BELLSOUTH TELECOMMUNICATIONS, INC.
DIRECT TESTIMONY OF JOSEPH B. THOMAS, PH.D.
BEFORE THE TENNESSEE REGULATORY AUTHORITY
FILED FEBRUARY 25, 2005
DOCKET NO. 04-00150

Q. PLEASE STATE YOUR NAME, YOUR POSITION WITH BELLSOUTH TELECOMMUNICATIONS, INC. ("BELLSOUTH") AND YOUR BUSINESS ADDRESS.

A. My name is Joseph B. Thomas. My business address is 675 West Peachtree Street, Atlanta, GA 30375. I am the Director of Statistics for BellSouth Telecommunications, Inc. In this position, I am responsible for statistical standards, methodology, and analysis for reported measurements of BellSouth's performance that is provided to competitive local exchange carriers ("CLECs").

Q. PLEASE SUMMARIZE YOUR BACKGROUND AND EXPERIENCE.

A. I have been a professional statistician for more than 9 years. I obtained a BS in mathematics in 1990 from Samford University, an MBA from Samford University in 1993, an MA in pure mathematics from the University of Alabama in 1995, an MS in applied statistics from the University of Alabama in 1996, and a Ph.D. in applied statistics from the University of Alabama in 2002 with a concentration in multivariate quality control.

Since June 1998, I have been employed by BellSouth. I am currently a Director in the Interconnection Services division of Network Services for the Measurements and Analysis Group specializing in statistical analysis. In this role I am responsible for statistical standards and analysis as they apply to reporting metrics for BellSouth. I also provide general statistical consulting across BellSouth. Prior to this position, I was a Director of Process Excellence for BellSouth where I advised process management and improvement teams. Before joining the Network Services organization I was Director of Statistical Standards and Analysis in the Corporate Strategy and Planning (CS&P) group of BellSouth Corporation. In this position I was responsible for various statistical analysis projects working with such diverse areas as our Consumer, Small Business, Broadband, Network, Information Technology, Corporate Measurements, Billing and BAPCO divisions. My group in CS&P was also responsible for the Corporate Quarterly Business Reviews. Prior to joining CS&P I was a senior statistician in BellSouth Technology Services Inc. (BTSI) working on vendor contract negotiations and BTSI performance metrics. Preceding my time at BTSI, I joined BellSouth as a statistician in Corporate Measurements performing time series analysis, forecasting, factor analysis, target setting, and resolving survey sampling issues.

Before coming to BellSouth, I taught at the University of Alabama in Tuscaloosa in the School of Engineering, the School of Arts and Sciences and the Culverhouse College of Commerce and Business Administration. Prior to that, I worked in the Actuarial department for Liberty National Fire and Casualty and in the Project and Performance Management department for AmSouth Bank.

I am a member of the American Statistical Association ("ASA"), the Institute of Mathematical Statistics ("IMS"), the American Society for Quality ("ASQ"), the ASA Alabama and Atlanta Chapters, the ASQ Greater Atlanta Section, the ASQ Electronics & Communication, Service Quality, and Statistics Divisions. I am also on the Advisory Board for the Center for International Standards & Quality in the Economic Development Institute at the Georgia Institute of Technology.

My interests and research have led to special expertise in multivariate quality control with particular experience in statistical process control, Six Sigma and process management, multivariate analysis, statistical computing and graphics, time series analysis and forecasting, performance measure comparisons, sampling, and design of experiments.

## I. INTRODUCTION

Q WHAT IS THE PURPOSE OF YOUR TESTIMONY?

A. The purpose of my testimony is to explain statistical rationale for some of the changes BellSouth seeks to make in the current Tennessee SEEM plan. I will begin by presenting the results of a statistical experiment that was performed on the current Tennessee SEEM plan. This experiment allows us to assess the true ability of the current SEEM plan to evaluate BellSouth's parity performance and levy appropriate remedies. Next, I will address the appropriateness of the truncated Z test in evaluating parity performance. Then, I will put forth the statistical justification for several of the proposed changes BellSouth has requested in the SEEM plan. I will conclude with a presentation of BellSouth's logical new procedure for calculating the number of transactions that caused a particular CLEC submetric combination to fail the parity test. This is the number of transactions that must be remedied for BellSouth to achieve parity for a particular failed submetric. While the statistical analysis about which I testify relates to measurement of "parity" services, I do not testify about Bellsouth's actual performance. That subject is addressed by Mr. Varner.

Q. HOW DO YOUR SPECIALTY AREAS IN STATISTICS RELATE TO THE ISSUES ADDRESSED IN YOUR TESTIMONY?

A. The SQM and SEEM plans were established to measure the quality of BellSouth's processes and whether or not any bias exists in BellSouth's processes. Having a proficiency in statistical process control allows me to objectively evaluate statistical techniques used to monitor processes. In this capacity I have been able to make determinations that the current SEEM plan does not appropriately assign remedy payments for BellSouth's parity performance.

## **EVALUATION OF THE CURRENT PLAN**

Q. IS THE USE OF A STATISTICAL SIMULATION EXPERIMENT AN ACCEPTED METHOD OF TESTING PARTICULAR HYPOTHESES?

A. Yes. The application of simulation techniques to evaluate statistical procedures such as those use in SEEM has gained momentum in statistical research over the past several decades with the availability and increase in power of personal computers. Many statistical disciplines now routinely use statistical simulation to create models of data that reflect predetermined industry criteria and use that data to test their procedure's effectiveness and validity. It is also used to compare several alternative techniques.

The approach of using a statistical simulation experiment is analogous to many everyday activities. For example, let's say I was shopping for a car, and wanted to see how different models looked in several different colors and trim packages. I don't need to go out and find actual cars in these different colors and trim packages. Instead, I can go to a car manufacturer's website, select the model I want and view different colors and trim packages to see how they look.

Q. HAVE YOU BEEN ABLE TO EVALUATE THE CURRENT PLAN USING STATISTICALLY PROVEN METHODOLOGIES?

A. Yes. BellSouth performed a statistical simulation experiment on the SEEM plan currently in use in Tennessee and Florida. A statistical simulation experiment allows BellSouth to apply this same approach to evaluating how well the current SEEM plan performs. In this case, BellSouth developed a model of how it performs certain activities, for example responding to repair troubles reports. Using the actual SEEM plan in place today, the model allowed BellSouth to determine the level of penalties that would be paid if it repaired troubles just as well for CLECs as it did for its retail customers, repaired them better for CLECs or worse for CLECs. Basically, what the statistical simulation does is allow us to create a situation where all parties would agree that BellSouth was providing parity service to CLECs, and then process the underlying data supporting the parity finding. If the SEEM plan is working correctly, there should be no SEEM penalties. If the SEEM plan generates a penalty even when parity service is being provided, then it follows that there is a problem with the SEEM plan that must be corrected.

Q. WHY IS A STATISTICAL SIMULATION EXPERIMENT PARTICULARLY USEFUL TO EVALUATE THE CURRENT SEEM PLAN?

A. The primary problem with evaluating the performance of the SEEM plan is there may not be full agreement about the true state of the process being measured and evaluated by the Plan (BellSouth's performance). For instance, BellSouth believes that performance data clearly shows that it is providing the CLECs with parity service. (Mr. Varner's testimony addresses this point.) On the other hand, the CLECs may contend that BellSouth is not providing service at parity. If BellSouth is correct; i.e., the true state of BellSouth's process is that the CLECs are receiving parity, or better than parity, service, and the SEEM plan still generates high penalties, then the plan is flawed. An important consequence of this conclusion is that BellSouth cannot avoid penalties simply by giving parity service to CLECs. However, if the CLECs are correct; i.e., that BellSouth is truly giving discriminatory service to the CLECs, then the existence of high SEEM penalties means that BellSouth simply needs to provide parity service to avoid penalties. If parties disagree on whether Bellsouth is in fact providing parity service, then the actual payments under SEEM do not permit us to draw any conclusions about the validity of the plan.

Here is where the statistical simulation experiment is useful. Instead of continuing to debate whether BellSouth is providing service at parity, and whether payments or lack of payments under the SEEM plan proves which side of the argument is correct, we can build a model of BellSouth's performance where parity of service provided between retail and CLEC customers is a given fact. Using the model, we can see what payments the SEEM plan produces in that case. The basic principle behind a statistical simulation experiment is to generate a controlled set of data that models desired conditions in industry, in this case parity of performance between BellSouth and the CLECs. This data can then be used to evaluate the SEEM plan's ability to perform in the expected manner. Since the true, underlying state of the process is controlled in the simulation, we can reasonably expect the statistical test to produce results consistent with the controlled state of the process. Said another way, since there is no longer any debate about whether the selected service is provided at parity, the SEEM plan should produce little if any penalty payments for that measure, if the plan is working properly.

Basically, the Plan is intended to establish consequences if BellSouth were to backslide on its wholesale performance. If the SEEM plan operates properly, then the Plan would require payment of penalties on why BellSouth's performance was not at parity. The purpose of the simulation is to evaluate whether the Plan is operating correctly. I am testing whether the payment of SEEM penalties does mean that BellSouth is not providing service at  parity, or whether the Plan is not accurately measuring whether  BellSouth's performance is below parity.

Through a statistical simulation, parity data can be generated and run through the entire SEEM plan just as real, production data would be run through the plan.  In addition, various situations where CLECs receive both better and worse than parity service can also be simulated. The performance of the plan can then be evaluated by how it treats the generated parity data.  If the Plan operates properly, then when the test uses data reflecting parity, or better, service, the SEEM plan should not require any penalties.  Likewise, when the test uses data that reflects below parity service,  then the SEEM plan should require penalties.

Q. BRIEFLY DESCRIBE HOW THE EXPERIMENT USED IN THIS ANALYSIS WAS DEVELOPED

A A full description of the simulation along with all of the results is presented in Attachment 1.  I will only provide a reasonably high level overview of the simulation experiment here.  I do want to reinforce, however, that all of the information that any statistician should need to understand what we have done, is found in the attachment.

The simulation was performed for three maintenance and repair measures, OOS (Percent Out of Service), PMRA (Percent Missed Repair Appointments) and PRT (Percent Repeat Troubles) using data for the month of February 2004.  The objective was to determine whether the SEEM plan correctly characterized the level of service that CLECs received, when compared to BellSouth's retail customers. The test evaluated how the Plan operated using data reflecting that BellSouth provided the service provided by these measures at parity, less than parity, or better than parity. The data for that month was validated to ensure that it was a good representation of the maintenance and repair functions that BellSouth typically performs when it receives a trouble report from a customer. To ensure that the data would not be biased (which would occur if BellSouth purposefully performed better for it's own customers that for the CLEC customers or vice versa) the data was randomly distributed to both BellSouth and the CLECs. This point is not intuitive, so I want to explain in more detail what we did, using the PMRA example. Assume that all customer trouble reports should be closed out by the date promised to the customer. PMRA (Percent Missed Repair Appointments) is the measure that tells how many of the reports are actually closed out by the committed time. To create data that could be used to evaluate how the Plan operates when BellSouth provides parity to CLECs, I used the following process.

Conceptually, in a completely unbiased process, whether or not an order is closed out on time or not is independent of whether the customer belongs to BellSouth or to a CLEC, meaning that BellSouth is just as likely to complete the repair as scheduled for a CLEC customer as for a BellSouth retail customer. To create data that clearly would represent such an unbiased process, the simulation experiment takes all of the customer trouble reports for both BellSouth and CLECs for February 2004 and combines them into one big group. This group represents each and every way that BellSouth handled customer trouble reports for all CLECs and retail customers in that month. To create a situation where the data would, without debate, reflect that BellSouth has provided this service to CLECs at parity service, I randomly assigned the actual February 2004 transactions to BellSouth and to the CLECs, without regard as to whether the individual result is actually associated with BellSouth or a CLEC. This means that, if a CLEC had 25 customers trouble reports that month then 25 customer trouble reports are randomly chosen from this big group and assigned to that CLEC. If another CLEC only had one customer trouble report that month then that CLEC only gets one randomly assigned customer trouble report. This procedure is done for every CLEC and BellSouth until each company is randomly assigned customer trouble reports equal to the actual number of customer trouble reports they had in February 2004.

An illustration of this random assignment process follows. Assume that each customer trouble report is a marble, red if the time limit was missed and blue if the appointment occurred as scheduled. Now we put all of the marbles (customer trouble reports) into a bag. Next I ask a CLEC to blindly pull a marble from the bag and record whether the marble is blue or red. The CLEC would then place the marble back and pick out another, until it had blindly selected one marble for each customer trouble report that they received during February, 2004. Then we go to the next CLEC and repeat the same procedure until all CLECs who had a trouble report in that month have performed this exercise. Next BellSouth would do the same thing until it has selected a marble with a value of blue or red for every customer trouble report that it received from its retail customers during that month. Notice that the entire bag of marbles is the TOTAL unbiased distribution of how BellSouth performed on customer trouble reports for all customers, both CLEC and retail for that month. As the marbles are redistributed randomly, they are dispersed without any inherent bias in the process. Each CLEC customer and each BellSouth retail customer's trouble report has the exact same chance of being a 'miss' or a 'made'. The data resulting from this exercise, as a statistical matter, will reflect parity because the likelihood of assignment of a "miss" or "made" (blue or red) marble is equal for BellSouth retail and for the CLECs.

To be clear, assuming there were a total of 2,000 trouble reports in February 2004 between BellSouth and the CLEC customers, those 2,000 trouble reports, some missed and some made, would now be distributed, in proportion to the total that each CLEC and BellSouth had for such month. Since the distribution was completely random, by definition every CLEC and BellSouth would be on a equal footing with regard to trouble reports for February 2004. Therefore, it is not debatable that the data created reflects that BellSouth provided this service at parity between its retail customers and the CLEC customers, is met.

To complete the experiment, this new 'parity' data is run through the SEEM methodology as if it were actual data. That is, the trouble reports assigned to CLEC A are processed as if they were CLEC A's trouble reports, and so forth. As noted earlier, the result should be that BellSouth passes all of the SEEM parity tests, because the data reflects a system where both retail and CLEC customer have the same opportunity for BellSouth to make or miss their appointment, in other words, parity service is being provided. If the SEEM methodology worked correctly in this circumstance, the Plan should not impose penalties on the basis of this randomly redistributed February 2004 data.

The second part of the simulation experiment was to artificially bias the data in favor of the CLECs by taking the parity data and increasing the number of missed customer trouble reports for BellSouth enough to equate to a percentage of better service for the CLECs. Biasing the data against the CLECs was also constructed. Then, this "biased" data, reflecting "above" or "below" parity performance was used to evaluate how the SEEM plan operated in either situation.

Q. HOW WAS THE STATISTICAL SIMULATION EXPERIMENT USED HERE EVALUATED?

A. In BellSouth's simulation, we used the principles of developing an appropriate experiment as set forth in Doug Montgomery's widely used book, Design and Analysis of Experiments (volume 4). These include, among others, replication, randomization and control. BellSouth took every effort to assure the validity of these simulations and I am confident that anyone duplicating this experiment will produce similar results.

Q. WHAT RESULTS DID THE STATISTICAL SIMULATION EXPERIMENT PRODUCE WITH REGARD TO HOW THE CURRENT SEEM PLAN PERFORMS?

A. The simulation study performed on Tennessee data produced startling results. Under the current Tennessee SEEM plan, BellSouth would have been required to make payments of $415,650 for the three measures included in the simulation, OOS (percent out of service), PMRA (percent missed repair appointments) and PRT (percent repeat troubles) even though the data reflected absolute parity service to all CLECs for the single month of February 2004 (as explained above.)

Even more bizarrely, when using the "better than parity service" data, the Plan still imposed penalties. Specifically, for the same measures, when giving CLECs 20% better service than BellSouth, the simulation indicated that penalty payments of $215,160 would be required. Even when giving a premium service to the CLECs that was 80% better than retail, BellSouth would still be required to pay a penalty. These results provide clear evidence of a disparity between penalty payments versus performance which points to a flaw in the SEEM plan. The test demonstrates that the design of the SEEM plan requires BellSouth to pay unreasonably high penalties, even when the service provided is at, or even above, parity. Importantly, this means that BellSouth cannot avoid unreasonable penalties by improving performance, because even if CLECs received far better service than BellSouth's retail customers the SEEM plan still generates high penalty payments.

Q. BASED ON THE STATISTICAL SIMULATIONS, DO YOU THINK THAT THE AMOUNT OF PENALTIES PAID BY BELLSOUTH IN TENNESSEE UNDER THE CURRENT SEEM PLAN IS A GOOD INDICATOR OF BELLSOUTH'S PERFORMANCE?

A. No. The test demonstrates that any conclusion that paying SEEM payments suggest that BellSouth is not performing adequately, is not accurate. For comparison with the simulation, BellSouth paid $225,900 for these measures in February 2004 in Tennessee (this is the same month from which the random data for the simulation was pulled). The simulation results indicate that BellSouth would pay on average, $302,888 in penalties for 75 failures when giving 10% better service to the CLECs, and $215,160 in penalties for 54 failures when giving 20% better service to the CLECs. Notice that the true penalty amount, $225,900, is close to the value for 20% better service, which indicates that BellSouth is paying penalties for service that exceeds parity.

It is important to remember why these plans were established as well as all of the compromises made in developing and finalizing the plan. A great statistician, George Box said, "All models are wrong. We make tentative assumptions about the real world which we know are false but which we believe may be useful." The current SQM and SEEM plans were the best we could do at the time. Many assumptions had to be made about what the competitive conditions would be in the future and how the plans would react to certain changes in the market. Given the time and experience we have gained, BellSouth believes the plans can now be improved.

Q. WHAT CHANGES TO THE CURRENT SEEM PLAN ARE SUGGESTED BY THE SIMULATION AND YOUR EVALUATION?

A. I believe the following changes would improve the statistical validity of the

SEEM plan.

- There is a need to reduce as many of the disaggregations as possible to significantly improve statistical validity of the truncated Z test.
- A method of monitoring BellSouth's overall performance against backsliding using well-known statistical techniques should be established.
- A sufficient, fixed materiality constant needs to be determined rather than an erratic estimate.
- Excessively eliminating good data needs to be controlled by removing the trimming rule.
- BellSouth should not be required to pay penalties until a measure has failed for two consecutive months.
- A reasonable way to calculate the number of transactions that caused the truncated Z test to fail has to be established for a transaction based plan.

I will address each of these changes in my testimony. Mr. Varner will address all other changes proposed by BellSouth.

Q. WOULD YOU PROPOSE TO CHANGE THE BASIC STATISTICAL METHODOLOGY USED IN THE CURRENT SEEM PLAN TO DECIDE WHETHER PARITY EXISTS.

A. No. The current Tennessee SEEM plan utilizes many statistical techniques established by a team of statisticians that were looking for a common basis upon which to build a parity plan. The statistical methodology they agreed upon was called the truncated Z test. It is not necessary to discard the truncated Z test in order to improve the plan.

BellSouth does not advocate completely replacing the current SEEM plan, because that would eliminate some tools that are useful and unnecessarily "scrap" all of the effort that went into creating the current statistical test. Many aspects of the Current Plan are not the subject of any dispute. For instance, neither BellSouth nor the CLECs have proposed to alter the basic truncated Z parity test. While the truncated Z parity test does neglect much of BellSouth's superior service to the CLECs, BellSouth understands this, and has compromised with the CLECs to address their concern about masking poor performance in one area with exceptional performance in another area. Specifically, the truncation that occurs in the truncated Z test essentially negates the superior service that BellSouth provides to the CLECs so that it does not overshadow those areas where the performance was inferior. However, this aspect of the current plan has a downside. That is, it allows the CLECs to receive far superior service in many areas and still receive penalty payments for small discrepancies in other areas unless this propensity is mitigated by other features of the plan. Over time we have been able to identify many problems, such as this, and can now propose alternatives to enhance the plan without throwing away all aspects of the Current Plan.

Q. DOES TRUNCATION OF GOOD PERFORMANCE USED IN THE CURRENT PARITY TEST HARM BELLSOUTH IN THE PARITY CALCULATIONS?

A. Yes. Any truncation of data, especially to only one side of the data as the truncated Z test does, will bias results. In this case the analysis is biased in favor of the CLECs. However, as noted above, BellSouth is willing to accept this in order to ensure that the Plan reasonably addresses CLEC concern that exceptional performance by BellSouth might mask any poor performance observed.

A principal reason for truncating positive modified Z statistics is to be able to statistically combine individual cell Z statistics into a single overall parity statistic for a SEEM submetric called the truncated Z statistic, without masking poor parity performance. An individual cell Z statistic compares BellSouth's performance for its retail customers to CLEC's customers at a very granular level for a SEEM submetric. For example, using the SEEM submetric Percent Missed Repair Appointments for the UNE-P products, an individual cell Z statistic would reflect this performance in a specific wire center, but only for orders where a technician had to be dispatched to the premises during the first half on the month for a report where the repair was needed on fewer than 10 circuits. Another cell might be for the same wire center in the same first half of the month where repair was still needed on fewer than 10 circuits, but where a technician did not have to be dispatched to the customer's premises. Each individual cell would reflect a different combination of these four variables ( wire center, half of month, whether the repair was need on more or less that 10 circuits and dispatch type) until all repair appointments for that CLEC and BellSouth's retail customers in Tennessee had been assigned to a cell. Because these four variables are the same for both the CLEC and retail customers, we are said to be making "like-to-like comparisons" at the cell level. I discuss the importance of making like-to-like comparisons at the cell level below. .

The truncated Z test allows the performance for each individual cell (such as those I just described) to be combined into one statistic for that CLEC for a SEEM submetric (PMRA for UNE-P in this example) to see if parity service is being provided without allowing good performance in one cell to offset poor performance in another cell.

Q. WOULD YOU PROPOSE ANY CHANGES TO THE TRUNCATION OF GOOD PERFORMANCE?

A. No. At this time there is no contention over the use of the truncated Z test of parity between BellSouth and the CLECs. BellSouth realizes the CLEC's concern over masking poor parity service and is willing to accept the risk associated with truncating good performance. However, BellSouth urges the TRA to balance this inherent risk in the test by altering the Plan to have the statistical benefit resulting from better aggregating sub-metrics to higher volumes per rollup. This can only be achieved by eliminating unnecessary disaggregations in the SEEM plan.

## **BELLSOUTH PROPOSES TO ELIMINATE UNNECESSARY DISAGGREGATIONS**

Q. FROM A STATISTICAL PERSPECTIVE, DOES THE CURRENT SEEM PLAN HAVE PROBLEMS PRESENTED BY THE DEGREE OF DISAGGREGATION?

A. Yes. There are two problems introduced by not aggregating more cells together. First, there are not a sufficient number of cells in 77% of the truncated Z tests. Therefore, the truncated Z tests are not as reliable as they could be, because an insufficient number of cells are being aggregated together. Second, the benefit afforded by using the truncated Z test (that benefit being that poor performance in one cell is not being masked by good performance in another cell) is not being fully utilized today. The point is, the test eliminates BellSouth's good performance in cells precisely so cells could be aggregated without fear of masking poor performance. The test fails, however, to aggregate the cells to a meaningful level. Thus, the purpose for which the test disregards BellSouth's good performance (by truncating off the individual cell Z statistics that exceeded parity), is only minimally employed. Both of these problems affect the plan's ability to give an accurate determination of BellSouth's parity performance. However, the first problem, of not having a sufficient number of cells in each sub-metric roll up, also detrimentally affects the CLECs.

Current disaggregations split the data into such small groupings that the worth of the parity tests are compromised. These small sample sizes result in unreliable conclusions. The decision is, which is more important? Should we have enough data grouped appropriately, to assure the statistical tests can be performed adequately? Or do we sacrifice statistical validity so we can sufficiently disaggregate data to such a point as to eliminate even the most remote chance of potential masking? I submit that without statistical validity, the potential masking is a moot point.

Q. WHY DO SMALL SAMPLE SIZES GIVE UNRELIABLE CONCLUSIONS?

A. The validity of a statistical test is based on certain assumptions. In fact, several statistical tests may be applicable to the same situation with one of the tests being superior if its assumptions are met and another test being superior if other assumptions are met. There are several statistical tests performed at the cell level in the SEEM plan. In the example discussed earlier where we were comparing the time it takes BellSouth to complete a repair for its retail versus CLEC customers, the individual cell Z test, also called cell Z test, is based on an assumption that the underlying data follows a normal distribution. The nice thing about the Z test is that if we have enough sample data, this test can still be performed with reasonable accuracy, but there is a minimum threshold below which the test can't consistently be considered reliable. Specifically, to be statistically significant, the sample that is used must be assumed to have a normal distribution. For example, if you had a barrel of red and blue marbles, and you wanted to know what percentage of the marbles were red and what percentage were blue, you could take a sample of the marbles by selecting marbles at random. If the sample you select, of if the group of samples that you select have a normal distribution (that is, on average they reflect the mix of the marbles in the barrel) a statistician can draw a conclusion about the mix in the barrel. However, you don't know whether the samples you select have a "normal" distribution. To try and mitigate against the possibility that the distribution of the sample is not normal, there is usually a minimum number of observations that you have to make.

The actual minimum amount of data required for a statistical test to be reliable depends on the actual underlying distribution of the population, which is not known. Thus, the most frequently used 'rule-of-thumb' is to have at least 30 observations from each of the entities being compared. This figure is roughly based on a well-known statistical concept called the Central Limit Theorem and a lot of simulation research. Similar situations exist for the other SEEM submetrics.

Q. DID THE STATISTICIANS THAT DEVELOPED THE TRUNCATED Z STATISTICAL TEST SEE THAT SMALL SAMPLE SIZES COULD CAUSE A PROBLEM WITH THE SEEM ANALYSIS AND CONCLUSIONS?

A. Yes, the statisticians that developed the test recognized this problem. However, they chose to balance the need for adequate sample sizes with the need for having true like-to-like comparisons in cell level testing. Like-to-like comparisons mean that CLEC transactions are compared to similar retail transactions (resulting in what is commonly referred to as an "apples-to-apples" comparison). The concern was that too much aggregation in the data could create problems of masking bad performance with good performance. Every effort was made to determine true like-to-like comparisons at the cell level that alleviated the masking concern but created the small sample size problem. The small sample size problem was to be resolved by ensuring that each SEEM submetric had a sufficient number of cells. There is some indication that the statisticians were viewing each measure, for example all missed repair appointments <u>for all products combined</u> as a single SEEM submetric. In contrast, the current SEEM plan disaggregates missed repair appointments into 40 individual SEEM submetrics.

Q. WHY IS THE NUMBER OF CELLS USED IN CALCULATING THE TRUNCATED Z STATISTIC IMPORTANT?

A. The truncated Z statistical test is bound by the same assumptions that govern the individual cell Z test. If the assumptions, such as normality, are not met then the effectiveness of the test is limited. Not only can the results be inaccurate, but another important feature of the truncated Z test, called the Type I and Type II error balancing may be compromised. Type I and Type II error balancing essentially means that the truncated Z parity test would ensure that the probability of failing the test when BellSouth is providing parity service equals the probability that the test would pass when BellSouth is discriminating against the CLECs. Since the former error (Type I) is important to BellSouth and the latter (Type II) is important to the CLECs the probability of these errors occurring are set to be equivalent. This error balancing is compromised when an inadequate number of cells is sued.

Q. WHAT WOULD BE AN AMPLE NUMBER OF CELLS WITH BOTH CLEC AND BELLSOUTH ACTIVITY FOR A SUFFICIENT PARITY TEST?

A. The traditional rule of thumb is to have at least 30 transactions for each party.

Q. WOULD THE LEVELS OF DISAGGREGATION PROPOSED BY BELLSOUTH IMPROVE THE STATISTICAL VALIDITY OF THE PARITY

CALCULATIONS?

A. Yes. They will make the conclusions of the tests that are performed more accurate and reliable because of the additional data that will be used in each truncated Z statistical test. Earlier I described how disaggregating too much can provide inconclusive results to statistical tests. With the current granularity in the SEEM, as well as the SQM disaggregations, the results can often be unreliable. However, if more cells can be included in the statistical test, then the truncated Z test will meet the appropriate assumptions for the statistical test and give more accurate results.

Another positive aspect of having fewer disaggregations comes from a common sense perspective; too much data can cause "paralysis of analysis". In most quality control and process management procedures, one of the first steps in controlling and monitoring a process is to identify the key metrics necessary to sufficiently measure the process. A general rule of thumb is to identify between, 1 and 3 key metrics. This allows continuous monitoring of the process without getting lost in the weeds of the secondary measures that are not as important.

Q. THE CLECS HAVE VOICED A CONCERN OVER SOMETHING CALLED MASKING CAUSED BY NOT DISAGGREGATING ENOUGH IN THE SEEM PLAN. IS THERE ANY JUSTIFICATION IN THIS CLAIM?

A. No, and this is an odd position for the CLECs to have. First, masking is the phenomenon that would occur if combining cells allowed poor performance in one cell to be offset, or masked by good performance in another cell. As previously discussed, the truncated Z statistical test was designed by statisticians representing both BellSouth and AT&T specifically to accommodate combining cells into an overall statistical test without masking. That is the principal reason that BellSouth's good performance is truncated, hence the name of the test. The concerns about chronic or systemic masking, caused by aggregating cells, are unfounded. A two year study of whether such masking occurs was performed jointly by statisticians representing BellSouth and the CLECs as a part of a Louisiana PSC sponsored review of the measurement and enforcement plan in effect in that state. The results of this study do not conclude that such masking occurs. The latest version of the Louisiana heterogeneity paper is presented in Attachment 2. In short, there is scant, and far from reliable, data that supports the assertion that masking of poor performance occurs when using the Truncated Z methodology.

## <u>ESTABLISH A METHOD FOR MONITORING BELLSOUTH'S OVERALL PERFORMANCE USING WELL PROVEN, STATISTICAL PROCESS CONTROL TECHNIQUES.</u>

Q. CAN YOU EXPLAIN THE STATISTICAL ASPECTS OF BELLSOUTH'S PROPOSED TRIPWIRE MECHANISM?

A. Yes. Mr. Varner describes the tripwire mechanism in his testimony, but I will include a brief description here for convenience. Basically to establish a baseline, BellSouth will calculate the percent of sub-metrics that would have been passed for the closest available 12 months preceding introduction of this plan. As closely as practicable, that percentage will be calculated as if the current plan had been in effect for that year. Each month, BellSouth will compare the overall percent of sub-metrics met to that baseline percentage and the result will determine the fee schedule that will apply to any fees paid. If performance is within 3 standard deviations either above or below the baseline percentage, the standard fee schedule will apply. If performance is more than 3 standard deviations below the baseline, a much higher fee schedule will apply, and if performance improves by 3 standard deviations above the baseline, no penalty will apply for that month.

BellSouth, through the use of a tripwire mechanism, is trying to identify systemic discrimination and separate it from random discrepancies. The most common way to do this is to use tools used in the field of statistical process control. Simply, statistical process control means that you measure a process. BellSouth has proposed to use the percentage of submetrics met as an overall measure of whether systemic discrimination might exist in the process of providing nondiscriminatory service to CLECs. With any such measurement you can expect some variation from month to month, for a host of reasons like weather, holidays, or just seasonal variation. Since we are trying to determine whether the measures are changing due to systemic discrimination, we need some means to identify how much the measurement result has to change before there is a high likelihood that systemic discrimination is occurring. That is where the "control" part of statistical process control comes in; it simply means that we statistically establish predetermined limits, and as long as the measure stays within those limits there is no cause for concern. For statistical process control, these limits are stated in terms of how many "spread" measures (standard deviations) about a central tendency statistic (mean). The "spread", or standard deviation, is a way of quantifying the degree of expected change in a measurement result. For SEEM, BellSouth is proposing to set these limits equal to 3 standard deviations. This means that as long as the overall percent of submetrics met stays within 3 standard deviations either above or below the baseline there is no cause for undue concern. These control limits will be used as a determination for both backsliding and significantly improved performance over the levels currently observed by BellSouth, which I understand are at least as good as performance levels when BellSouth was given interLATA authority.

The main theme here is to determine if the performance of BellSouth has significantly deteriorated or significantly improved over time. Since backsliding can only occur over time it makes sense to measure current performance against a baseline of quality service. There are many statistical tests to determine if this has occurred, but the use of a 3 standard deviation control limit was the foundation for the field of statistical quality control and is still the most widely used method of detecting if a process shifts from its consistent, predictable state. Also, most of the methods for determining if a process's performance backslides or improves requires more than one month of data to ascertain a conclusion, while the 3 standard deviation rule allows us to make a conclusion every month.

To be able to derive the 3 standard deviation control limits, an established, in-control and predictable state of a process must be established. Control limits will be constructed using this data with a 3 standard deviation spread. Each month a similar overall performance measure will be calculated and compared to the tripwire's control limits. If BellSouth's overall performance in the given month is below the lower 3 standard deviation control limit, then the proposed higher fee schedule will be used to pay per transaction for that month. If BellSouth's overall performance is within the tripwire's control limits then BellSouth will pay remedies based on the proposed standard, more rational fee schedule. If BellSouth improves its overall performance to the point it passes the upper, high performance control limit, then BellSouth is rewarded by not having to pay any penalties for that month.

This procedure also creates a positive incentive for BellSouth to improve service. If BellSouth can continue to develop performance to significantly better levels, then it is rewarded with not having to pay any penalties. Such an incentive does not exist in the current plan.

## <u>USE A FIXED DETERMINATION OF MATERIALITY: DELTA FOR TESTING MEAN MEASURES AND PSI FOR TESTING PROPORTION MEASURES</u>

Q. WHAT IS MATERIALITY AND HOW DO THE VARIABLES DELTA AND PSI RELATE TO IT?

A. As discussed in Mr. Varner's testimony, materiality establishes how much "noise" to allow before a test statistic becomes cause for concern. It is similar in concept to the control limits that I discussed in the previous section. Materiality is the level of difference between BellSouth and CLEC performance that the TRA determines is too big to have occurred by chance. It should also be what difference in performance would hinder a CLEC's ability to compete. Delta and psi are the variables used in the statistical formulas to incorporate this concept into the truncated Z test.

Q. WHAT DOES THE CURRENT PLAN'S DELTA FUNCTION DO?

A. The delta function attempts to make a business decision on materiality appear to be more scientific. The function tries to establish lower levels of delta at large volumes of data and higher levels of delta at smaller volumes of data. Materiality is not a statistical or mathematical decision and it should be constant without regard for the volume of data. The higher the level of delta, the more often BellSouth will be seen as giving parity performance. Under similar conditions, the lower the level of delta, the more often BellSouth is seen as harming a CLEC. The delta function was an attempt at trying to make this determination mathematical. It tries to use the fact that there is more randomness apparent in small sample sizes to say that the delta should be larger in those cases. It also tries to assign smaller materiality, delta values, to bigger volumes.

Q. IS USE OF THE DELTA FUNCTION APPROPRIATE?

A. No. Use of the Delta Function overlooks the fact that the large variances for small sample sizes and small variances for large sample sizes are already accounted for in the truncated Z test. Simply said, it attempts to correct what has already been corrected. Trying to vary the delta value for increased sample sizes in addition to the already diminishing variances accounted for in the truncated Z test makes the test far too sensitive to small variations in performance particularly when volumes are large. Since the variance in the Z statistic is already accounted for in the truncated Z test, having both effects incorporated in the calculations only serves to excessively magnify each other. Again remember that, materiality is not a statistical determination and having a function to evaluate the materiality factor is statistically unnecessary and cumbersome.

## **BELLSOUTH PROPOSES TO STOP EXCESSIVELY TRIMMING GOOD, USEFUL DATA**

Q. WHAT IS TRIMMING?

A. Trimming is another word for eliminating data from a statistical test. Such data is often referred to as 'outliers.'

Q. WHY IS TRIMMING USED?

A. Trimming was introduced as yet another safeguard against the possibility of excessively poor performance provided to BellSouth retail customers masking poor performance for the CLECs. As I explained earlier the truncated Z test already guards against this possibility and as a consequence, trimming is not necessary.

Q. WHAT IS THE TRIMMING RULE IN THE CURRENT SEEM PLAN?

A. The current trimming rule is fairly vague. It only requires that 3 conditions be met. One, that it can be applied in a production setting. Two, that any 'trimmed' data be examined for possible use in the final decision making before it is discarded. Three, that it should only occur on performance measures that are sensitive to "outliers".

Q: WHAT MEASUREMENTS HAVE OUTLIERS WHERE TRIMMING WOULD BE EMPLOYED?

A: Trimming would be used for the mean measurements since they are the only measurements that have outliers. Mean measurements are measurements of time duration. Maintenance Average Duration (MAD) which measures how long it takes to clear a trouble report, and Order Completion Interval (OCI) which measures how long it takes to install service are the only mean measures in SEEM.

Q: HOW IS TRIMMING APPLIED IN THE CURRENT SEEM PLAN?

A: All of the data for a particular sub-metric for a given month are separated into two groups, BellSouth retail and all CLEC transactions. The CLEC data is sorted and the largest CLEC value identified. BellSouth's retail data is sorted and any data value larger than the largest CLEC value is eliminated or trimmed from parity testing.

Q. HOW MUCH ACTUAL DATA IS BEING ELIMINATED USING THE CURRENT TRIMMING RULE?

A. Attachment 3 has the amount of data trimmed in Tennessee during 2004. In 2004, over 7,796 BellSouth retail trouble tickets and service orders have been eliminated from parity testing calculations for the two measures, Maintenance Average Duration (MAD) and Order Completion Interval (OCI). Similarly, in BellSouth's entire nine state region, 48,130 MAD and OCI, BellSouth retail transactions have fallen victim to the trimming rule.

Q. DOES BELLSOUTH PROPOSE THE USE OF ANY TRIMMING RULE?

A. No. The trimming rule in SEEM should be eliminated because SEEM measurements that had the potential for excessive influence from outliers already include an inherent safeguard. This safeguard is the cell level truncation of good performance to the CLEC, such as might occur when the BellSouth performance that was used as the comparison had an unusually long duration. This essentially caps the effect of a BellSouth retail transaction with an enormous value. Any other trimming appears to be redundant.

A single, unusually large value can only affect one cell level Z score. If the performance for this value is really poor for BellSouth retail such that the cell level statistical test indicates the CLEC received superior service compared to BellSouth retail, then the cell's Z score is truncated to reflect the same performance as a cell that was exactly at parity. Doing this only affects that single cell and does not counter any  poor service provided in the other cells with activity. Thus, the outlier's value has little effect on the aggregate truncated Z score which, ideally, should have at least 20 cells incorporated into it. On the other hand, any values that indicate discrimination toward the CLEC would not result in the truncation of a cell's modified Z score. Just one of these unusually long values could cause a WHOLE SUBMETRIC to fail the parity test when otherwise parity has been provided for the remainder of the transactions in the submetric.

In short, the way the rule is applied today; it is eliminating a significant amount of BellSouth performance data. This can cause a situation reflecting disparity while parity is achieved when using all of the data. An outlier, should it exist, should be included in this statistical test because the test can accommodate it. Eliminating the trimming rule will allow the SEEM calculations to remain in a production setting with little need for the human intervention required to analyze each trimming candidate before it is discarded. Consequently the trimming rule in SEEM should be eliminated.

## **BELLSOUTH PROPOSES A THRESHOLD OF TWO MONTHS FAILURE BEFORE PENALTY PAYMENTS ARE REQUIRED**

Q. MR. VARNER DISCUSSES PAYING PENALTIES ONLY IF THE PERFORMANCE STANDARD IS MISSED FOR TWO CONSECUTIVE MONTHS. IS THERE A STATISTICAL PRECEDENT FOR THIS?

A. Yes. As I mentioned earlier, a bias in performance is what would occur if BellSouth purposefully performed better for it's own customers that for the CLEC customers. If there is a bias inherent to a process then it is systemic and the bias would be expected to impact performance for several months, rather than just one. While better performance for a particular month could randomly swing toward the CLECs or BellSouth, an inherent bias in the process will reappear month after month. Much of the field of study called statistical quality control is built on this principle.

Q. FOR TIER I MEASUREMENTS, DOES THE CURRENT SEEM PLAN ADHERE TO THIS PRINCIPLE USED IN STATISTICAL QUALITY CONTROL?

A. No. The current plan evaluates performance each month, independent of the preceding month. While the current plan does apply escalating penalties for failures in successive months, the current plan only uses one month of data to assess whether a Tier 1 penalty should apply. When the plan was developed several years ago, it was believed that bias could be identified through the use of a snapshot of data for one month. Where this method has the benefit of immediate problem recognition, the trade-off is that of not looking at data over a period of time, which causes more uncertainty to exist in the conclusions of the statistical tests.

In contrast to Tier I, Tier II payments do take into account the possibility of random occurrences beyond BellSouth's control. For Tier II analysis, three months of data are used and three consecutive months have to fail before any penalties are administered.

Q. IS THERE ANY STATISTICAL JUSTIFICATION FOR NOT PAYING PENALTIES UNTIL A MEASURE HAS FAILED THE PARITY TEST FOR TWO CONSECUTIVE MONTHS?

A. Yes. One concern is in how often BellSouth is erroneously identified as giving disparate service to a CLEC when, in fact, BellSouth is providing parity or better service. Each time the same data is viewed there is a potential error of saying that BellSouth is discriminating when, in fact, we are not. As discussed earlier in my testimony, this is Type I error. There is also a potential error of determining that BellSouth is not discriminating when, in fact, we are (Type II error). The single truncated Z test was developed to balance Type I and Type II error probabilities.

However, with each additional test performed on the same process, the probability of Type I error goes up while the probability of Type II error goes down. Let me give an example. If a statistical test has 5% as both the probability of a Type I and a Type II error, then two statistical tests of the same data will have a 9.75% probability of a Type I error (failing a test when actually providing parity service) with only a 0.25% probability of a Type II error (passing both tests when actually providing discriminating service). These percentages are general statistical rules of thumb. This effect grows with each additional test that is performed on the same data.

Recall that Tier I SEEM data is tested with the truncated Z test. When all of the Tier I data is combined, the same data used for Tier I analysis is also tested again in Tier II analysis. Further, an individual transaction may appear in several different SEEM measurements where the data is, again, subject to a separate statistical test. For example the transactions tested for the SEEM measurement Out of Service > 24 hours are also tested separately in the SEEM measurement Maintenance Average Duration (MAD). Each one of these separate tests of the same process data increases the likelihood that test will indicate a failure, when, in fact, BellSouth provided parity service.

In Bellsouth's proposal, not paying penalties until a measure has failed for two consecutive months is an attempt at reducing the Type I errors caused by having so many duplicative tests of disparity. This is similar to the principle behind the Tier II requirement of failing 3 consecutive months before paying penalties.

## **METHOD FOR DETERMINING NUMBER OF TRANSACTIONS UPON WHICH PENALTIES SHOULD APPLY**

Q. IS IT NECESSARY TO PROPOSE A NEW METHODOLOGY FOR DETERMINING THE NUMBER OF TRANSACTIONS UPON WHICH PENALTIES SHOULD APPLY?

A. Yes. The SEEM plan currently in effect in Tennessee is a per-measurement plan. This type of plan does not require a determination of the number of transactions upon which penalties should apply. However the per-transaction plan proposed by BellSouth does require the determination of the number of transactions upon which penalties should apply.

Q. GIVEN THAT BELLSOUTH HAS PROPOSED A TRANSACTION BASED PLAN THAT REQUIRES SUCH A DETERMINATION, WHAT PROCEDURE DOES BELLSOUTH PROPOSE TO DETERMINE THE NUMBER OF TRANSACTIONS?

A. In general, BellSouth's methodology is designed to rank-order the CLEC transactions from worst performance to best, then start at the worst performance and add up the transactions to the point where the CLEC performance and retail performance 'break-even' in terms of parity. In more specific terms, BellSouth proposes a method based on the principle of systematically correcting negative cell Z scores until the truncated Z test, which is the only approved method of determining if BellSouth is giving parity service, passes. For brevity, I refer to this as the zero-out procedure.

Q. PLEASE DESCRIBE THE ZERO-OUT PROCEDURE THAT BELLSOUTH IS PROPOSING?

A. For each failed measure, BellSouth will determine the number of transactions that it will pay remedies on by successively changing the cells within that measurement where BellSouth gave the least parity (or the poorest service) into a state of parity. Specifically, the cells with the largest disparity (poorest service), as determined by the smallest (or largest negative) cell Z score, will be temporarily assigned a parity indicator, where the Z score equals zero, until the overall measure passes the truncated Z parity test. Since it is often not necessary to resolve all of the transactions in the final cell to be manipulated, the last cell will be interpolated to determine how many transactions are required to move BellSouth into a parity situation. At the end of this process, the sum of the changed transactions is the amount upon which penalties will be calculated.

This procedure is consistent with the objective of the remedy procedure which is to determine how many transactions are required for BellSouth to achieve parity. To accomplish this objective, BellSouth must temporarily assign enough transactions as if they were in parity to produce an overall passing parity test result for the measurement being evaluated. There are many potential methods to adjust transactions to obtain an overall parity result, but BellSouth determined that the best way is to address the most damaging out-of-parity situations first and then, if parity is still not obtained, to successively address the less damaging out-of-parity situations. This approach is consistent with the view that presumably a CLEC and the CLEC's end-user would be more negatively impacted by poorer performance than better performance. Arguably, the customer may be completely unaware of lesser out-of-parity situations which are usually caused by random events and do not indicate any discrimination of service.

Q. IF THE ZERO-OUT METHODOLOGY IS USED IN SEEM, WILL THE RESULTING SEEM PAYMENTS BE FOR THE TRANSACTIONS THAT ARE OUT OF PARITY AND SOME TRANSACTIONS THAT ARE ALREADY IN PARITY?

A. Yes. This method will often produce more remedy payments than actually required to achieve parity.

Q. PLEASE EXPLAIN HOW THIS IS LIKELY TO OCCUR.

A. The zero-out methodology is based on changing whole cells rather than the individual transactions within the cell. Changing whole cells instead of individual transactions can improperly inflate the number of remedy transactions required by including transactions in the cell that reflect parity or even better than parity. For example a cell can contain a mixture of transactions that are at parity, above parity, slightly below parity or well below parity. When the cell is changed, all of the transactions are counted. However, BellSouth chooses to err on the side of caution.

The final cell brought into parity regularly does not require all transactions to be addressed for parity to be achieved. An appropriate action is to interpolate how many of the transactions would need to be changed to bring the entire sub-metric into a parity situation.

Q. IS THERE JUSTIFICATION FOR ONLY PAYING ON THE NUMBER OF DISPARATE TRANSACTIONS REQUIRED TO ALLOW THE TRUNCATED Z TEST TO PASS?

A. Yes. Passing or failing the truncated Z test is the only established point at which parity detection has been agreed upon in the SEEM plan. Paying on any transaction that reduces the truncated Z score past this point causes BellSouth to pay on transactions in cases where it is uncertain that there has been any disparity in treatment and if there was a disparity the TRA has already determined that such degree of disparity is immaterial. In other words there is no statistical basis for a conclusion that CLECs have been harmed.

Q. FOR A FAILED MEASUREMENT, SHOULD THE MATERIALITY THRESHOLD USED FOR DETERMINING THE NUMBER OF TRANSACTIONS TO BE PAID BE THE SAME MATERIALITY THRESHOLD USED TO DETERMINE IF THE MEASUREMENT FAILED?

A. Yes. BellSouth insists that the definition of materiality should be the same for the pass/fail determination and for the calculation of penalties. It would not make sense to use one materiality definition to determine pass/fail and then switch to another one to determine the penalties for the failed measurement. The reason is because variability exists in any process, the truncated Z parity test was designed to help differentiate between "chance" and a true difference. In this light, as a part of this Docket, the TRA will order materiality values, delta and psi , that sufficiently distinguish a material difference from an immaterial difference; one that separates a true difference in performance from one that is merely "chance" or one that is immaterial. This material difference is used to determine the point where the truncated Z test will delineate a material difference in performance as specified by the TRA. However, the converse must also be true, it also indicates the point below which any difference is considered to be an <u>immaterial difference</u> in performance by the TRA.

Because of this reasoning, only the transactions that actually **cause** the material difference should be penalized. Transactions that, if corrected, would drive the truncated Z statistic below this level are not considered to be material in the pass/fail determination. The established materiality region allows all other transaction's differences to be considered immaterial since they do not cause the truncated Z test used in the pass/fail determination to fail. To be consistent, when the truncated Z test does detect a failure, these transactions should not be considered material when penalties are being calculated.

Q. ARE THERE OTHER REASONS FOR NOT PAYING ON ANY MORE TRANSACTIONS THAN ARE NECESSARY TO PASS THE TRUNCATED Z TEST?

A. Yes. To require payment for any additional transactions eliminates any materiality determination from the penalty calculation and penalizes BellSouth as if these transactions were causing material failures. The contradiction in paying penalties on these transactions is that the initial pass/fail determination does not treat these transactions as if they cause material failures.

Variation always exists in data. In the statistical literature and in practice, the variation can generally be classified two ways. Some variation is due to either "common causes" which result in immaterial differences. Common causes are evident in the randomness within a single distribution. Common causes simply happen. They do not require a change in the process. The second classification of variation is referred to as "special causes." Special causes are indications that the variation is due to a difference in process possibly as a result of systems and processes that have lost efficiency. In these instances the systems and processes should be corrected where possible.

It would be inconsistent to simply say that common causes are special causes and require BellSouth to pay penalties for immaterial differences. BellSouth should not be responsible for any "noise", or "chance" transaction, that occurs in the process. Yet this is exactly the implication if BellSouth is required to pay for transactions that do not have a material influence on the failure of the initial truncated Z parity test used to determine if the measurement passed or failed.

Q. EARLIER IN YOUR TESTIMONY YOU DISCUSSED BALANCING OF TYPE I AND TYPE II ERRORS. IF THE SEEM PAYMENT WERE TO INCLUDE THESE ADDITIONAL TRANSACTIONS WOULD AFFECT THIS CONCEPT?

A. Yes. This is yet one more reason to pay penalties only for transactions that are necessary to pass the truncated Z test; it is the only way to consistently balance the likelihood of a Type I or Type II error occurring in the truncated Z test. In fact, **the point at which the truncated Z test indicates whether the test passed or failed is the only point where this occurs.** In evaluating the affected volume by zeroing-out cells until the test passes, we continue to balance Type I and Type

II errors as was done in the initial pass/fail determination. This consists of zeroing-out cells, which models what would happen if BellSouth provides equal to or better performance to the CLEC in that cell, without changing the probability of a Type I or Type II error occurring.

In other words, BellSouth's proposed zeroing-out method keeps the hypothesis test's integrity in place by continuing to zero-out cells only until the cell's data that caused the initial determination of failure are fixed.

This methodology proposed by BellSouth is a new application in the SEEM plan. However it is inspired by methodology frequently used in area of statistics called multivariate quality control, which is the area of statistics in which I specialize.

Q. PLEASE SUMMARIZE YOUR TESTIMONY.

A. The plan is intended to determine whether or not BellSouth is providing parity of performance when service provided to CLECs is compared to service provided to BellSouth retail customers. Consequently, when BellSouth provides parity service, there should no SEEM payments. My testimony shows that the current SEEM Plan does not function as intended. The statistical simulation of the current Tennessee SEEM plan using actual data demonstrates that BellSouth pays substantial penalties even when providing better than parity service to CLECs. Accordingly, any conclusion that the payment of SEEM penalties somehow suggests that BellSouth is not providing adequate (i.e. parity) service to CLECs is inaccurate. From a statistical validity perspective, the SEEM plan can be improved by increasing the number of observations associated with determining whether BellSouth is providing parity service. This can be done by reducing the level of disaggregation in the plan; using a fixed or constant delta function; and removing the existing trimming rule. Because the parties propose retaining the Truncated Z test, CLEC concerns about good performance masking poor performance are unfounded (because superior service provided to CLECs is "truncated" or not counted when making parity determinations). Further, payment of SEEM only after BellSouth has failed a given measure for two consecutive months will help ensure that SEEM payments are made when there are systemic performance misses rather than paying penalties because of random occurrences.

BellSouth is proposing a transaction-based SEEM plan. A transaction-based plan requires a determination regarding the number of transaction upon which penalties should apply. To make such a determination, BellSouth is proposing the "zero-out" procedure. Under the zero-out procedure, to determine the number of transactions that BellSouth will pay SEEM fees, BellSouth will rank CLEC transactions (or cells) from worst to best, then will correct negative cell Z scores until the parity test is passed (i.e. the Truncated Z test). From a statistical perspective, the zero-out procedure ensures that BellSouth pays SEEM penalties only when BellSouth fails to provide parity service, which is, of course, the sole purpose of the SEEM plan.

Q. DOES THIS CONCLUDE YOUR TESTIMONY

A. Yes.

BellSouth Telecommunications, Inc.
Tennessee Regulatory Authority
Dpcket No. 04-00150
Direct Testimony of Joseph B. Thomas, Ph.D.
Exhibit No. UBT-1

# EVALUATION OF TENNESSEE SEEM PLAN USING A STATISTICAL SIMULATION EXPERIMENT

## GOAL

The goal of the simulation experiment is to evaluate BellSouth's SEEM enforcement plan for situations such as: parity, better than parity, and worse than parity. Parity will be defined as BellSouth providing nondiscriminatory or unbiased service to the CLECs. Better than parity will occur when BellSouth biases performance in favor of the CLECs. Worse than parity will occur when BellSouth discriminates performance against the CLECs. The underlying hypothesis for the testing will be that BellSouth should not pay any penalties when either parity or above parity service is given to the CLECs. Also, BellSouth should pay adequate penalties for discriminatory service.

A more narrowly defined goal was to evaluate the number of parity test failures and determine the amount of remedies to be paid when parity exists.

## UNDERLYING ASSUMPTIONS

This simulation experiment was performed in accordance with standard principals of a statistical experiment as stated in Doug Montgomery's book, Design and Analysis of Experiments. These include replication, randomization, and control.

**Replication** – We want to be assured that the conclusions drawn through the process of this experiment will not be based on an anomaly. Therefore, we ran the entire experiment 30 times to get better picture of the possible outcomes. Most of our conclusions were drawn from the average of the 30 experiment runs.

**Randomization** – We want to be assured that there is not any bias in the process that was not introduced by the experiments. To accomplish this, for each measurement, we randomized all of the underlying data before we redistributed it back to the CLECs and to BellSouth in accordance with the original volumes for the CLECs and BellSouth respectively.

**Control** – We can also keep unintentional bias from impairing our experiment by controlling all of the outside influences. We have accomplished this by randomly assigning unbiased observations from a controlled distribution of data from the measurement. The controlled distribution consists of ALL of the ways BellSouth performed the particular process being measured during a particular month. Controlled bias was introduced into the experiment by the 'below parity data runs' and the 'above parity data runs'.

## TECHNICAL SPECIFICS

In this simulation we generated a month run using data representing activity during February 2004. This was done separately for three maintenance and repair measures, OOS (Percent Out of Service), PMRA (Percent Missed Repair Appointments) and PRT (Percent Repeat Troubles) for the state of Tennessee. We then created several scenarios:

- A parity situation
- A discriminatory situation where BellSouth was given 20% better service than the CLECs
- Eight (8) above parity situations, equivalently spaced between 0% and 80% better service for the CLECs than for BellSouth.

Volumes in this simulation were designed to replicate real CLEC volumes by using actual CLEC activity during February 2004.

The simulation allows us to determine how much we would have paid under the current plan in each of the aforementioned scenarios. For each existing CLEC that had activity in a particular measure in February 2004 in Tennessee, a comparison against the actual penalties paid that month can be done. The comparison should indicate whether or not BellSouth pays penalties when giving parity service to the CLECs for the measurement as designed by the experiment. Each of our 3 measures was compared in a separate simulation run.

## GENERAL OVERVIEW:

For each measurement, a parity situation was first "created" for CLEC and BellSouth data. This was done by gathering, into a pool, the actual BellSouth transactions and CLEC transactions for the measurement. Transactions were then randomly selected, with replacement, from that pool and assigned to be either CLEC or BellSouth transactions. These transactions are assigned in an unbiased, random way, thus assuring that the data observed in each CLEC's group of transactions and in BellSouth's group of transactions will come from an unbiased process and have to be considered unbiased data. ALL of the simulated data was distributed WITHOUT any distinction possible between BellSouth and CLEC data, therefore parity can be assumed for the data randomly selected from the pool of all original production data which is the actual data from the month of February 2004.

The method for creating the pool of all original production data can easily be explained using Tier 2 cells for a measurement. Recall that the CLEC side of a Tier 2 cell already combines/aggregates all the CLEC activity comparable to the appropriate BellSouth activity within a given wire center for the measurement. Therefore each pool involved in our simulation could have been generated by gathering all the CLEC and BellSouth transactions that fall into the same Tier 2 cell. Note that since Tier 2 cells exist only when there is both CLEC and BellSouth activity for that cell's characteristics, the total number of pools should equal the number of Tier 2 cells for these measures in February 2004.

Generating the disparity situations posed a unique problem. Because BellSouth stores these numbers in integer numerator and denominator form, we chose to allow the numerators of the observations to be real numbers represented by decimals. To create data reflecting CLECs levels of service ranging from 20% worse to 80% better, BellSouth or CLEC numerator was altered accordingly. If we wanted the situation where BellSouth had 20% better service than the CLECs, then we would multiply the BellSouth numerator by 0.8. To clarify, each of the three measurements used in this simulation are in the category where a lower number is better. For instance for the measurement Percent Missed Repair Appointments, the fewer missed repair appointments there are, the better. Hence to achieve a 20% improvement in Missed Repair Appointments, we reduce the missed appointments in the numerator by 20% - i.e. multiply by 0.8. Similar explanations apply to OOS and PPT. If we wanted the situation where BellSouth gave 20% better service to the CLECs, then we would multiply the CLEC numerator by 0.8 (to favor the CLECs). The denominators could not be altered without affecting the total number of transactions during a month. The numerator only affects how those transactions were categorized as pass or fail.

Because the volume of transactions and the geographical distribution for both the CLEC and ILEC activity may affect the results, selection was important. Using the wrong model, volume and/or geographical distribution, may have led to suspect information. To make the simulated CLECs as similar as possible to actual CLECs, we used the real CLEC's company codes, volumes and geographical distribution to create the CLECs used in the simulation. For any real CLEC that had activity in the measures OOS, PMRA and PRT in Tennessee, a similar CLEC with the exact same geographical distribution and exact same volume, in exactly the same wire centers, was created. We matched the simulated number of transactions with the real data all the way down to the cell within each wire center. In each wire center, the same volume of data, for the same half of the month, product group, service order type, circuit type etc. as there was in February 2004, in Tennessee for the measurements OOS, PMRA or PRT was generated. This "perfect" replication was done on both BellSouth's and the CLEC's side. This avoided the need to model the CLEC/BellSouth geographical distribution and volume of activity and provided us with meaningful simulations of CLEC and BellSouth activity. Another advantage of "mirroring" CLECs and BellSouth geographic activity and volume so closely is that we can compare with actual results. Because the simulated data mirrors the real data with parity introduced, the results of the simulation can be compared with the actual results for February 2004. To summarize, we can tell how much we would pay, if in parity, and compare with how much we actually paid.

An important note is that the data used in the disparity situations is the same data that was randomly selected to create the parity situation. In other words, we randomly selected once for each cell, per simulation run, for all ten of the situations, i.e., the data was the same for BellSouth and the same for the CLEC. Then the disparity situations were introduced using the same random parity data. This allows us to compare the parity situation with the other two known disparity situations. A new random selection would have added some ``noise`` and could have skewed the two disparity situations.

After the simulations were generated for CLECs and BellSouth, a regular run of PARIS production code was performed separately using each of these new sets of data. In the "production run", each of the parity situations was treated independently.

As mentioned earlier, to provide statistically sound results we generated thirty (30) such runs. By averaging the results for each CLEC across these thirty runs we reduce the impact of potential ``outliers`` (data that falls far outside the normal range) giving us a more stable picture.

## RESULTS

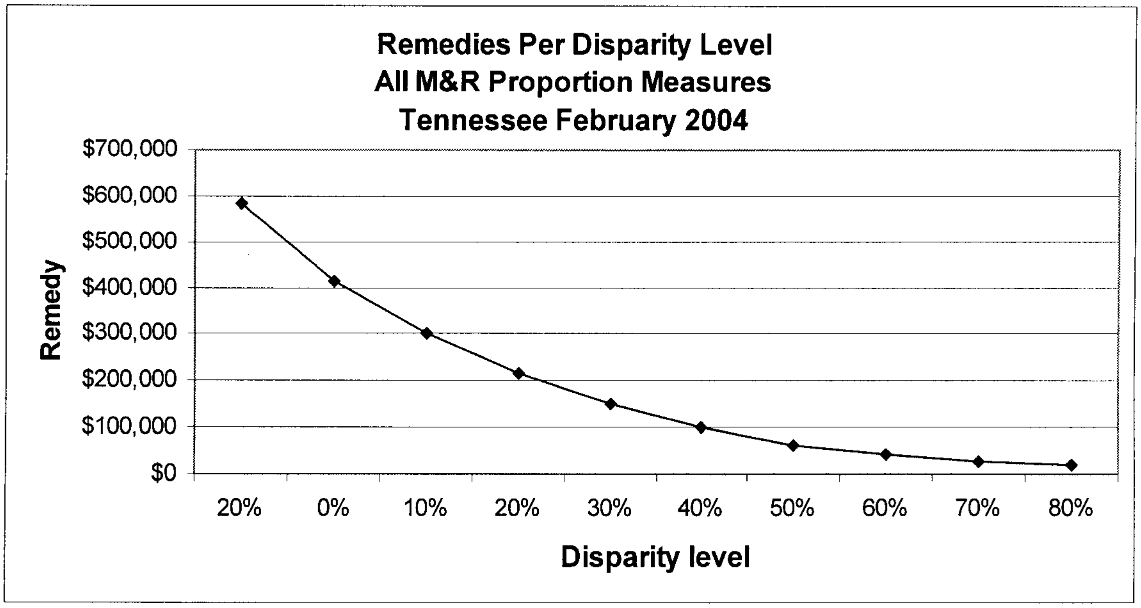
1. The summary results of the simulation are presented here. The first table shows the average number of failures and average amount of remedies expected for the simulated parity situation (disparity is 0%) and all of the aforementioned disparity situations. The disparity of -20% represents when BellSouth's process is biased against the CLECs. The disparities of 10% through 80% represent when BellSouth is giving the CLECs better service than it does to its own customers. Recall that an assumption of the SEEM plan is that BellSouth should not pay anything at parity and yet the simulation shows BellSouth having to pay an average of $415,650 per month under the current SEEM plan.

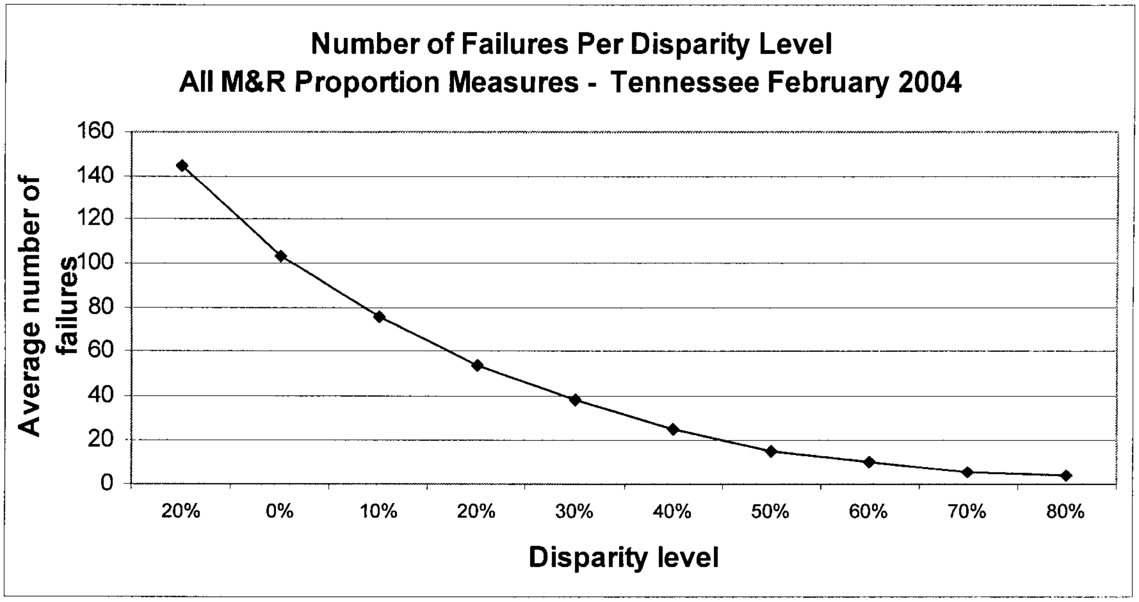
The averages are across all 30 times that the simulation was performed.

ALL M&R PROPORTION MEASURES
TENNESSEE - FEBRUARY 2004

| Disparity | Avg Failures | Avg Remedy |
|---|---|---|
| -20% | 144.43 | $583,818 |
| 0% | 103.2 | $415,650 |
| 10% | 75.4 | $302,888 |
| 20% | 53.87 | $215,160 |
| 30% | 37.9 | $150,915 |
| 40% | 25 | $99,817 |
| 50% | 15.73 | $63,255 |
| 60% | 10.53 | $43,525 |
| 70% | 6.03 | $26,292 |
| 80% | 4.2 | $19,240 |

**The next two charts graphically depict what was shown in the previous table.**



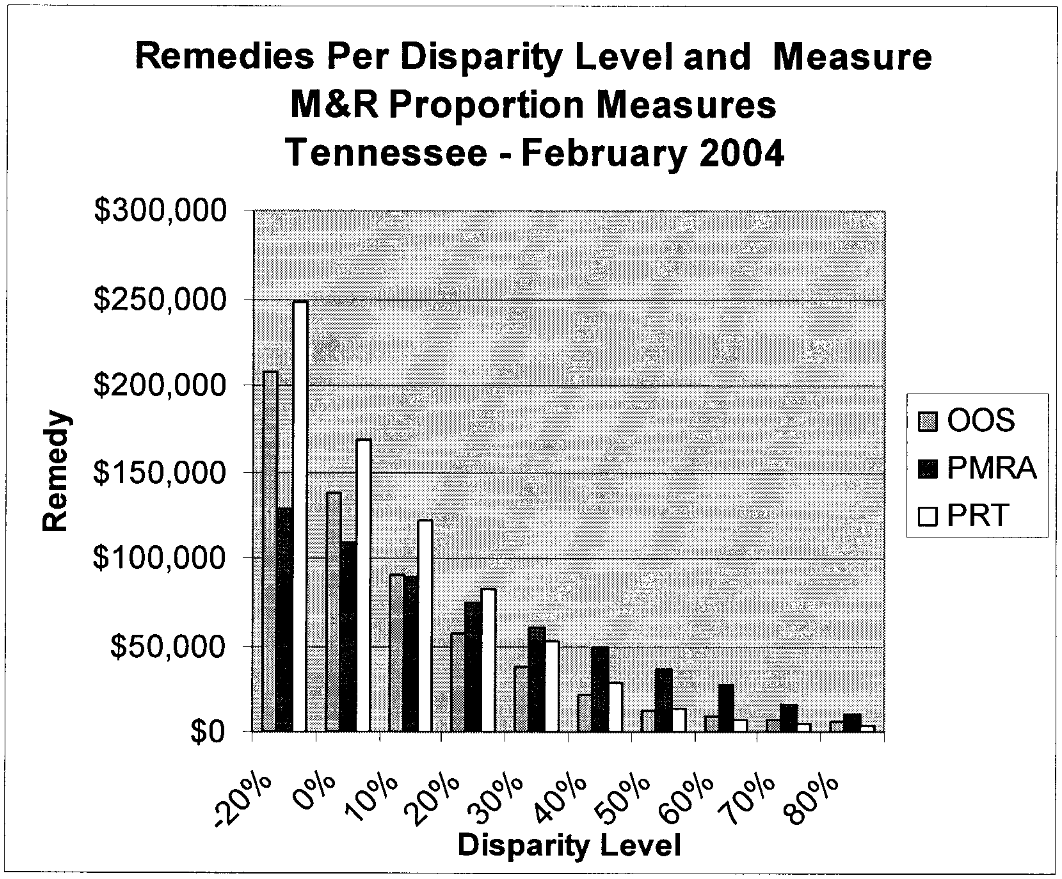


Note that the two graphs above look very similar. This is due to the plan being measure based. For every submeasure that fails a truncated Z test, a base remedy is charged to BellSouth. In a transaction based plan this would look different.

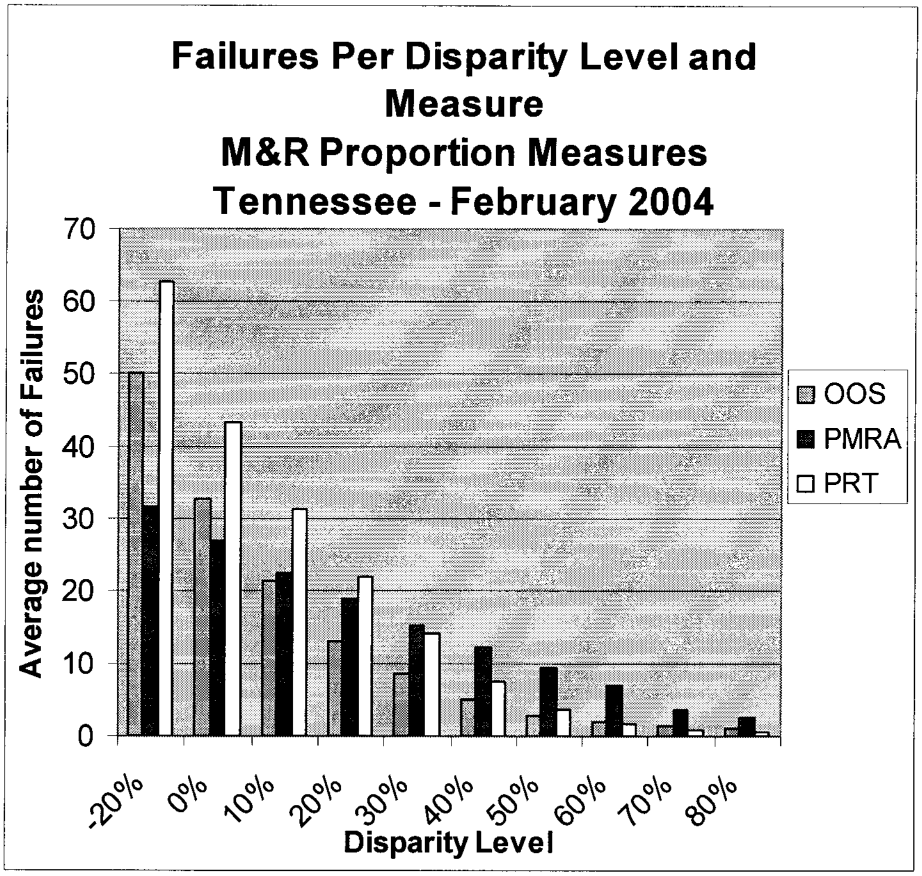
2. The next table has the same data as given in table 1 but separated into the three proportion measures used in the simulation, OOS, PMRA and PRT. Notice once again that BellSouth is required to pay remedies when providing parity service. Also note that even when providing sufficiently better service for the CLECs, BellSouth still pays some penalties. The two charts graphically depict the data in the table.

| | OOS | | PMRA | | PRT | |
|---|---|---|---|---|---|---|
| Disparity | Avg Failures | Avg Remedy | Avg Failures | Avg Remedy | Avg Failures | Avg Remedy |
| -20% | 50.1 | $207,800 | 31.57 | $128,168 | 62.77 | $247,850 |
| 0% | 32.9 | $138,052 | 27.03 | $108,883 | 43.27 | $168,715 |
| 10% | 21.5 | $90,647 | 22.43 | $89,992 | 31.47 | $122,250 |
| 20% | 13.13 | $56,940 | 18.9 | $74,983 | 21.83 | $83,237 |
| 30% | 8.57 | $37,615 | 15.27 | $60,565 | 14.07 | $52,735 |
| 40% | 5.1 | $22,213 | 12.33 | $48,880 | 7.57 | $28,723 |
| 50% | 2.87 | $12,788 | 9.37 | $36,800 | 3.5 | $13,667 |
| 60% | 1.93 | $9,065 | 6.9 | $27,568 | 1.7 | $6,892 |
| 70% | 1.4 | $6,650 | 3.73 | $15,603 | 0.9 | $4,038 |
| 80% | 1.13 | $5,383 | 2.4 | $10,690 | 0.67 | $3,167 |

*Note that the baselines do not account for the missing value – 10% disparity.

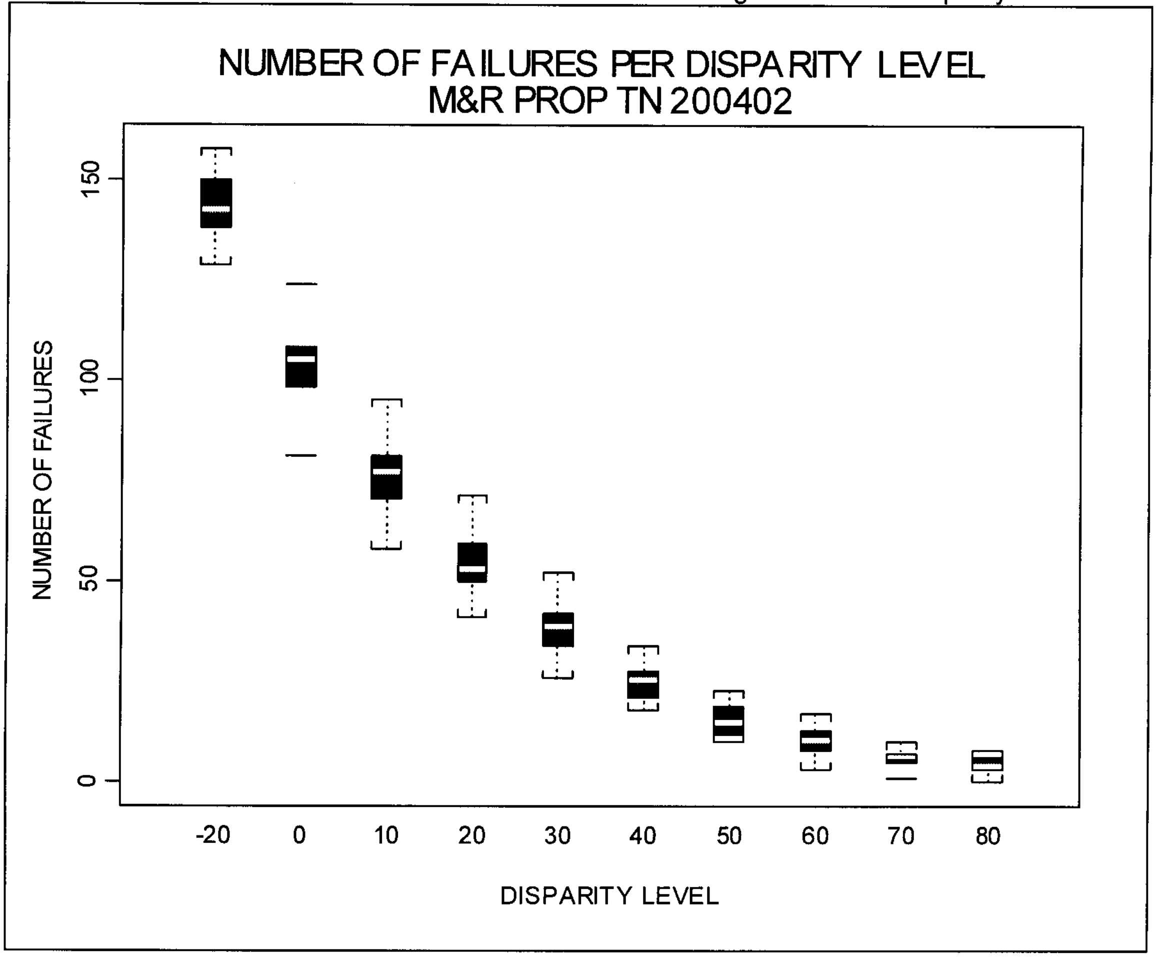


BellSouth Telecommunications, Inc.
Tennessee Regulatory Authority
Dpcket No. 04-00150
Direct Testimony of Joseph B. Thomas, Ph.D.
Exhibit No. UBT-1
Page 6 of 7



3. This last chart shows the spread of each disparity simulated for each of the 30 replications (runs) performed during the simulation. Although this chart depicts failures, it is important to note that not any of the 30 random runs allowed BellSouth to not pay penalties when providing parity service. In fact, when giving as much as 60% better service to the CLECs, BellSouth was always required to pay some remedies. Also note that the number of failures in the current measure-based SEEM plan only change slightly with large improvements in service.

*Note that the baselines do not account for the missing value – 10% disparity.



# Statistical Analysis of SEEM Disaggregation and Reaggregation
## Follow Up to BellSouth Statistical Team's Report Filed April 21, 2003

Over the last year an in-depth analysis of the statistical components of BellSouth's Louisiana Self-Effectuating Enforcement Mechanism (SEEM) system has been undertaken jointly by BellSouth Telecommunications, Inc. (BellSouth) and Competitive Local Exchanges Carriers (CLECs). BellSouth filed a report of the analysis on April 21, 2003. That report suggested that more analysis needed to be completed before recommendations concerning changing or not changing the SEEM system should be made.

This report explains the subsequent analyses that have taken place. This report is organized into four sections. Section I provides background information about the SEEM plan and why the Louisiana Public Service Commission (LPSC) staff requested the analysis. A summary of the results of January 2002 through April 2003 performance measurement data analysis is provided in Section II. An outline of BellSouth's and the CLECs' recommendations for future actions is provided in Section III. Section IV provides descriptions of supporting documents that are attached to this report as appendices. There are four appendices attached to this report.

## I. Background

SEEM is a system that performs agreed upon calculations in order to assess when the service provided by BellSouth to CLEC customers is as good as the service BellSouth provides to its own customers. When the system's calculations where retail analog standards apply indicate sufficient evidence supporting a disparity in service quality, additional calculations are performed to determine a penalty amount that BellSouth pays. Some of the calculations performed within the SEEM system are based on statistical hypothesis testing methods, and are referred to as *parity testing* calculations.

The parity testing methods in the SEEM plan try to answer the question "Are CLEC customers receiving service that is (significantly) worse than that received by similar BellSouth customers?" In order to do this, performance measurement data first must be disaggregated to insure that CLEC transactions are compared with similar BellSouth transactions (like-to-like comparisons). The statisticians refer to this as disaggregation to the *cell-level*. The cell-level is generally a very deep disaggregation level, and it is not necessarily the level at which parity judgments should be made.

Statistical reaggregation techniques are used within the SEEM system for many reasons that are associated with sound statistical practices. For example, CLEC sample sizes are sometimes very small for individual cells, and this can lead to "noisy" (imprecise) comparisons at the cell-level. In the reaggregation stage, cell-level measures of evidence about the service relative to parity received by CLEC customers (*modified Z-scores*) are combined to produce a single test statistic for a submeasure (*a truncated Z-score*). Comparison of the truncated Z-score with the balancing critical value produces a single compliance determination for a submeasure.

CLECs have voiced concern to the LPSC that the current reaggregation levels used in the SEEM plan potentially mask discrimination. Reaggregation may combine cells that differ substantially from each other in terms of the quality of service received by CLEC customers *relative to the service received by BellSouth customers.* For example, for a given submeasure, assume that CLEC customers with dispatched orders systematically receive better service than that received by BellSouth customers in the corresponding "like-to-like" cells. On the other hand, assume that CLEC customers with non-dispatched orders systematically receive poorer service than BellSouth's customers in the corresponding cells. In this case, there is not a single correct answer to the question posed above (Are CLEC customers receiving service that is (significantly) worse than that received by similar BellSouth customers?). For dispatched orders, the answer is "no," but for non-dispatched orders the answer is "yes."

In response to the CLECs' concerns, the LPSC staff asked a team of statisticians, representing both BellSouth and the CLECs, to review SEEM performance measurement data and determine if there was any statistical evidence of masking that would call for changes in the way the data are disaggregated at the cell-level, and reaggregated at the submeasure level in the parity testing process. Two forms of masking were defined as follows:

> Masking of Discrimination. There is the potential *masking of discrimination* where BellSouth passes the test when the subgroups are not split out, but BellSouth would have failed one of the tests had the subgroups been split out.
>
> Masking of Parity. There is also the potential *masking of parity* where BellSouth fails the test when the subgroups are not split out, but BellSouth would have passed one or both of the tests had the subgroups been split out.

AT&T statistician Dr. Robert Bell represented the CLECs in this process, and BellSouth had PricewaterhouseCoopers LLP statistical consultant Dr. Edward Mulrow, one of the authors of the Louisiana Statisticians' Report,[1] participated in the analysis. A team of statisticians from Ernst & Young LLP, including Dr. Mary Batcher, Ms. Susan Garille Higgins and Ms. Ru Sun, also participated in the analysis, and provided most of the data processing work. BellSouth also requested that Dr. Fritz Scheuren, another author of the Louisiana Statisticians' Report, join the team partway through analysis. Other representatives from BellSouth, AT&T, as well as a LPSC staff representative also provided input at various stages.

There were no statistical tools available to assess whether or not masking occurred in the SEEM system, so the statisticians applied related concepts and developed two diagnostic tools to assess the situation. The main diagnostic studied by the statisticians was a test for heterogeneity. The statisticians used the following definition of heterogeneity:

[1] "Statistical Techniques For The Analysis And Comparison Of Performance Measurement Data." Submitted to the LPSC, Docket U-22252 Subdocket C. Revised February 28, 2000.

Heterogeneity. *Heterogeneity* is a systematic tendency for relative performance on a submeasure to be better for one subset of transactions (group of cells) than for another subset.

Since all cell-level Z-scores are produced on a standardized scale, distinguishing homogeneity and heterogeneity was difficult but in the end turned out to be doable. The team developed a test statistic, $Z_{AB}$, which is designed to have a standard normal distribution for an arbitrary split of a homogeneous group of cells. However, when heterogeneity exists $Z_{AB}$ should systematically deviate from zero. A diagnostic graphical tool was also developed to assess when masking was taking place. These two diagnostic tools together allowed the statisticians to see if there was any association between heterogeneity and masking. Section II of Appendix 1 provides more detail on these concepts.

Through the use of the $Z_{AB}$ statistic to determine if heterogeneity was present and diagnostic graphical tools, the statisticians explored SEEM performance measure data from the January 2002 through April 2002 time period. This exploration enabled Dr. Bell to lay out a set of criteria for judging when heterogeneity was systematically present. This set of criteria was then applied to the May 2002 through April 2003 time period. Additionally, the diagnostic graphic tools were used to determine when masking was present during the same time period. The results and conclusions of this analysis are presented below.

## II. Results

The work done jointly by the CLEC and BellSouth statisticians began with an exploratory phase where the SEEM performance measurement data for the period January 2002 through April 2002 were examined. These results were reported to the LPSC on April 21, 2003.[2] The two diagnostic analysis tools already mentioned were created during this period, but because so little data had been examined, only four months, there was insufficient information to draw conclusions about individual submeasures. A further test of 12 more months was agreed to with data from May 2002 through April 2003. It is the results from these additional 12 months that will be focused on in this report.

As in the original analysis of January 2002 through April 2002 data, only those situations where cell counts were generally[3] at least 20 for the CLECs and BellSouth were

[2] Over the 4-month period from January 2002 through April 2002, there were 128 combinations of measure, mode, factor, and month that are examined. Descriptions of these combinations can be found in Table 1. In over half (57%) of these combinations, there is no heterogeneity detected and in all but 1 of these cases there is no evidence of potential masking. Of the 55 (43%) cases where heterogeneity is detected, 34 (62%) are cases where there appears to be no evidence of potential masking, 6 (11%) cases of potential masking of parity service, and 15 (27%) cases of potential masking of discriminatory service. Of these 15 cases, 9 (60%) occur for PMIA, Mode 1 for various categories. The other 6 cases are distributed more or less evenly among ACNI, MAD, PT30, and RT30.

[3] The January 2002 through April 2002 analysis included one case of a cell count of less than 20: PMIA, Mode 1, Non-Residence, March 2002 had a cell count of 15. The May 2002 through April 2003 analysis

examined.[4] (See Appendix 2.) Table 1 details the cases that meet this criterion. Also, as was done in the analysis of the January 2002 through April 2002 data, since the transaction count in Tier I cells can frequently be less than 20, only Tier II cells were considered for analysis.

**Table 1. Measure† – Mode‡ – Factor Combinations Analyzed**

| Measure | Dispatch Status: Dispatched & Non-Dispatched | Order Type: Change & New or Transfer | Product Group: Residence & Non-Residence |
|---|---|---|---|
| **PMIA** | Modes 1, 4 | Modes 1, 4 | Mode 1 |
| **ACNI** | Modes 1, 4 | Modes 1, 4 | Mode 1 |
| **OCI** | Modes 1, 4 | Modes 1, 4 | Mode 1 |
| **PT30** | Modes 1, 4 | Modes 1, 4 | Mode 1 |
| **MRA** | Modes 1, 4 | | Mode 1 |
| **MAD** | Modes 1, 3, 4 | | Mode 1 |
| **RT30** | Modes 1, 3, 4 | | Mode 1 |
| **CTRR** | | | Mode 1 |

. Measure Abbreviations: PMIA = Percent Missed Installation Appointments; ACNI = Average Completion Notice Interval; OCI = Order Completion Interval; PT30 = Provisioning Troubles Within 30 Days; MRA = Missed Repair Appointments; MAD = Maintenance Average Duration; RT30 = Repeat Troubles Within 30 Days; CTRR = Customer Trouble Report Rate.

. Mode Abbreviations: Mode 1 = Resale POTS; Mode 2 = Resale Design; Mode 3 = UNE Loops; Mode 4 = UNE Loops and Port Combos; Mode 5 = Interconnection Trunks; Mode 6 = UNE xDSL; Mode 7 = UNE Line Sharing.

To move from the early exploratory phase in our first analysis, Dr. Bell developed a set of criteria to be used to confirm whether heterogeneity existed for a given measure – mode – factor combination. The criteria, which require statistically significant patterns of $Z_{AB}$ values in the anticipated direction, were designed to sharply limit the likelihood of finding heterogeneity where none existed. This confirmatory analysis used May 2002 through April 2003 data to test pre-specified hypotheses suggested by the evidence of heterogeneity in the January 2002 through April 2002 data. The analysis determined that heterogeneity was present for 15 combinations of measure, mode, and heterogeneity factor (e.g., dispatched versus non-dispatched), involving 12 distinct submeasures.[5] (See Appendix 3.)

included six cases of cell counts less than 20: PMIA, Mode 1, Non-Residence, December 2002, February 2003, March 2003, and April 2003 had cell counts of 17, 12, 13, and 16, respectively. RT30, Mode 3, Non-Dispatched, May 2002 and December 2002 had cell counts of 16 and 18, respectively.

[4] While analyzing this system over the past few years, Dr. Mulrow determined through computer experiments (that is, statistical simulations) that, for many situations, 20 is an acceptable number of cells in order to have a truncated Z-score without severe skewness problems.

[5] The 15 combinations of measure, mode, and factor (shown in parentheses) are ACNI, Mode 1 (Dispatch Status and Order Type); ACNI, Mode 4 (Dispatch Status and Order Type); PMIA, Mode 1 (Dispatch Status and Order Type); MAD, Mode 1 (Product Group); MAD, Mode 4 (Dispatch Status); MRA, Mode 1 (Dispatch Status); MRA, Mode 4 (Dispatch Status); OCI, Mode 1 (Order Type); PT30, Mode 1 (Order Type); PT30, Mode 4 (Order Type); CTRR, Mode 1 (Product Group); and RT30, Mode 1 (Product Group).

Over the 12-month period from May 2002 through April 2003, there are 384 combinations of measure, mode, factor, and month that are examined. Descriptions of these combinations can be found in Table 1. In over half (65%) of these combinations, there is no heterogeneity detected. In all but one of these cases there is no evidence of potential masking.

Of the 134 (35%) cases where heterogeneity is detected, there are 112 (84%) with no evidence of potential masking, 2 (1%) cases of potential masking of parity service, and 20 (15%) cases of potential masking of discriminatory service. Of these 20 cases, 11 (55%) occur for PMIA, Mode 1 for various categories. The other 9 cases are distributed more or less evenly among ACNI, MAD, MRA, and RT30. In short, of the 384 combinations of measure, mode, factor, and month, there were 21 (5%) cases of potential masking.

For Tier II, a penalty payment is computed only if BellSouth fails for three consecutive months for a given measure – mode combination. The team looked for instances where masking of discrimination eliminated a situation where penalty payments should have been calculated. In other words, were there any combinations of failure and potential masking of discrimination that occurred for three consecutive months? During the 12-month period from May 2002 through April 2003, potential masking of discrimination occurred just once for three consecutive months (November 2002 – January 2003). This was for PMIA, Mode 1, Product Group.

In addition, repeated potential masking of discrimination occurred, although not in two consecutive months, for three of the 12 submeasures identified as heterogeneous. (See Appendix 4.)

- PMIA, Mode 1, Dispatch Status: 3 out of 12 months
- MRA, Mode 4, Dispatch Status: 4 out of 12 months
- RT30, Mode 1, Product Group: 2 out of 12 months.

## III Recommendations

The recommendations provided in this section are of two types: (1) Recommendations for changes in the basis system itself, and (2) Recommendations for further research on SEEM.

### *Action Recommendations*

The statisticians agree on the findings reported in the Results section. However, there is a lack of consensus about the appropriate action to recommend based on these results. The table below details areas of agreement and disagreement of various recommendations.

| **Recommendation #1: Split Three of the Existing Submeasures Further[6]** | |
|---|---|
| **Dr. Bell's Position** | **Dr. Scheuren's Position** |
| 1. For each of these three submeasures, the current aggregation has masked strong evidence of subparity performance on multiple occasions from May 2002 through April 2003. Depending on the submeasure, truncated Z-score values of less than -2.2 for a pre-specified subgroup of cells were masked two, three, or four times in twelve months. For each submeasure, at least one truncated Z-score value reached -3.30 (corresponding to a P-value of less than 1 in 2,000). Consequently, there is no need for and nothing to gain by continued analysis of more months of data for these submeasures. | 1. Dr. Bell's concerns about two and maybe all three of these submeasures may be warranted. This is true despite the fact that the link anticipated between heterogeneity and masking of parity turned out to be weaker than expected. Also, there seems to be little evidence that masking of discrimination for these three measure-mode-factor combinations might become more frequent in the future. In fact, the masking of discrimination for these three became relatively less frequent in the May 2002 through April 2003 period (25%) than it had been in the January 2002 through April 2002 period (42%). |
| 2. Whether it is a good idea to collapse some submeasures is a question that requires business expertise beyond that of the statisticians. Presumably, the decision to create separate submeasures for each of the seven modes was based on a business judgment that these distinct sets of products involved distinct service processes that should not be combined for performance measurement. | 2. We agree with Dr. Bell's observation that the decision to create separate new measures, whether by combining them or further splitting them, should be based mainly on a business decision. Therefore, we asked BellSouth to use their business judgment to propose three cases where collapsing three current submeasures would make sense, so as to balance the three measures that might have to be split.[7] |
| 3. On the other hand, data analysis can shed light on the assertion that Recommendation #1 would inappropriately increase the probability of Type I errors (suggesting a need to counter this with the collapse of three pairs of submeasures). Past data indicate that the probability of a Type I error for any of these submeasures has been essentially zero because the truncated Z-score statistic was being driven by a group of cells with very good service (see Appendix 2 and Table 1 of Appendix 4). As long as this remains the case, the only type of error that is possible for the other group of cells is a Type II error. Consequently, there is no need to compensate for any submeasures that are split. | 3. We believe BellSouth should prepare to implement Dr. Bell's proposal but worry about the possible increase in Type I error and that is why we are recommending a period in which a compensating change be made to keep the number of measures unchanged. We do not agree with the reasoning underlying Dr. Bell's position regarding Type I error. Instead we feel that a period of further testing, where an alternative is considered alongside what is now being done would be prudent. The key phrase in Dr. Bell's observations is the qualifier to his opinion that begins "as long as this remains the case." Without the presence of further evidence, due diligence would suggest the need to compensate for any measures that are split. |

[6] Split three submeasures into six submeasures: (1) Split PMIA–Mode 1 into PMIA–Mode 1–Dispatched and PMIA–Mode 1–Non-Dispatched or into PMIA–Mode 1–New or Transfer Orders and PMIA–Mode 1–Change Orders. (2) Split MRA–Mode 4 into MRA–Mode 4-Dispatched and MRA–Mode 4-Non-Dispatched. (3) Split RT30–Mode 1 into RT30–Mode 1–Residence Products and RT30–Mode 1–Non-Residence Products.

[7] BellSouth proposed that they would (1) collapse the two submeasures Resale POTS (Mode 1) and Resale Design (Mode 2) into Resale and (2) collapse the three submeasures UNE Loops (Mode 3), UNE xDSL (Mode 6), and UNE Line Sharing (Mode 7) into UNE Loops.

| **Recommendation #2: Split Seven of the Existing Submeasures Further**[8] | |
|---|---|
| **Dr. Bell's Position** | **Dr. Scheuren's Position** |
| 1. While masking by the formal definition did not occur from May 2002 through April 2003 for any of these submeasures, there were instances of large negative truncated Z-scores in the hypothesized direction that were masked (-2.50 for OCI, Mode 1; -4.22 for PT30, Mode 4; and -3.61 for CTRR, Mode 1). Furthermore, there is the potential for masking in the future. If service deteriorates in coming months, there would be little or no chance to detect it using the current submeasure aggregations. Simply monitoring these submeasures for nine more months means that poor performance could easily go unremedied for a year or more. | 1. For these measures masking arguably happened so infrequently that the problem is "in the noise" and not warranting adjustment, unless a wholesale series of changes were to be made. (Potential masking of discrimination did not occur during May 2002 through April 2003 for MAD–Mode 4, OCI–Mode 1, PT30–Mode 1, PT30–Mode 4, CTRR–Mode 1, or MAD–Mode 1. Potential masking of discrimination occurred one time out of 12 months for MRA–Mode 1, Product Group.) We agree that masking may occur in the future but propose that only further regular monitoring be done and that this be done in a timely manner, perhaps quarterly. |
| 2. There is no reason not to split these seven measures. As with the three submeasures listed in Recommendation #1, the risk in terms of increased Type I error is very small. In contrast, two other submeasures for which systematic heterogeneity was observed, ACNI Modes 1 and 4, are excluded from this recommendation because splitting them would increase the probability of Type I errors. | 2. There seems to be little evidence that masking of discrimination for these seven measure-mode-factor combinations might become more frequent in the future. In fact, the masking did not occur for these seven in the May 2002 through April 2003 period (0%) as it did in the January 2002 through April 2002 period (4%). To reiterate, only regular monitoring is proposed, using the same approach that was taken on data from January 2002 through April 2003. |

[8] (1) Split MAD–Mode 4 into MAD–Mode 4–Dispatched and MAD–Mode 4–Non-Dispatched. (2) Split MRA–Mode 1 into MRA–Mode 1–Dispatched and MRA–Mode 1–Non-Dispatched. (3) Split OCI–Mode 1 into OCI–Mode 1–New or Transfer Orders & OCI–Mode 1–Change Orders. (4) Split PT30–Mode 1 into PT30–Mode 1–New or Transfer Orders & PT30–Mode 1–Change Orders. (5) Split PT30–Mode 4 into PT30–Mode 4–New or Transfer Orders & PT30–Mode 4–Change Orders. (6) Split CTRR–Mode 1 into CTRR–Mode 1–Residence Products & CTRR–Mode 1–Non-Residence Products. (7) Split MAD–Mode 1 into MAD–Mode 1–Residence Products & MAD–Mode 1–Non-Residence Products. Note that this recommendation only discusses masking of discrimination. As noted in the Results Section of this report, masking of parity also occurs for certain submeasures. No recommendations for masking of parity are being proposed.

***Research Recommendations***

The joint statistical work has been a success and should continue at a modest level, if only as a matter of due diligence. After all, the methods currently used by BellSouth in SEEM to compare the service it provides to its customers with the service that it provides to the CLECs' customers are very complex. Occasionally SEEM appears to have small failures favoring either BellSouth or all CLECs in total. A way to discover and assess these is needed and to determine what (if any) repairs are warranted. To this end there are three specific consensus recommendations offered:

> Regular Monitoring. Because of the complexity of SEEM a joint team of BellSouth and CLEC statisticians should monitor results regularly. Every twelve months appears sufficient. Initially this would be done by continuing the current examination of heterogeneity and masking but eventually, depending on the two further recommendations made below the monitoring might shift to other system factors. Short reports from this monitoring would be produced regularly for the LPSC by the joint statistical team.
>
> Tier I Masking. Heterogeneity and masking have only been examined on a subset of Tier II data because there is not a sufficient amount of data at the Tier I level to perform the analysis. But we should be careful in drawing conclusions about Tier I based on Tier II analysis; it does not necessarily follow that heterogeneity and masking exist at the Tier I level even if it exists at the Tier II level. There is a consensus among the statisticians that further work here might be useful, if only to develop new diagnostic tools similar to those employed at the Tier II level in the current analysis.
>
> Distributional Concerns. There are several distributional issues that exist in the current system. For example the current SEEM model assumes a normal distribution of the truncated Z-scores. In fact, the distribution may be skewed. As has been proposed in the past, this should be researched to determine whether this weakness is big enough to warrant a fix. There are many cases where small numbers of cells are employed in the calculations, challenging distributional assumptions. These distributional concerns may need research attention. A systematic research effort on extreme values seems needed. The definition of these anomalies and some root cause analysis should be performed. For example, there are unexpected extreme $Z_{AB}$ values that are frequently observed with the ACNI measure. Each of these examples individually and collectively raises concerns of normality of the test statistic ($Z_{AB}$) under the null hypothesis and under the alternative hypothesis.

## IV. Supporting Documents

The following appendices are supplied as supporting documents to this report:

Appendix 1: Statistical Analysis of SEEM Disaggregation and Reaggregation (*Appendix 1 – LA Stat Analysis Summary-21Apr2003 final - changes accepted.doc*)

This document was filed with the LPSC on April 21, 2003 as the Statistician's Report. It summarizes the results of analysis performed on January 2002 through April 2002 data and was submitted in response to LPSC Docket Number U-22252-C. This document includes the following appendices:

- Appendix A: Louisiana Disaggregation Analysis (*Appendix A-LA Disaggregation-2Apr2003.doc*)
- Appendix B: An Analysis of the Time of Month Characteristic: A Report of Some Work in Progress (*Appendix B-Time of Month Results-4Apr2003.doc*)
- Appendix C: Heterogeneity and Masking Appendix (*Appendix C-Heterogeneity and Masking-15Apr2003.doc)*

Appendix 2: Heterogeneity and Masking May 2002 – April 2003 (*Appendix 2 – Heterogeneity and Masking-2003_1119-DRAFT.doc*)

This document provides details of the analysis of May 2002 through April 2003 data. It is an updated version of the Appendix C-Heterogeneity and Masking-15Apr2003.doc file that was submitted as Appendix C to the April 21, 2003 filing.

Appendix 3: Results of Heterogeneity Assessment Associated with Pre-Specified Hypotheses for May 2002 to April 2003 (*Appendix 3 - Results of Heterogeneity Assessment-2003_0902.doc*)

This document was prepared by Dr. Robert Bell. It summarizes Dr. Bell's assessment of heterogeneity for pre-specified submeasures based on the information provided in Appendix 2: Heterogeneity and Masking May 2002 – April 2003.

Appendix 4: Assessment of Masking for Submeasures Previously Determined to be Heterogeneous (*Appendix 4 - Assessment of Masking-2003_0918.doc*)

This document was prepared by Dr. Robert Bell. It summarizes Dr. Bell's analysis of masking based on the information provided in Appendix 2: Heterogeneity and Masking May 2002 – April 2003.