BELLSOUTH TELECOMMUNICATIONS, INC.

DIRECT TESTIMONY OF EDWARD J. MULROW, PH.D.

BEFORE THE KENTUCKY PUBLIC SERVICE COMMISSION

DOCKET NO. 2001-105

MAY 18, 2001

Q. PLEASE STATE YOUR NAME, WHO YOU WORK FOR, AND YOUR BUSINESS ADDRESS.

My name is Edward J. Mulrow. I am employed by Ernst & Young LLP as a Senior Manager in the Quantitative Economics and Statistics Group. I have been retained by BellSouth Telecommunications, Inc. ("BellSouth") as a statistical advisor. My business address is 1225 Connecticut Ave., NW, Washington, DC 20036.

Q. WHAT IS YOUR PROFESSIONAL EXPERIENCE AND EDUCATIONAL BACKGROUND?

A. My career as a statistical consultant spans over 13 years. While at Ernst & Young, I have been involved in a number of regulatory issues for several telecommunications companies. Prior to my employment at Ernst & Young, I was a senior scientist at Science Applications International Corporation (SAIC) where I was involved in the analyses of current and future defense systems. I also have worked as a senior sampling statistician at the National Opinion Research Center (NORC) at the University of Chicago, a mathematical statistician for the Internal

Revenue Service, and an assistant professor of mathematics for Southern Illinois University. I received a BA in mathematics from Illinois Wesleyan University, an MS in mathematics from the University of Utah, and a Ph.D. in statistics from Colorado State University.

Q. WHAT IS THE PURPOSE OF YOUR TESTIMONY?

A. I am here to address statistical issues relevant to the permanent penalty plan. Mr. Varner addresses the permanent performance measures to which the penalty plan applies. I will speak to issues involving the appropriate methodology for determining whether BellSouth is providing parity: 1) to individual Competitive Local Exchange Carriers ("CLECs") (Tier I), and 2) to the CLEC community as a whole (Tier II).

Q. PLEASE SUMMARIZE YOUR TESTIMONY.

A. The appropriate methodology for determining whether or not BellSouth is providing parity service to a CLEC's customers consists of the following three elements.

1. The Truncated Z statistical test should be used when transaction level data is available and a BellSouth retail analog exists.
2. The statistical testing methodology should balance Type I and Type II error probabilities.
3. The same methodology should be used for both Tier I and Tier II testing.

I will address each of these points in more detail in my testimony.

Q. CAN YOU PROVIDE A BRIEF OVERVIEW OF WHAT WE ARE TRYING TO ACCOMPLISH WITH THE STATISTICAL ANALYSIS THAT YOU ARE GOING TO DESCRIBE IN YOUR TESTIMONY?

A. Yes. What we are talking about here is the situation where BellSouth provides a service of some sort to its competitors, the CLECs. BellSouth also, at the same time, is providing a similar, or at least an analogous service, to its own retail operations. The question is whether BellSouth is favoring its retail operations in the provision of the particular service, or whether it is providing the same level of service to its competitors as its provides to itself.

For instance, assume that CLECs purchased widgets from BellSouth and BellSouth also provided widgets to its own retail operations which then used the widgets to provide service to BellSouth's own retail customers. If BellSouth provided the widgets to the CLECs on a two-day interval every time, and provided the widgets to its own retail operations on a two-day interval every time, then anyone could conclude that BellSouth was providing parity to the CLECs.

Similarly, if BellSouth were furnishing the widgets to the CLECs on a one-day interval, and furnishing the widgets to its own retail operations in two days, it would be evident that BellSouth wasn't providing parity, but was providing better service to the CLECs than to its own retail operations. Presumably the CLECs would not be upset with that.

The problem arises when BellSouth, in a given month, provides the widgets to its retail operations on average in two days, and provides widgets to the CLECs, on average, in 2.2 days. The question is whether the difference is attributable to random chance, or whether the difference is attributable to either some systemic problem with BellSouth's operations or some intentional act on BellSouth's part. The purpose of the statistical analysis is to provide the tools that the Kentucky Public Service Commission ("Commission") can use to make an informed judgment about whether the difference I just described is something to be concerned about or rather is simply the result of the sample used and therefore meaningless. The specific tool that I am going to describe in my testimony is a test that can be applied whenever the Commission wishes to compare two outcomes to determine whether any perceived difference in the outcomes is real or not. While the test is a statistical one, and involves statistical concepts, I believe that what we have is very workable and understandable.

Q. WHAT IS THE APPROPRIATE STATISTICAL METHODOLOGY THAT SHOULD BE EMPLOYED TO DETERMINE IF BELLSOUTH IS PROVIDING COMPLIANT PERFORMANCE?

A. The appropriate methodology to use is called the Truncated Z method with error probability balancing. Dr. Colin Mallows, a recently retired statistician from AT&T Research Labs, created the Truncated Z statistic, and then Dr. Mallows together with Ernst & Young statisticians, including myself, developed the actual Truncated Z methodology. The methodology is distinguished from the statistic in that we jointly took Dr. Mallows' formula that yielded the statistic and complemented it with such things as the error probability balancing. The collaborative effort was the result of a request by the Louisiana Public Service Commission (LPSC), lasted over nine months, and concluded in the filing of a "Statisticians' Report" with the LPSC in September of 1999 (revised February 2000 -- attached as Exhibit No. EJM-1).[1]

Q. CAN YOU EXPLAIN IN LAYMAN'S TERMS WHAT THE TRUNCATED Z METHODOLOGY DOES?

A. I can. Remember that what we are doing is comparing two outcomes to see if there is any difference. Therefore, one of the first things that must be done is to separate all of our observations into identical, or substantially identical categories. For instance, let's assume that what we are trying to compare is the performance of BellSouth with regard to order completion intervals. That is, we want to know whether the order completion intervals for BellSouth's retail operations are statistically the same as the order completion intervals ("OCI") for the CLECs. You would not want to compare a BellSouth retail residential order that requires a dispatch with a CLEC resale residential order that did not require a dispatch. The requirements for provisioning the different orders would be different.

Obviously, you can carry this concept of granularity to an extreme, but the point is that the first thing we have to do is to separate the individual observations into

[1] Typographical error corrections are attached as Exhibit No. EJM-2.

enough categories so that the comparison we are going to make is as close to being an apples-to-apples comparison as we can reasonably get it.

In our work, we call these classifications “cells.” For any particular measurement contained in BellSouth’s performance measurement plan, there could be thousands of these “cells.” Once we have these cells identified and populated with observations, we apply statistical tests to the information in the cells to put the conclusions we draw about every cell on a common footing. To make this illustration as clear as possible, I will assume that I have a cell for residential dispatched orders during the first half of the month. For illustrative purposes, I will assume that BellSouth has one observation that took 2 days, and the CLECs had a single observation that took 2.2 days, the times I used above. We would then apply a statistical calculation to those two observations, as is described in Appendix A of Exhibit EJM-1, and we would derive a “cell z-value” of -0.67. [2] The calculation of this value is not subject to a simple explanation, but is done through standard statistical analysis with which no statistician should disagree. Obviously, as the number of observations in the cell increases, the “cell z-value” may change.

I have described briefly what we would do for the individual cell. In actuality, we would make this same type of calculation for every cell (or more plainly stated, for each of the apples-to-apples comparisons that we had identified in connection with the specific measurement).

[2] The ~~z~~ score is based on the permutation methodology described in Appendix A of EJM-1.

Q. WHAT HAPPENS NEXT?

A. When we are done, we would have a large number, potentially thousands of numbers, each representing the "cell z-value" for each individual cell. The "cell z-values" would be either positive, or negative, or in some cases would be zero. The cells that have a negative "cell z-value" would represent those cells where, continuing my example from above, it appears that the interval for the CLECs was longer than for BellSouth. The cells that had a positive "cell z-value" would represent those cells where, again continuing my example, it appears that the interval for the CLECs was shorter than for BellSouth. Where the "cell z-value" was zero, there would be no apparent difference in the intervals.

Q. WHAT DO YOU DO WITH THESE THOUSANDS OF "CELL Z-VALUES?"

A. We move to the next step in the analysis, which is to analyze the "cell z-values" using a normal distribution curve. If BellSouth were providing parity, one would expect that the distribution of the values over the entire range of the cells would look just like the normal bell curve with which we should all be familiar.

This is where the idea of "truncating" the z statistic comes into play. We have z statistics for every cell. Some are positive, meaning they fall on the right side of the normal bell curve. Some are negative, which means that they are on the left side of the normal bell curve. One concern we would have is that if all of the z-values were included in the analysis, the positive z-values, if there were enough of them, might mask one or more significant negative z-values when averaging the z-values across all cells. That is, if there were a thousand cells, and 800 of them had positive z statistics, the sheer number of positive observations might hide significant negative values. Therefore, in order to prevent this, the Truncated Z methodology simply sets every positive value to zero, hence the “truncation.” By setting the positive observation to zero, it forces us to concentrate on the negative values on the left side of the bell shaped curve.

Q. WHAT DO YOU DO NEXT?

A. Remember we are now only concentrating on the lower half of the normal bell-shaped curve, and what we are going to try to do, in layperson’s terms, is to determine how far the observations we have made fall from the normal bell curve I have been talking about. You would not expect the observations to lie down perfectly on the curve. There are going to be variations and the question is how much is too much. Consequently, the next step is to calculate a Z statistic for all the cells, including those formally positive cells whose value has now been set to zero. Assuming that a statistician understood the purpose of truncating the positive values, and the selection of the cells weights, the calculation of the Z statistic for the truncated observations (the positive ones set to zero and the remaining negative observations left as they were found) should not be subject to dispute. This calculation will leave you with a single number that represents the truncated Z statistic value for the particular measurement contained in BellSouth’s plan for which the observations were made.

Q. DOES THIS CALCULATED Z STATISTIC BY ITSELF REPRESENT A STATISTICALLY SIGNIFICANT DIFFERENCE IN THE PERFORMANCE BELLSOUTH PROVIDED TO ITS RETAIL OPERATIONS AND THE CLECS?

A. No, generally you can't draw any conclusion from the Z statistic itself. It is just a number. However, if the number turns out to be positive (which, even though it seems illogical because of changing the positive values to zero, could occur) you could just ignore the result. If it is negative, however, you still have to have a number to compare the Z statistic to, in order to determine whether the difference represented by the Z statistic is significant.

Q. ONCE YOU HAVE THIS NEGATIVE Z STATISTIC, THEN, WHERE DO YOU GET THE NUMBER THAT IT IS COMPARED WITH IN ORDER TO DETERMINE WHETHER THERE IS A STATISTICALLY SIGNIFICANT DIFFERENCE IN THE SERVICE PROVIDED TO THE CLECS AND THE SERVICE BELLSOUTH PROVIDES TO ITSELF WITH REGARD TO THE SPECIFIC ITEM THAT YOU ARE MEASURING?

A. There are several ways of determining the number that is used for comparison. Given the constraints of a self-effectuating system, the best way, in my opinion, is to use what we call "Error Probability Balancing." Using this approach allows the observer to determine both that the observed difference is statistically significant, and that it is material. I will discuss this in more detail subsequently in my testimony.

Q. WHAT ARE SOME OF THE OTHER WAYS?

A. The most common statistical method used is what we call the "fixed critical value." Let me explain what this is, and why it shouldn't be used here. One of the main issues statisticians have to face in determining whether there is a statistical difference between two numbers is controlling the probability that the observed difference indicates a failure to provide parity when in fact parity has been achieved. We call these kind of errors, where it appears that there is a statistically significantly difference when there is in fact not one, a Type I error. To illustrate this point, consider the situation where a person is flipping a coin. Everyone knows that on average, heads should come up the same number of times as tails. Suppose you flip the coin five times, and just as a matter of chance, tails comes up every time. You might then conclude that something is wrong with the coin, that the coin is somehow biased toward tails because it is not acting in accord with what we know to be correct. In fact, the coin may be perfectly okay, and what we are seeing is simply a Type I error.

One way to determine the "critical value" that is to be compared to the Z statistic that we have been talking about is to determine what the acceptable level of a Type I error is, and when that is done, a "critical value" can be calculated using standard statistical tools. For instance, if you wanted the probability of a Type I error occurring limited to less than a 5 percent chance, the calculated "critical value," based on a standard normal distribution, would be –1.645. Every statistician in the world would agree with the calculation of that number given the criteria we have laid out.

Q. WHAT WOULD YOU DO WITH THIS "CRITICAL VALUE" IF THAT WERE THE APPROACH TAKEN?

A. This is what is called a "fixed critical value." All you would have to do is compare the truncated Z statistic that we obtained as described above, with this value. If the truncated Z statistic were positive or closer to zero than the "fixed critical value" a statistician would conclude that the observed difference was not statistically significant and that there was no actual difference between the observed measurements.

Q. IF IT IS THAT SIMPLE TO USE A "FIXED CRITICAL VALUE" WHY DON'T WE JUST AGREE TO THAT APPROACH?

A. The problem is that while the "fixed critical value" can tell you whether the observed differences are statistically significant, it cannot tell you whether the differences are material. Let's use an example. Suppose the observed interval for residential dispatched orders furnished to BellSouth's retail operations is 4.1 days. Suppose the observed interval for the CLEC is 4.3 days. Using a "fixed critical value" it might be possible to get a truncated Z statistic for these measurements that was less than –1.645, that is, that was much larger in magnitude (farther from zero in the negative direction). That would tell you that the two numbers were statistically different. However, someone would then have to look at the actual numbers, 4.1 days versus 4.3 days, and determine whether the difference is material. Did it really make a difference to the CLEC or the CLEC's customers that it took two-tenths of a day longer, on average, to provide service to the CLEC's customer? Maybe it does and maybe it doesn't. Using the "fixed critical value" cannot answer that question, which means that another analysis will have to be made in each case where there is a statistically significant difference observed. This is not practical for a self-effectuating system that is supposed to determine parity on a timely basis.

Q. DOES THE USE OF THE "ERROR PROBABILITY BALANCING METHOD" FIX THIS PROBLEM?

A. It does. Using "Error Probability Balancing" we determine a "balancing critical value" which allows you to determine whether an observed difference is statistically significant and material all at the same time. Therefore, there is no need for another analysis and no dispute as to whether two-tenths of a day is material or not. The application of the "balancing critical value" provides both answers.

Q. CAN YOU TELL US MORE ABOUT THE DIFFERENCE BETWEEN THE "FIXED CRITICAL VALUE" AND THE "BALANCING CRITICAL VALUE?"

A. Certainly. I have already described how the "fixed critical value" is determined. The "balancing critical value" introduces another dimension and that involves what we call Type II errors. A Type II error is where the observed data suggests that parity has been achieved, but in fact it has not. In the simplest terms, a Type I error hurts the ILEC because it says the ILEC didn't provide parity when in fact it did. A Type II error hurts the CLEC because it says that BellSouth provided parity when it did not. What the “Error Probability Balancing” method does is make the probability of committing either of the two different types of errors equal. You will recall when I was discussing the “fixed critical value” I talked only about having the probability of a Type I error at a level less than 5 percent. With a “balancing critical value,” we are saying that the number we are using to compare to the Z statistic reflects the probability that there will be just as many Type II errors as there are Type I errors. In other words, we don’t worry about whether there is a 5 percent chance of a Type I error or a 30 chance of a Type II error. Rather, we derive a figure that yields an equal probability of either type of error. There are formulae that are used to make the calculation that yields a single number that can be then compared to the Z statistic we talked about earlier.

Q. CAN YOU DISCUSS THESE FORMULAE?

A. The formulae are outlined in Appendix C of Exhibit EJM-1, and are difficult to describe in a short statement. The formulae are dependent upon the type of performance measure (mean, proportion, rate), the number of BellSouth and CLEC transactions, and the “delta” that is selected for use in the formula.

In a simple scenario with a large number of BellSouth transactions, an approximate value can be calculated by taking the negative of the square root of the number of CLEC transactions and multiplying it times the “delta” divided by 2. I know that this is not intuitive, but once again these formulae are ones that a well-trained statistician would agree are appropriate, and would yield a critical value that represents a balancing of the Type I and Type II error probabilities. For instance, if we selected a “delta” of 1, and we had 25 CLEC observations, the appropriate critical value to compare the truncated Z statistic to would be -2.5. If the Z statistic were less than –2.5 (that is, it is further from zero than –2.5) there would be a statistical difference and it would be material, thus avoiding the problems associated with the “fixed critical value” approach.

If the Z statistic were greater than –2.5 (that is, the Z statistic was closer to zero or positive), it would indicate that the difference was not statistically significant and the analysis would be at an end.

Q. WHY IS THIS METHODOLOGY APPROPRIATE?

A. First of all, Dr. Mallows created the truncated Z statistic so that it possesses five important properties.

1. It is a single, overall index on a standard scale; that is, you can use a normal bell shaped curve to make judgments.
2. If transaction counts for BellSouth and the CLECs across comparison cells (classifications) are exactly proportional, the aggregate index should be very nearly the same as if we had not disaggregated. This means that if the granular disaggregation I have discussed really wasn’t necessary, you will still get the same results.
3. The contribution of each cell depends on the number of transactions in the cell.

4. As far as possible, systematic discriminatory performance in some cells is not masked by good performance in other cells.
5. The final result does not depend critically on minor details in the data; that is, small changes in transaction values only induce small changes in the final result.

Second, the methodology follows the four key principles that Dr. Mallows and the Ernst & Young team laid out.

1. Like-to-Like Comparisons. When possible, data should be compared at appropriate levels; for example, CLEC transactions that are "new" provisioning orders should be compared with "new" BellSouth provisioning orders.

2. Aggregate Level Test Statistic. Each performance measure of interest should be summarized by one overall test statistic giving the decision maker a rule that determines whether a statistically significant difference exists.

3. Production Mode Process. The decision system must be developed so that it does not require intermediate manual intervention.

4. Balancing. The testing methodology should balance Type I and Type II error probabilities. A Type I error adversely affects BellSouth; a Type II error adversely affects an CLEC. Balancing the error probabilities ensures that both sides assume the same level of uncertainty in the decision process.

Q. CAN YOU EXPLAIN WHAT THE TERM "DELTA" ENCOMPASSES?

A. "Delta" is a factor that is used to identify whether a meaningful difference exists between the BellSouth and CLEC performance, in addition to a statistically significant difference. It is a rather complex concept so let me try to use a very simple example to illustrate what "delta" does. I want to caution you that this is a simplistic example that I am offering just to try to illustrate this complex point. Let's assume that for a given month, the mean (average) time that BellSouth took to provision a dispatched residential retail order was 5 days. Assume further that the standard deviation associated with that mean or average was also 5 days. If this were a normally distributed data set, then about 68 percent of all of these services were provisioned for BellSouth customers within a period of 0 days to 10 days. In actuality, the data set is not normally distributed because no provisioning time can be less than 0 days, but as I said, I am oversimplifying the example to illustrate a point. If we continued to assume a normally distributed data set, the remaining 32 percent of BellSouth's customers would fall equally above and below that spread of 0 to 10 days. (Again, I recognize that the interval cannot be less than 0 days. I am trying to use a simple example to illustrate the point.) Let's now assume that the "delta" or the materiality factor we choose was "1." This means that as long as the average time taken to provide the relevant service to the CLECs did not exceed the BellSouth mean (5 days) plus one-half of the standard deviation I mentioned (5 days), the difference would not be material. That is, if the mean for the CLECs for this period were 7.5 days or less, the difference would not be material. I arrived at the conclusion that the difference could not be more than one-half of the BellSouth standard deviation by dividing the “delta” of one by two, as I set out in my formula above.

Let’s consider another very simple example to illustrate what happens when “delta” is reduced. Assume the exact same facts as above, but use a “delta” of “0.5.” In that case, the difference between the BellSouth average for the month and the CLEC average for the month for the same measure could only be 6.25 days, instead of 7.5 days. The question that the selection of “delta” raises is how close is close enough in terms of materiality. Is it material that BellSouth took 2.5 days longer on average to provide service to the CLEC than to its own retail services? Is it material that BellSouth took 1.25 days longer, on average?

Q. HAVE THE STATISTICIANS DETERMINED THE APPROPRIATE VALUE FOR “DELTA”?

A. No. While statistical science can be used to evaluate the impact of different choices of these parameters, there is not much that an appeal to statistical principles can offer in directing specific choices. BellSouth witness Varner will discuss the appropriate delta. Specific choices should be made based on economic/business judgment.

Q. DO ANY ASPECTS OF THE STATISTICAL METHODOLOGY NEED TO BE CHANGED FOR TIER II ENFORCEMENT MECHANISMS?

A. No. The statistical methodology for comparing the service experience of all CLEC customers to BellSouth customers remains the same. One may want to consider changing the value of "delta" however. When the statisticians were putting together the "Statisticians' Report" for Louisiana, it was thought that it might be prudent to use a smaller value of "delta" for Tier II testing. The reasoning behind this is that when one combines all CLEC transactions together, poor service to a few small CLEC's could be masked by better service to the rest of the CLECs. One way to try to avoid such masking is to use a small materiality threshold. Whether or not this is necessary, and how much smaller "delta" should be for Tier II compared with Tier I, are questions subject matter experts and regulators should answer. As was stated before, the statistician should still play a role in this process so that the impact of various choices can be assessed.

Q. DOES THIS CONCLUDE YOUR TESTIMONY?

A. Yes.