EXHIBIT A

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

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.

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.

> 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.

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.

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.

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.

> "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 4

Bruce Spencer
<bruce.spencer@sbcglobal.net>

02/17/2010 03:00 PM

To wvandaele@lecg.com
cc Scott Cowan <swcowan@JonesDay.com>, Laurens Wilkes <jlwilkes@JonesDay.com>
bcc
Subject Instructions for SAS Analysis

Dear Walter,

I would like you to run some analyses and send the output to me. Use the email address bruce.spencer@sbcglobal.net. If you have any questions please let me know.

The analysis requests are in 6 parts.

Part 1. Run the SAS programs provided by Dr. Levy. They are found in ORCLX-AACG-000004. Instructions for creating the data file are found in "readme.docx".

Part 2. In the code for "analysis.sas" in ORCLX-AACG-000004, provide me with formulas to show how the point estimates, standard errors, and confidence intervals are computed by SAS for measures 128, 129, 131, 132.

Part 3. Carry out the following procedure separately for Retrofit and Critical Update sample and population.

Procedure: Calculate a z statistic defined as

$$z = (\text{sample mean minus population mean})/\text{s.e.},$$

where s.e. is the ratio of the standard deviation in the population to the square root of the sample size. The standard deviation in the population is the sum of squared deviations about the population mean, divided by the population size. (Note: the divisor is not the population size minus 1.) The z statistic should be calculated for each of the numeric variables whose values are known for the entire population. Count how many variables (measures) had a z statistic exceeding 1.64485 in absolute value.

Run the procedure for the sample drawn by Dr. Levy.

Then draw 10,000 samples with replacement (using the same sample sizes as Dr. Levy) and run the procedure each time. Have SAS calculate summary statistics for the number of variables whose absolute z values exceeded 1.64485. I also want to know how many of the 10,000 samples had counts as large or larger than the sample of Dr. Levy.

Part 4. Do the following separately for Retrofit and Critical Update sample and population.

Procedure: Calculate the 90% confidence interval based on the t statistic as used by Dr. Levy and as revised to not incorporate the finite population correction. That is, use the critical value based on the t distribution with degrees of freedom equal to sample size minus 1 for Critical Updates and degrees of freedom equal to infinity for Retrofit; of course, the t distribution with infinite degrees of freedom is simply the normal distribution. Use the estimate of standard error that does not incorporate the finite population correction - calculate it exactly as he does in his revised SAS code. The confidence interval should be calculated for each of the numeric variables whose values are known for the entire population. Count how many variables had a confidence interval that did not include the population average.

Run the procedure for the sample drawn by Dr. Levy.

Then draw 10,000 samples with replacement and run the procedure each time. (These samples can be the same as the other 10,000,) Have SAS calculate summary statistics for the number of variables whose confidence intervals did not include the population average. I also want to know how many of the 10,000 samples had counts as large or larger than the sample of Dr. Levy.

Part 5a. Do the same simulation as in part 3 except that you should do it separately for each variable. That is, for each measure, count how many times out of the 10,000 samples the population average is not included in the confidence interval of the form sample average plus or minus 1.64485*s.e., where s.e. is the standard error as defined in part 3. Or, equivalently, for each variable, count how many times (out of 10,000) its z statistic does not exceed 1.64485 in absolute value.

Part 5b. Do the same simulation as in part 4 except that you should do it separately for each variable. That is, for each measure, count how many times out of the 10,000 samples the population average is not included in the 90% confidence interval as calculated by Levy.

Part 5b. For the Retrofit samle only, do the same simulation as in part 5 except that you should use the correct degrees of freedom for the t distribution instead of assuming infinite degrees of freedom as Levy did.

Part 6. In the file “repeatsample.sas” in ORCLX-AACG-000004 there is some SAS code that generates random numbers and random sample selections for 10,000 with-replacement subsamples of size 238. The code includes

```
do i=1 to 238;
   rownum = int(ranuni(110309+&j)*238);
       output;
  end;
```

I want you to run this and produce 3 rectangular output files containing, respectively, (i) the seeds used for every random number that is generated, (ii) the random numbers that are generated, and (iii) the sample selection numbers corresponding to the random numbers. The files only need to include the results for the first 20 of the 10,000 samples. Each column can refer to a sample, so there would be 20 columns, and 238 rows.

Again, if you have any questions about this please get in touch with me.

Thanks,
Bruce