciences (summer 1997)
Fellow, American Statistical Association (1992)
AT&T Research Fellowship (1992-93)
Journal of the American Statistical Association Invited Applications Paper (1991; see article A38, below)
Palmer O. Johnson Memorial Award from the American Educational Research Association (1983)

### Memberships in Professional Organizations

American Statistical Association
Institute of Mathematical Statistics
Population Association of America
Royal Statistical Society

## PUBLICATIONS

### Books

**B1** *Statistical Demography and Forecasting.* New York: Springer, 2005 (with Juha M. Alho).

**B2** *Statistics and Public Policy.* (editor) Oxford: Oxford University Press, 1997.

**B3** *Benefit-Cost Analysis of Data Used to Allocate Funds.* New York: Springer, 1980.

**B4** *Estimating Population and Income of Small Areas,* Panel on Small-Area Estimates of Population and Income, Committee on National Statistics, National Research Council, Washington, D.C.: National Academy Press, 1980.

This book includes the Panel's report and nine papers with attributed authorship. As Study Director for the Panel, I was largely responsible for drafting the report. I also wrote seven papers in the book, which are listed below (with page references).

"Models for error in postcensal population estimates." (pp. 217-228)

"A note on the use of postcensal population estimates in employment and unemployment measures." (pp. 215-216)

"Postcensal per capita income estimation methods of the Census Bureau: Summary." (pp. 188-193)

"Revenue sharing allocations and the effects of data errors." (pp. 205-214)

"Effects of biases in census estimates on evaluation of postcensal estimates." (pp. 232-236)

"Postcensal population estimation methods of the U.S. Bureau of the Census." (pp. 131-187)

"The role of judgement in postcensal estimation." (pp. 194-199)

### Articles and Reports (* denotes refereed journal)

**A1 *** "Sufficiency, Minimal Sufficiency, and the Lack Thereof" *The American Statistician* **30** 34-35, 1976 (with Allan R. Sampson).

**A2** "Implications of Equity and Accuracy for Undercount Adjustment: A Decision-

Theoretic Approach." Pp. 204-216 in U.S. Bureau of the Census. *Conference on Census Undercount: Proceedings of the 1980 Conference*. Washington, D.C.: U.S. Department of Commerce, 1980.

**A3** "Issues of Accuracy and Equity in Adjusting for Census Undercount" *Proceedings of the American Statistical Association, Social Statistics Section,* Washington, D.C., 1981.

**A4** "On Estimating Population and Income for Local Areas" *Statistical Reporter* **81** 377-381, 1981 (with Evelyn M. Kitagawa).

**A5 *** "The Sense and Nonsense of School Effectiveness" *Journal of Policy Analysis and Management* **1** 43-52, 1981 (with David E. Wiley).

**A6** "Small-Area Analysis." Pp. 607-614 in J.A. Ross, editor, *International Encyclopedia of Population,* New York: The Free Press, 1982 (with Robert J. Lapham).

**A7 *** "A Note on Statistical Defensibility" *The American Statistician* **36** 208-209 (with comments 209-216), 1982.

**A8 *** "Feasibility of Benefit-Cost Analysis of Data Programs" *Evaluation Review* **6** 649-672, 1982.

**A9 *** "Concerning Dubious Estimates of the Effects of Census Undercount Adjustment of Federal Aid to Cities" *Urban Affairs Quarterly* **18** 145-148, 1982.

**A10 *** "Technical Issues in Allocation Formula Design" *Public Administration Review* **42** 524-529, 1982.

**A11** "Feasibility of Benefit-Cost Analysis of Public Data" *Proceedings of the American Statistical Association, Social Statistics Section,* Washington, D.C., 1982.

**A12** "Distribution of Federal Benefits According to Statistical Formulas: Discussion of Paper" *Proceedings of the American Statistical Association, Social Statistics Section,* Washington, D.C., 1982.

**A13 *** "On Interpreting Test Scores as Social Indicators: Statistical Considerations" *Journal of Educational Measurement* **20** 317-334, 1983.

**A14 *** "Test Scores as Social Statistics: Comparing Distributions" *Journal of Educational Statistics* **8** 249-270, 1983.

**A15** "Toward Conducting Benefit-Cost Analyses of Data Programs" *Proceedings of the American Statistical Association, Social Statistics Section* pp. 46-51, 1984.

**A16** "Simplifying Complex Samples with the Bootstrap" *Proceedings of the American Statistical Association, Survey Research Section,* pp. 484-488, 1984.

**A17** * "Report of the ASA Technical Panel on the Census Undercount". *The American Statistician* **38** 252 - 256, 1984 (as Panel member).

**A18** * "Uncertain Population Forecasting" *Journal of the American Statistical Association* **80** 306-314, 1985 (with Juha M. Alho).

**A19** * "Avoiding Bias in Estimates of the Effect of Data Error on Allocations of Public Funds" *Evaluation Review* **9** 511-518, 1985.

**A20** * "Optimal Data Quality" *Journal of the American Statistical Association* **80** 564-573, 1985.

**A21** * "Statistical Aspects of Equitable Apportionment" *Journal of the American Statistical Association* **80** 815-822, 1985.

**A22** "Test Score Decline: What and How Well Does It Measure?" *Proceedings of the American Statistical Association, Social Statistics Section,* pp. 57-64, 1985.

**A23** "Conceptual Issues in Measuring Improvement in Population Estimates." pp. 393-407 in *Second Annual Research Conference.* Washington, D.C.: U.S. Bureau of the Census, 1986.

**A24** "Toward Conducting Benefit-Cost Analyses of Data Programs." Pp. 38-59 in R. W. Pearson and R. F. Boruch (eds.) *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits.* New York: Springer-Verlag, 1986.

**A25** "Sampling Problems in Merging a Cross Sectional and a Longitudinal Program". Pp. 117-140 in George H. Brown and Elizabeth M. Faupel (editors) *The National Assessment of Educational Progress and the Longitudinal Studies Program; Together or Apart?* Report of a Planning Conference, December 11, 1986. Washington, D.C.: Center for Education Statistics.

**A26** * "Comment" on "Census Undercount Adjustment and the Quality of Geographic Population Distributions" by Allen L. Schirm and Samuel H. Preston, *Journal of the American Statistical Association* **82** 984 - 986, 1987.

**A27** "Total Error of Dual-System Estimates of Population Size", in *Fourth Annual Research Conference* Washington, D.C.: U.S. Bureau of the Census, 1988 (with Mary H. Mulry).

**A28** * "Total Error in the Dual System Estimator: The 1986 Census of Central Los Angeles County", *Survey Methodology* **14** 241 - 263, 1988 (with Mary H. Mulry).

**A29** "Assessing Total Error in Dual System Estimates" *Proceedings of the American Statistical Association, Survey Research Section,* pp. 535-540, 1988 (with Mary H. Mulry).

**A30** Aaron, Joan, Billard, Lynn, Brookmeyer, Ron, Menken, Jane, Moses, Lincoln, Spencer, Bruce, Stoto, Michael, and Turner, Charles "Monitoring the Epidemic", chapter 1 (Pp. 31- 72) of: Turner, Charles, Miller, Heather, and Moses, L. (Eds.) *AIDS, Sexual Behavior, and Intravenous Drug Use: Report of the Committee on AIDS Research and the Behavioral, Social, and Statistical Sciences* National Academy of Sciences. Washington, D.C.: National Academy Press, 1989.

**A31** "On the Accuracy of Current Estimates of the Numbers of Intravenous Drug Users". Pp. 429 - 446 of: Turner, Charles, Miller, Heather, and Moses, L. (Eds.) *AIDS, Sexual Behavior, and Intravenous Drug Use: Report of the Committee on AIDS Research and the Behavioral, Social, and Statistical Sciences. National Academy of Sciences* Washington, D.C.: National Academy Press, 1989.

**A32** * "Effects of Targets and Aggregation on the Propagation of Error in Mortality Forecasts". *Journal of Mathematical Population Studies* **2** 209-227, 1990 (with Juha M. Alho).

**A33** * "Error Models for Official Mortality Forecasts" *Journal of the American Statistical Association* **85** 609-616, 1990 (with Juha M. Alho).

**A34** * "Needed Data Expenditure for an Ambiguous Decision Problem" *Journal of the American Statistical Association* **85** 1099-1104, 1990 (with Lincoln E. Moses).

**A35** "Total Error in Post Enumeration Survey (PES) Estimates of Population: The Dress Rehearsal Census of 1988." Pp. 326-361 (with discussion 362-366) in *1990 Annual Research Conference* Washington, D.C.: U.S. Bureau of the Census, 1990 (with Mary H. Mulry).

**A36** Spencer, Bruce, Frankel, Martin, Ingels, Steven, Rasinski, Kenneth, and Tourangeau, Roger, *Base Year Sample Design Report, National Education Longitudinal Study of 1988.* Washington, D.C.: National Center for Education Statistics, 1990.

**A37** * "Sampling Probabilities for Aggregations, with Applications to NELS:88 and Other Educational Longitudinal Surveys." *Journal of Educational Statistics* **16** 21-34, 1991 (with Wendene Foran).

**A38** * "Total Error in PES Estimates of Population: The Dress Rehearsal Census of 1988". *Journal of the American Statistical Association, 86:* 839-854, 1991 (with discussion 855-863). (with Mary Mulry)

**A39** "A Critique of Sampling in the 1990 Trial State Assessment". Pp. 1-18 in *Assessing Student Achievement in the States: Background Studies.* Studies for the Evaluation of the NAEP Trial State Assessment Commissioned for the National Academy of Education Panel Report on the 1990 Trial. Stanford: National Academy of Education, 1991.

**A40** "Eligibility/Exclusion Issues in the 1990 Trial State Assessment". Pp. 19-49 in *Assessing Student Achievement in the States: Background Studies.* Studies for the Evaluation of the NAEP Trial State Assessment Commissioned for the National Academy of Education Panel Report on the 1990 Trial. Stanford: National Academy of Education, 1991.

**A41** * "Population Forecasts as a Database" *Journal of Official Statistics* 7 295-310, 1991 (joint with Juha Alho).

**A42** "Shrinkage Weights for Unequal Probability Samples", *Proceedings of the American Statistical Association, Survey Research Section,* 1991 (with Theodora Cohen).

**A43** "Accuracy of 1990 Census Undercount Estimates for the Postcensal Estimates", *Proceedings of the American Statistical Association, Survey Research Section,* (with Mary Mulry).

**A44** * "On Comparing Distributions of Poverty Gaps" *Sankhyā, Series B,* **54** 114-126, 1992 (with Stephen Fisher).

**A45** "Accuracy of Undercount Estimates for the 1990 Census", *1992 Annual Research Conference Proceedings*, U.S. Bureau of the Census, pp. 3 - 17 (with Mary Mulry).

**A46** "A Study of Eligibility Exclusions and Sampling: 1992 Trial State Assessment". Pp. 1-68 in *The Trial State Assesssment: Propsects and Realities: Background Studies.* Studies of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1992 Trial State Assessment. Stanford: National Academy of Education, 1992.

**A47** "Evaluating Accuracy of Population Estimates for Small Areas", Pp. 27-36 in *Small Area Statistics and Survey Designs* Volume II: Contributed papers and panel discussion. Warsaw: Central Statistical Office, 1993 (with Mary Mulry).

**A48** "Discussion of Germain-Julien and Fay-Thompson", *1993 Annual Research Conference Proceedings*, U.S. Bureau of the Census, pp. 92 - 95.

**A49** * "Accuracy of the 1990 Census and Undercount Adjustments", *Journal of the American Statistical Association* **88** 1080-1091, 1993 (with Mary Mulry).

**A50** * "Pancarcinoma T/Tn Antigen Detects Human Carcinoma Long Before Biopsydoes and Its Vaccine Prevents Breast-Carcinoma Recurrence. *Annals of The New York Academy of Sciences* **690** 355-357, 1993 (with Springer, G.F., Desai, P.R., Tegtmeyer, H., Scanlon, E.F.)

**A51** * "Sensitivity of Benefit-Cost Analysis of Data Programs to Monotone Misspecification". *Journal of Statistical Planning and Inference* **39** 19-31, 1994.

**A52** "Optimally Weighted Means in Stratified Sampling". *Proceedings of the American Statistical Association, Survey Research Section*, 863-866, 1994 (with Jiahe Qian).

**A53** "Evaluating Accuracy of Population Statistics for Small Areas", *Statistics in Transition*, December 1994 (with Mary Mulry).

**A54** * "T/Tn Antigen Vaccine Is Sage and Highly Effective in Preventing Recurrence of Advanced Breast Carcinoma". *International Journal of Oncology* **5** 372-373, #C-1 (with Springer, G. F., Desai, P. R., and Tegtmeyer, H.)

**A55** "Adjusting for Noncoverage of Nontelephone Househods in the National Immunization Survey". *Proceedings of the American Statistical Association, Survey Research Section*, 678-689, 1995 (with M.P. Battaglia, D.J. Malec, D.C. Hoaglin, and J. Sedransk).

**A56** * "T/Tn antigen vaccine is effective and safe in preventing recurrence of advanced breast carcinoma". *Cancer Detection and Prevention* **19** 374-380 (with Springer, G. F., Desai, P. R., Tegtmeyer, H., Carlstedt, S. C. and Scanlon, E. F.).

**A57** * "The Practical Specification of the Expected Error of Population Forecasts", *Journal of Official Statistics* **13** 203-226, 1997 (with Juha Alho).

**A58** * "A Conversation with I. Richard Savage", *Statistical Science* **14** 126-148, 1999 (with Sampson, Allan. R. and Savage, I. Richard)

**A59** * "Sampling-Based Adjustment of the 2000 Census – A Balanced Perspective", *Jurimetrics* **40**, 341-356, 2000 (with Anderson, M., Daponte, B. O., Fienberg, S. E., Kadane, J. B., and Steffey, D. L.).

**A60** * "An Approximate Design Effect for Unequal Weighting When Measurements May Correlate with Selection Probabilities", *Survey Methodology* **26** 137-138, 2000.

**A61** *Accuracy and Coverage Evaluation: Overview of Total Error Modeling and Loss Function Analysis.* DSSD Census 2000 Procedures and Operations Memorandum Series B-19. Washington, D.C.: U.S. Census Bureau, 2001 (with M. H. Mulry)

**A62** "Loss Function Analysis for A.C.E. Revision II Estimates of Census 2000 Coverage Error", *Proceedings of the American Statistical Association, Survey Research Section,* 2966-2971, 2003 (with M. H. Mulry and R. S. ZuWallack)

**A63** *Statistical Issues in Allocation Funds by Formula.* Panel on Formula Allocations, Thomas A. Louis, Thomas B. Jabine, and Marisa Gerstein, eds. Washington, D.C.: National Academy Press, 2003. (as Panel member)

**A64** *Measuring International Trade on U.S. Highways.* Panel on Bureau of Transportation Statistics International Trade Traffic, National Research Council, 2005. (as Panel member)

**A65** "I. Richard Savage", *International Encyclopedia of Statistical Sciences*, 2nd ed. Wiley: New York, 2005, vol. 11, 7440-7443 (with Myles Hollander, Allan Sampson, Jayaram Sethuraman).

**A66** "Investigation of Extreme Estimates of Census Coverage Error for Small Areas", *Proceedings of the Survey Research Methods Section*, Washington, D.C.: American Statistical Association, 2005, 3414-3421 (with M. H. Mulry, E. Schindler, T. Mule, N. Nguyen).

**A67** * "Estimating the Accuracy of Jury Verdicts", *Journal of Empirical Legal Studies* **4**, 305-329, 2007.

**A68** "Verdict on Verdicts: Statistical Measurement of the Accuracy of Jury Verdicts", Pp. 2089-2099 in *Proceedings of the American Statistical Association,* [CD-ROM], Alexandria, VA: American Statistical Association, 2007.

**A69** *Surveying Victims: Options for Conducting the National Crime Victimization Survey,* Robert M. Groves and Daniel L. Cork, Editors, Panel to Review the Programs of the Bureau of Justice Statistics, National Research Council. Washington DC: National Academy Press, 2008. (as Panel member)

**A70** "Direct Estimates As a Diagnostic for Dual System Estimators Based on Logistic Regression". Pp. 1751-1758 in *Proceedings of the Survey Research Methods Section,* [CD-ROM], Alexandria, VA: American Statistical Association, 2008. (with M. H. Mulry, T. Mule, N. Nguyen).

**A71** *Ensuring the Quality, Credibility, and Relevance of U.S. Justice Statistics,* Robert M. Groves and Daniel L. Cork, Editors, Panel to Review the Programs of the Bureau of Justice Statistics, National Research Council. Washington DC: National Academy Press. (as Panel member)

**Presentations**

"Benefit-Cost Analysis of the Census", Kennedy School of Government, Harvard University, 1978.

"Benefit-Cost Analysis of the Census", Graduate School of Business, University of Chicago, 1978.

"Benefit-Cost Analysis of Data Used to Allocate Funds", U.S. Library of Congress, 1979.

"Benefit-Cost Analysis of Data", Energy Information Administration, U.S. Department of Energy, 1979.

"Benefit-Cost Analysis of Data Used to Allocate Funds", Research Conference on Public Policy and Management, Chicago, 1979

"Implications of Equity and Accuracy for Undercount Adjustment: A Decision-Theoretic Approach", U.S. Bureau of the Census Conference on Census Undercount: Proceedings of the 1980 Conference, Arlington, 1980.

"An Evaluation of Postcensal Population Estimation Methods for Substate Areas", Annual Meeting of the Population Association of America, Denver CO 1980.

"Benefit-Cost Analysis of Data Used to Allocate Funds", Department of Statistics, University of Chicago, 1980.

"Small-Area Estimates of Population and Income", Demography workshop, University of Chicago, 1981.

"Issues of Accuracy and Equity in Adjusting for Census Undercount", Annual Meetings of the American Statistical Association, Houston, 1981.

"Benefit-Cost Analysis of Data Programs", Numerical Data Advisory Board, National Research Council, 1982.

"Feasibility of Benefit-Cost Analysis of Public Data", Annual Meetings of the American Statistical Association, Social Statistics Section, Washington, D.C., 1982.

"Distribution of Federal Benefits According to Statistical Formulas: Discussion of Paper", Annual Meetings of the American Statistical Association, Washington, D.C., 1982.

"Models for Data Use: Implications for Managers", American Statistical Association, Chicago Chapter, 1983.

"Test Scores as Social Statistics", Stanford-Berkeley Quantitative Methods Seminar, University of California, Berkeley, 1984.

"Uncertain Population Forecasting", Biostatistics Workshop, Stanford University, 1984.

"Benefit-Cost Analysis of Data Programs", Statistics Department, Stanford University, 1984.

"Confidence Intervals for Population Forecasts", Morrison Research Seminar (Demography), Stanford University, 1984.

"Toward Conducting Benefit-Cost Analyses of Data Programs", Annual Meetings of the American Statistical Association, Philadelphia 1984.

"Simplifying Complex Samples with the Bootstrap" Annual Meetings of the American Statistical Association, Philadelphia, 1984.

"Benefit-Cost Approaches to Designing Longitudinal Studies", Social Science Research Council, Washington, D.C. 1984.

"Ordinal Measurement and Social Statistics", Annual Meeting of the American Educational Research Association/National Council on Measurement in Education, Chicago 1985.

"Test Score Decline: What and How Well Does It Measure?" Annual Meetings of the American Statistical Association, Las Vegas, 1985.

"Conceptual Issues in Measuring Improvement in Population Estimates", U.S. Bureau of the Census, Second Annual Research Conference, 1986.

"Benefit-Cost Analysis of Data Programs", Statistics Canada, Ottawa, 1986.

"Sampling Problems in Merging a Cross Sectional and a Longitudinal Program", Planning Conference on the National Assessment of Educational Progress and the Longitudinal Studies Program, Center for Education Statistics, U.S. Department of Education, 1986.

"On Benefit-Cost Analysis of Data Programs", International Symposium on Methodological Aspects of Empirical Sociological Research, Institute for Sociology, U.S.S.R. Academy of Sciences, Moscow, U.S.S.R., October 1988.

"On the Accuracy of Estimates of Numbers of Intravenous Drug Users in the United States", National Opinion Research Center, December 1988.

"Estimation of Sampling Probabilities for Tenth Grade Schools in NELS:88", National Center for Education Statistics, June 1989.

"Use of Expert Opinion in Mortality Forecasting", Panel on Active Life Expectancy, University of Chicago Medical School, July 1989.

"Total Error in Dual Systems Estimates of Population", University of Chicago Statistics Department, April 1990.

"Error in True Values", International Conference on Measurement Errors in Surveys, Tucson, Arizona, November 11-14, 1990.

"Total Error in PES Estimates of Population", *Journal of the American Statistical Association* Invited Applications Paper, annual meetings of American Statistical Association, Atlanta, August 1991.

"Total Error in Dual Systems Estimates of Population", Northeastern Illinois Chapter of the American Statistical Association, November 1991.

"Decision Theory and Statistical Agencies", Chicago Chapter of the American Statistical Association, February 1992.

"Choice of Loss Function", Committee on Adjustment of Postcensal Estimates (CAPE), Meeting of Experts, U.S. Bureau of the Census, July 1992.

"Eligibility Exclusions and Sampling Issues", National Academy of Education, Panel on the Evaluation of the NAEP Trial State Assessment Project, Chicago, May 1992.

"Total Error of the 1990 Undercount Adjustments", Sampling Workshop, Department of Statistics, University of Chicago, April 1992.

Discussion of "The Social Context of Individual Behavior" by James S. Coleman, Frontiers of Social Measurement Symposium, NORC, Chicago, March 1992.

Discussion of papers by Germain-Julien and Fay-Thompson, Annual Research Conference, U.S. Bureau of the Census, Washington, D.C. 1993.

"Evaluation of Design and Implementation of Sampling in the 1994 TSA", National Academy of Education, Panel on the Evaluation of the NAEP Trial State Assessment, Washington, D.C., June 1994.

"Combining State and National NAEP: Preliminary Thoughts and Plans", National Academy of Education, Panel on the Evaluation of the NAEP Trial State Assessment, Washington, D.C., October 1994.

"Combining State and National NAEP", National Academy of Education, Panel on the Evaluation of the NAEP Trial State Assessment, Washington, D.C., September 1995.

"Accuracy of Block-Level Estimates of Population", Census 2000 Committee on Statistical Policy, Bureau of the Census, June 1998.

Discussant, National Academy of Sciences workshops on the 2000 Census (October 6-7, 1999 and February 2-3, 2000).

"Total Error Modeling for Census 2000", Bureau of the Census, Washington, D.C. February 3, 2000.

"Evaluating Statistical Data: Plans for a Northwestern Research Center", Institute for Policy Research, Northwestern University, November 6, 2000.

"Census Undercount and Redistricting: Politics and Statistics of the Recent Adjustment Decision", Institute for Policy Research, Northwestern University, April 9, 2001.

"Total Survey Error and Randomized Social Experiments", Institute for Policy Research, Northwestern University, March 7, 2005.

"Total Error in Surveys, Forecasts, and Randomized Social Experiments: Modeling Approaches", National Institute of Statistical Sciences Workshop on Total Survey Error, March 17-18, 2005, Washington, D.C.

"On Assessing the Total Error of Large Scale Demographic Statistics, Including Surveys, Forecasts, and Randomized Social Experiments", Department of Statistics, Iowa State University, December 5, 2005.

"Statistical Demography with Applications", day-long course presented with Juha M. Alho at the American Statistical Association Meetings, Seattle, August 2006.

"Estimation of the Accuracy of Jury Verdicts", Royal Statistical Society, Belfast, September 10-14, 2006.

"Verdict on Jury Verdicts", Institute for Policy Research, Northwestern University, January 29, 2007.

"Verdict on Jury Verdicts: Statistical Measurement of the Accuracy of Jury Verdicts", Joint Statistical Meetings, Salt Lake City. July 30, 2007.

"On Measuring the Balance between Wrongful Convictions and Wrongful Acquittals in Criminal Trials", Second Conference on Empirical Legal Studies, NYU Law School, November 9-10, 2007.

Population Association of America, Chair of Session on PAA session on Statistical Applications in Population Research, New York, March 30, 2007

Discussant for Session on Statistical Modeling Issues in Population Research, and Chair of Session on PAA session on Statistical Applications in Population Research, Population Association of America, New York, March 30, 2007

Discussant of "Numerical Analysis in Least Squares Regression With an Application to the Abortion-Crime Debate" by William Anderson and Martin Wells, Second Conference on Empirical Legal Studies, NYU Law School, November 9-10, 2007.

"On Measuring the Balance Between Wrongful Convictions and Wrongful Acquittals in Criminal Trials", IPR Seminar on Performance Measurement in the Public and Nonprofit Sectors, Northwestern University, December 5, 2007

"On Statistical Estimation of Accuracy of Verdicts in Criminal Cases", Institute for Mathematical Behavioral Sciences, University of California, Irvine. May 1, 2008.

"Census Costs and Benefits", Meeting of Experts, November 9, 2009, U.S. Census Bureau, Suitland, M.D.

**TEACHING AND ADVISING**
**Courses taught since 1985**

Education C70 Topics in Statistics 1986-87
Education D06 Data Display 1986-87
Statistics C5O Regression Analysis 1987-88 N=9
Statistics C25 Survey Sampling 1987-88 N=14
Statistics C25 Survey Sampling 1988-89 N=13
Statistics C25 Survey Sampling 1989-90 N=8
Statistics C25 Survey Sampling 1990-91 N=13
Statistics B03 Statistics & Public Policy 1990-91 N=l0
Statistics C25 Survey Sampling 1991-92 N=13
Statistics B03 Statistics & Public Policy 1991-92 N=l0
Statistics C25 Survey Sampling 1992-93 N=11
Statistics B03 Statistics & Public Policy 1992-93 N=50
Statistics C25 Survey Sampling 1993-94 N=13
Statistics B03 Statistics & Public Policy 1993-94 N=15
Statistics C25 Survey Sampling 1994-95 N=15
Statistics B03 Statistics & Public Policy 1994-95 N=27

Statistics C25 Survey Sampling 1995-96 N=11
Statistics B03 Statistics & Public Policy 1995-96 N=40
Statistics C25 Survey Sampling 1996-97 N=15
Statistics B03 Statistics & Public Policy 1996-97 N=25
Statistics C25 Survey Sampling 1997-98 N=19
Statistics C59 Topics: Statistical Demography and Forecasting 1997-98 N=9
Statistics B03 Statistics & Public Policy 1997-98 N=45
Statistics C25 Survey Sampling 1998-99 N=13
Statistics B06 Elementary Statistics for Research 1998-99 N=50
Statistics D98 Supervised Statistical Consulting 1998-99 N=4
Statistics 325 Survey Sampling 2000-01 N=15
Statistics 325 Survey Sampling 2001-02 N=13
Statistics 320-1,2 Statistical Theory and Methods I, II 2001-02 N=8
Statistics 325 Survey Sampling 2002-03 N=15
Statistics 320-1,2 Statistical Theory and Methods I, II 2002-03 N=5
Statistics 325 Survey Sampling 2003-04 N=13
Statistics 320-1,2 Statistical Theory and Methods I, II 2003-04 N=16
Statistics 325 Survey Sampling 2004-05 N=20
Statistics 320-1,2 Statistical Theory and Methods I, II 2004-05 N=8
Statistics 325 Survey Sampling 2005-06 N=15
Statistics 101-6 How Many People Can the Earth Support? 2005-06 (Freshman Seminar) N=13
Statistics 345 Statistical Demography and Forecasting 2005-06 N=10
Statistics 325 Survey Sampling 2006-07 N=18
Statistics 101-6 How Many People Can the Earth Support? 2006-07 (Freshman Seminar) N=13
Statistics 345 Statistical Demography and Forecasting 2006-07 N=4
Statistics 461 Adv. Topics in Statistical Demography and Forecasting 2006-07 N=2
Statistics 101-6 How Many People Can the Earth Support? 2007-08 (Freshman Seminar) N=16
Statistics 325 Sample Surveys 2007-08 N=22
Statistics 325 Sample Surveys 2008-09 N=20
Statistics 325 Sample Surveys 2009-10 N=20

**Director of Dissertation Committee:**

Juha Alho, Statistics 1983 Ph.D.
Kannikar Pinyakong, Education 1984 Ph.D.
Theodora Cohen, Statistics 1990 Ph.D.
Jiahe Qian, Statistics 1992 Ph.D.
Haoliang Song, Statistics 2007 Ph.D.

**Northwestern University Committees and Service (since 1990)**

Search Committee, interschool faculty position in demography, 2007-present
Search Committee, Institute for Policy Research and School of Education and Social Policy, 2005-2008

Program Review Council, 2005-present
Chair (2008-present)
Subcommittee on Office of General Counsel (Chair), 2007-2008
Subcommittee on Department of African-American Studies (Chair), 2006-2007
Subcommittee on Technology Transfer (Chair), 2005-2006
Committee on Athletics and Recreation, 2005-2008
Committee to Plan for Immersion Experiences for Undergraduates (Provost's Office), 2005
Director of Undergraduate Studies, Department of Statistics, 2004-present
Search Committee, Institute for Policy Research and School of Education and Social Policy, 2005-present
Ad hoc committees on tenure and promotion (various) 1986-present
Committee on Research in the Post 9/11 Environment, Office of the V.P of Research 2004-2005
University Hearing Board (UHAS Appeals Board), 2004-2007
Faculty Mentor for Beth Andrews, Department of Statistics , 2003-present
Faculty Mentor for Spyros Konstantopoulos, School of Education and Social Policy 2003-2008
Committee on the Program in Mathematical Experience for Northwestern Undergraduates (MENU), 1993-present
Search Committee, Institute for Policy Research and Department of Statistics, 2003-05
Program Review Subcommittee, Jewish Studies Program, 2002-3
Chair, Graduate Admissions in Statistics, 2001-04
Change Management Committee (Provost's Office), 2000 - 2005
Chair, Department of Statistics 1986-1999, 2000-2001
University Advisory Committee for the Institute of Health Services Research and Policy Studies, 1998 - 2001
Committee on Personnel Procedures, College of Arts and Sciences (Chair) 1998-99
Ad Hoc Faculty Group on New Educational Initiatives (Provost's Office), 1997-98
Klopsteg Lecture Committee (Chair) 1995-97
College of Arts and Sciences Dean Search Committee (Chair) 1995-97
Graduate School Fellowship and Scholarship Committee, 1995-96
Advisory Committee in the Mathematical Sciences, College of Arts and Sciences, 1991-96
Faculty Search Committee, Department of Statistics (Chair), 1991-6 (various years)
Dean's Advisory Committee on Promotion and Tenure Issues (DACPTI), College of Arts and Sciences 1994-95
College of Arts and Sciences Committee on Promotion and Tenure, 1993-95
Executive Committee, Survey Research Laboratory, 1990-95
Graduate School Committee on Statistical Training in the Social Sciences (Chair) 1990-95
Program Review Subcommittee for Department of Electrical Engineering and Computer Science, 1994
Ad Hoc Advisory Committee to Assist in the Review of the Instruments and Procedures Used by the Course and Teacher Evaluation Council, 1991-94
Search Committee for Dean of School of Education and Social Policy, 1993
Ad Hoc Committee on Computers and the Undergraduate Student Experience (Chair), 1991-92
Building Renovation Committee, Department of Statistics (Chair), 1991