EXHIBIT F

TIER-1 AND TIER-2
ANNUAL ENFORCEMENT MECHANISM CAPS BY STATE

AL	$ 9,500,000.00
FL	$ 32,500,000.00
GA	$ 20,500,000.00
KY	$ 6,000,000.00
LA	$ 12,000,000.00
MS	$ 6,500,000.00
NC	$ 12,500,000.00
SC	$ 7,500,000.00
TN	$ 13,000,000.00

Statistical Techniques For The Analysis And Comparison Of Performance Measurement Data

Submitted to Louisiana Public Service Commission (LPSC) Docket U-22252 Subdocket C

1. Introduction and Scope

The Louisiana Public Service Commission (LPSC) staff has requested Drs. S. Hinkins, E. Mulrow, and F. Scheuren of Ernst & Young LLP (consultants for BellSouth Telecommunications), and Dr. C. Mallows of AT&T Labs-Research to set out their views on the application of a statistical analysis to performance measurement data. The present report is intended to provide a detailed statistical report on appropriate methodology.

The setting for the analysis is crucial to the interpretation of any statistical significance that might be found. There is no doubt that, to quote the Commission staff, "statistical analysis can help reveal the likelihood that reported differences in an ILECs performance toward its retail customers and CLECs are due to underlying differences in behavior rather than random chance" (Staff Final Recommendation, LPSC Docket No. U-22252 - Subdocket C, dated August 12, 1998, pages 15 - 16).

To frame our presentation the next paragraph from the LPSC Docket U-22252 is quoted in its entirety.

> **"Statistical tests are effective in identifying those measurements where differences in performance exist. The tests themselves cannot identify the cause of the apparent differences. The differences may be due to a variety of reasons, including: 1) when the ILEC and CLEC processes being measured are actually different and should not be expected to produce the same result, 2) when the ILEC is employing discriminatory practices, or 3) when assumptions necessary for the statistical test to be valid are not being met."** (Ibid., page 16)

Apparent statistically significant differences in BellSouth and CLEC performance can arise when

- the ILEC and CLEC processes being measured are actually different and should not be expected to produce the same result
- the ILEC is employing discriminatory practices, or
- assumptions necessary for the statistical test to be valid are not being met.

To meet the Louisiana Commission's purpose, we will recommend techniques that are robust in the presence of possible assumption failure, carefully examine BellSouth Telecommunications (BST) and CLEC performance so "like" is compared only to "like,"

and are still able, in a highly efficient manner, to detect differences. Upon investigation any differences detected might lead to concerns about possible discriminatory practices.

The LPSC staff also states "that a uniform methodology which identifies those items which need to be measured, how they are to be measured, and how the results are to be reported is also desirable and would be beneficial to all parties" (Ibid., page 16). We agree with this goal as well, stipulating only that the use of a single method may not be desirable while a single methodology (or a set of methods) could be.

The statistical process for testing if CLEC and ILEC customers are being treated equally involves more than just a mathematical formula. Three key elements need to be considered before an appropriate decision process can be developed. These are

- the type of data,
- the type of comparison, and
- the type of performance measure.

When examining the various combinations of these elements, we find that there is a set of testing principles that can be applied uniformly. However, the statistical formulae that need to be used change as the situation changes.

To be responsive to the Commission, we have divided our discussion into four sections and five appendices. The contents of each of these are briefly mentioned below -- first for the main report and then for the extensive supporting appendix materials.

For the main report, this section (Section I) introduces our work and sets out the required scope. The next two sections (Sections II and III) discuss the type of comparisons that need to be identified, and the appropriate testing principles. The final section (Section IV) provides an overview of appropriate testing methodologies, based on what we have learned from our examination of BellSouth's performance measure data in Louisiana.

The five appendices provide technical details on the statistical calculations involved in the Truncated Z statistic (Appendix A), the implementation of the methodology for the trunk blocking performance measure (Appendix B), the calculations involved in computing the balancing critical value of a test (Appendix C), ways to present the results using detailed statistical displays so that results can be audited (Appendix D), and the technical details involved in data trimming (Appendix E).

2. Data Considerations, Comparisons, and Measurement Types

This section makes general distinctions which apply to the performance measures. These distinctions will be important in the determination of appropriate methodologies.

Data Set Types. The type of statistical methodology used depends on the form of the data available. In general, there are two ways to classify the data used for performance measure comparisons. These are:

- transaction level data, and
- aggregated summaries.

Records in a transaction level data set represent a single transaction, e.g. an individual customer order, or the record of a specific trouble reported by a customer. This type of data set allows for deep like-to-like comparisons, and may also allow one to identify the root cause of a problem. A testing methodology needs to be carefully chosen so that it incorporates the comparison levels and does not cover up problem areas.

Records in an aggregated summary data set are typically summaries of related transactions. For example, the total number of blocked calls in a trunk group during the noon hour of a day is a summary statistic. This type of data set may not contain as much information as a transaction level data set, and it therefore needs to be treated differently. While a general methodology may be determined for a transaction level data set, it may not be possible to do so for aggregated summaries. Testing methodology needs to be developed on a case-by-case basis.

Comparison Types. An ILEC's performance in providing services to CLEC customers is tested in one of two ways:

- by comparing CLEC performance to ILEC performance when a retail analog exists, or
- by comparing CLEC performance to a benchmark.

The testing methodologies for these two situations will have similarities, but there are differences that need to be understood.

Table 1 categorizes those performance measures that E&Y has examined by data type and comparison type. The table shows that five performance measures with retail analogs have transaction level data, while three others with retail analogs only have summary level data. No performance measures using benchmarks have been studied.

Table 1. Classification of Performance Measures by Data and Comparison Type (only measures previously examined by E&Y are included)

Level of Data	Comparison Type	
	Retail Analog	**Benchmark**
Transaction Level	Order Completion Interval Maintenance Average Duration % Missed Installations % Missed Repair Trouble Report Rate	No Measures Examined
Summary Level	Billing Timeliness OSS Response Interval Trunk Blocking	No Measures Examined

Measurement Types. The performance measures that will undergo testing are of three types: means, proportions (an average of a measure that takes on only the values of 0 or 1), and rates.

While all three have similar characteristics, proportions and rates are derived from count data while a mean is an average of interval measurements. Table 2 classifies the performance measures by the type of measurement.

Table 2: Classification of Performance Measures by Measurement Type

Mean	Proportion	Rate
Order Completion Interval Maint. Ave. Duration OSS Response Interval	Percent Missed Installations Percent Missed Repairs Billing Timeliness Trunk Blocking	Trouble Report Rate

3. Testing Principles

This section describes five general principles which the final methodology should satisfy:

1. *When possible, data should be compared at appropriate levels, e.g. wire center, time of month, dispatched, residential, new orders.*
2. *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. *The decision system must be developed so that it does not require intermediate manual intervention.*
4. *The testing methodology should balance Type I and Type II Error probabilities.*
5. *Trimming of extreme observations from BellSouth and CLEC distributions is needed in order to ensure that a fair comparison is made between performance measures.*

Like-to-Like Comparisons. *When possible, data should be compared at appropriate levels, e.g. wire center, time of month, dispatched, residential, new orders.*

In particular, to meet this goal the testing process should:

- Identify variables that may affect the performance measure.
- Record important confounding covariates.
- Adjust for the observed covariates in order to remove potential biases and to make the CLEC and the ILEC units as comparable as possible.

It is a well know principle that comparisons should be made on equal footing: apples-to-apples, oranges-to-oranges. Statistical techniques that are addressed in most text books usually assume that this is the case beforehand. Some higher level books address the issue of "designed experiments" and discuss appropriate ways to structure the data collection method so that the text books' formulae can be used in analyzing the data.

Performance measure testing does not involve data from a designed experiment. Rather, the data is obtained from an observational study. That being the case, one must impose a structure on the data after it is gathered in order to assure that fair comparisons are being made. For example, it is important to disaggregate the data to a fine level so that appropriate like-to-like comparisons of CLEC and ILEC data can be made. Any statistical methodology that ignores important confounding variables can produce biased results.

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.*

To achieve this goal, the aggregate test statistic should have the following properties:

- The method should provide a single overall index, on a standard scale.

- If entries in comparison cells are exactly proportional over a covariate, the aggregated index should be very nearly the same as if comparisons on the covariate had not been done.
- The contribution of each comparison cell should depend on the number of observations in the cell.
- Cancellation between comparison cells should be limited, i.e., positive outcomes should not be allowed to cancel negative ones.
- The index should be a continuous function of the observations.

Since the data are being disaggregated to a very deep level, thousands of like-to-like comparison cells are created. It would be an extremely laborious task for a decision maker to sort through individual test results for each cell and determine if discrimination exists. An aggregate summary statistic is needed in order to make an overall judgment.

The aggregate level statistic should be insensitive to small changes in cells values, and its value should not be affected if some of the disaggregation for like-to-like cells is truly unnecessary. Furthermore, individual cell results should be weighted so that those cells with more transactions have larger effects on the overall result.

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

Two statistical paradigms are possible for examining performance measure data. In the exploratory paradigm, data are examined and methodology is developed that is consistent with what is found. In a production paradigm a methodology is decided upon before data exploration. For the production paradigm to succeed

- Calculations should be well defined for possible eventualities.
- The decision process should be based on an algorithm that needs no manual intervention.
- Results should be arrived at in a timely manner.
- The system must recognize that resources are needed for other performance measure-related processes that also must be run in a timely manner.
- The system should be both auditable and adjustable over time.

While the exploratory paradigm provides protection against using erroneous data, it requires a great deal of lead time and is unsuitable for timely monthly performance measure testing. A production paradigm will not only promptly produce overall test results but will also provide documentation that can be used to explore the data after the test results are released.

Error Probability Balancing. *The testing methodology should balance Type I and Type II Error probabilities.*

Specifically, what is required to achieve this goal is

- The probability of a Type I error should equal the probability of a Type II error for well-defined null and alternative hypotheses.
- The formula for a test's balancing critical value should be simple enough to calculate using standard mathematical functions, i.e. one should avoid methods that require computationally intensive techniques.
- Little to no information beyond the null hypothesis, the alternative hypothesis, and the number of observations should be required for calculating the balancing critical value.

The objective of a statistical test is to test a hypothesis concerning the values of one or more population parameters. Usually an inquiry into whether or not there is evidence to support a hypothesis, called the *alternative hypothesis*, is conducted by seeking statistical evidence that the converse of the alternative, the *null hypothesis*, is most likely false. If there is not sufficient evidence to reject the null hypothesis, then a case for accepting the alternative has not been made.

Two types of errors are possible in any decision-making process. These have been summarized in Table 3.

Table 3: Statistical Testing Errors

Decision Error	General Description	In terms of Performance Measure Testing
Type I	Rejecting the null hypothesis (accepting the alternative) when the null is true.	Deciding that BST favors its own customers when it does not.
Type II	Accepting the null hypothesis when the alternative is true.	Deciding that BST does not favor its own customers when it does.

In a controlled experimental study where the sample sizes are relatively small, it is generally desirable to control the Type I error closely to avoid making a conclusion that there is a difference when, in fact, there is none. The probability of a Type II error is not directly controlled but is determined by the sample size and the distance between the null and the alternative hypotheses. Thus, there is some kind of balance between Type I and Type II errors with Type I error usually controlled more closely.

If a standard of materiality is set by stating a specific alternative for the test, and the distribution of the test statistic under both the null and alternative hypotheses is understood, then a critical value can be determined so that the two error probabilities are equal.

Trimming. *Trimming of extreme observations from BellSouth and CLEC distributions is needed in order to ensure that a fair comparison is made between performance measures.*

Three conditions are needed to accomplish this goal. These are:

- Trimming should be based on a general rule that can be used in a production setting.
- Trimmed observations should not simply be discarded; they need to be examined and possibly used in the final decision making process.
- Trimming should only be used on performance measures that are sensitive to "outliers."

For the purpose of performance measure testing, trimming refers to removing transactions that significantly distort the performance measure statistic for the set of transactions under consideration. For example, the arithmetic average (or mean) is extremely sensitive to "outliers" since a single large value can significantly distort the average.

The term "outliers" refers to:

1) extreme data values that may be valid, but since they are rare measurements, they may be considered to be statistically unique; or
2) large values that should not be in the analysis data set because of errors in the measurement or in selecting the data.

Trimming is beneficial since it puts both ILEC and CLEC transactions on equal footing with respect to the largest value in each set. Note, though, that it is only needed for performance measures that are distorted by outliers. Of the three types of measures defined in Section 2, only mean (average) measures require trimming. Appendix E sets forth a trimming plan for mean performance measures.

4. Testing Methodology

This section details the testing methodology that is most appropriate for the various types of performance measures. First, transaction level testing will be discussed when there is a retail analog. Next, transaction level testing against a benchmark. Then, testing when only aggregated summaries are available.

Transaction Level - Retail Analog: The Truncated Z Statistic. When a retail analog is available CLEC performance can be directly compared with ILEC performance. Over the last year, for transaction level data, many test statistics have been examined. We now believe that the "Truncated Z" test statistic provides the best compromise with respect to possessing the desired qualities outlined in Section 3, above.

The Truncated Z is fully described in Appendix A, and formulae for calculation of a balancing critical value are found in Appendix C. The main features of this statistic are:

- A basic test statistic is calculated within each comparison cell.
- The value of a cell's result is left "as is" if the result suggests that "favoritism" may be taking place. Otherwise, the result is set to zero. This is called the truncation step.
- Weights that depend on the volume of both ILEC and CLEC transactions within the cell are determined, and a weighted sum of the "truncated" cell results is calculated.
- The weighted sum is theoretically corrected to account for the truncation, and a final overall statistic is determined.
- This overall test value is compared to a balancing critical value to determine if favoritism is likely.

The test statistic itself is based on like-to-like comparisons, and it possesses all five of the properties of an aggregate test statistic (Section 3). While the test requires a large amount of calculations, our studies of the process on some of BellSouth's performance measure data indicate that the calculations can be completed in a reasonable amount of time. Therefore, the process can be put into production mode. Finally, since a balancing critical value can be calculated, it is possible to balance the error probabilities.

Transaction Level - Benchmark. When a benchmark is used, CLEC performance is not compared with ILEC performance. Like-to-like comparison cells are not needed, thus greatly simplifying the testing process. Statistical testing can be done using a probability model, or non-statistical testing can be done using a deterministic model. No data for this data/comparison class has been studied at this point in time.

If one wants a method that is independent of the number of transactions, then statistical methods should be used to determine if observed performance below the benchmark is statistically significant. Once again, we want a procedure that adheres to the principles outlined in Section 3.

Aggregated Summary - Retail Analog or Benchmark. We cannot provide any one single set of rules for the analysis of data in this class. Data that is an aggregated summary of transactions may or may not present problems. For example, BellSouth's trunk blocking data is saved as summaries by hour of the day. Collectively, the summaries provide sufficient information to proceed with the Truncated Z methodology.

On the other hand, our examination of the data for the OSS response interval revealed that information necessary for computing a Truncated Z was not available. In this case, however, we were able to construct a satisfactory time series method to analyze the measure.

Each measure falling into this class needs to be handled on a case-by-case basis. If sufficient information is available to use the Truncated Z method, then we feel it should

be used. When the Truncated Z cannot be used, a testing methodology that adheres closely to the principles outlined in Section 3 should be determined and followed.

Appendix A. The Truncated Z Statistic

The Truncated Z test statistic was developed by Dr. Mallows in order to have an aggregate level test when transaction level data are available that

- provides a single overall index on a standard scale;
- is robust with respect to unnecessary disaggregation,
- incorporates the number of observations in a cell into the determination of the weight for the contribution of each comparison cell,
- limits the amount of "neutralization" between comparison cells, and
- is a continuous function of the observations.

The Ernst & Young statistical team and Dr. Mallows have studied the implementation of the statistic using some of BellSouth's performance measure data. This has resulted in an overall process for comparing CLEC an ILEC performance such that the following principles hold:

1) Like-to-Like Comparisons are made. (See Appendix B for an example based on the trunk blocking measure.)
2) Error probabilities are balanced. (See Appendix C)
3) Extreme values are trimmed from the data sets when they significantly distort the performance measure statistic. (See Appendix E)
4) The testing process is an automated production system. (Discussed here. See Appendix D for reporting guidelines.)
5) The determination of ILEC favoritism is based on a single aggregate level test statistic. (Discussed here.)

This appendix provides the details behind computing the Truncated Z test statistic so that principles 4 and 5 hold. We start by assuming that any necessary trimming of the data is complete, and that the data are disaggregated so that comparisons are made within appropriate classes or adjustment cells that define "like" observations.

Notation and Exact Testing Distributions

Below, we have detailed the basic notation for the construction of the truncated z statistic. In what follows the word "cell" should be taken to mean a like-to-like comparison cell that has both one (or more) ILEC observation and one (or more) CLEC observation.

L = the total number of occupied cells

j = 1,…,L; an index for the cells

n_{1j} = the number of ILEC transactions in cell j

n_{2j} = the number of CLEC transactions in cell j

n_j = the total number transactions in cell j; $n_{1j} + n_{2j}$

X_{1jk} = individual ILEC transactions in cell j; $k = 1,\ldots, n_{1j}$

X_{2jk} = individual CLEC transactions in cell j; $k = 1,\ldots, n_{2j}$

Y_{jk} = individual transaction (both ILEC and CLEC) in cell j

$$= \begin{cases} X_{1jk} & k = 1,\ldots,n_{1j} \\ X_{2jk} & k = n_{1j}+1,\ldots,n_j \end{cases}$$

$\Phi^{-1}(\cdot)$ = the inverse of the cumulative standard normal distribution function

For Mean Performance Measures the following additional notation is needed.

$\overline{X}_{1j}$ = the ILEC sample mean of cell j

$\overline{X}_{2j}$ = the CLEC sample mean of cell j

s^2_{1j} = the ILEC sample variance in cell j

s^2_{2j} = the CLEC sample variance in cell j

y_{jk} = a random sample of size n_{2j} from the set of $Y_{j1},\ldots,Y_{jn_j}$; $k = 1,\ldots,n_{2j}$

M_j = the total number of distinct pairs of samples of size n_{1j} and n_{2j};

$$= \binom{n_j}{n_{1j}}$$

The exact parity test is the permutation test based on the "modified Z" statistic. For large samples, we can avoid permutation calculations since this statistic will be normal (or Student's t) to a good approximation. For small samples, where we cannot avoid permutation calculations, we have found that the difference between "modified Z" and the textbook "pooled Z" is negligible. We therefore propose to use the permutation test based on pooled Z for small samples. This decision speeds up the permutation computations considerably, because for each permutation we need only compute the sum of the CLEC sample values, and not the pooled statistic itself.

A permutation probability mass function distribution for cell j, based on the "pooled Z" can be written as

$$PM(t) = P(\sum_k y_{jk} = t) = \frac{\textit{the number of samples that sum to } t}{M_j},$$

and the corresponding cumulative permutation distribution is

$$CPM(t) = P(\sum_k y_{jk} \le t) = \frac{\textit{the number of samples with sum} \le t}{M_j}.$$

For Proportion Performance Measures the following notation is defined

a_{1j} = the number of ILEC cases possessing an attribute of interest in cell j

a_{2j} = the number of CLEC cases possessing an attribute of interest in cell j

a_j = the number of cases possessing an attribute of interest in cell j; $a_{1j} + a_{2j}$

The exact distribution for a parity test is the hypergeometric distribution. The hypergeometric probability mass function distribution for cell j is

$$HG(h) = P(H = h) = \begin{cases} \dfrac{\binom{n_{1j}}{h}\binom{n_{2j}}{a_j - h}}{\binom{n_j}{a_j}}, & \max(0, a_j - n_{2j}) \le h \le \min(a_j, n_{1j}) \\ 0 & \text{otherwise} \end{cases},$$

and the cumulative hypergeometric distribution is

$$CHG(x) = P(H \le x) = \begin{cases} 0 & x < \max(0, a_j - n_{1j}) \\ \sum_{h=\max(0,a_j-n_{1j})}^{x} HG(h), & \max(0, a_j - n_{1j}) \le x \le \min(a_j, n_{2j}) \\ 1 & x > \min(a_j, n_{2j}) \end{cases}.$$

For Rate Measures, the notation needed is defined as

b_{1j} = the number of ILEC base elements in cell j

b_{2j} = the number of CLEC base elements in cell j

b_j = the total number of base elements in cell j; $b_{1j} + b_{2j}$

p_{1j} = the ILEC sample rate of cell j; n_{1j}/b_{1j}

p_{2j} = the CLEC sample rate of cell j; n_{2j}/b_{2j}

q_j = the relative proportion of CLEC elements for cell j; b_{2j}/b_j

The exact distribution for a parity test is the binomial distribution. The binomial probability mass function distribution for cell j is

$$BN(k) = P(B = k) = \begin{cases} \binom{n_j}{k} q_j^k (1-q_j)^{n_j-k}, & 0 \le k \le n_j \\ 0 & \text{otherwise} \end{cases},$$

and the cumulative binomial distribution is

$$CBN(x) = P(B \le x) = \begin{cases} 0 & x < 0 \\ \sum_{k=0}^{x} BN(k), & 0 \le x \le n_j \\ 1 & x > n_j \end{cases}.$$

Calculating the Truncated Z

The general methodology for calculating an aggregate level test statistic is outlined below.

1. **Calculate cell weights,** W_j. A weight based on the number of transactions is used so that a cell which has a larger number of transactions has a larger weight. The actual weight formulae will depend on the type of measure.

Mean Measure

$$W_j = \sqrt{\frac{n_{1j} n_{2j}}{n_j}}$$

Proportion Measure

$$W_j = \sqrt{\frac{n_{2j} n_{1j}}{n_j} \cdot \frac{a_j}{n_j} \cdot \left(1 - \frac{a_j}{n_j}\right)}$$

Rate Measure

$$W_j = \sqrt{\frac{b_{1j} b_{2j}}{b_1} \cdot \frac{n_j}{b_j}}$$

2. **In each cell, calculate a Z value,** Z_j. A standard normal Z statistic is needed for each cell.

   - If $W_j = 0$, set $Z_j = 0$.
   - When the cell sample sizes are sufficiently large, formulae based on a normal approximation can be used.
   - If cell sample sizes are not large enough for a normal approximation to hold, then exact testing methods must be employed. When this occurs, the results of the test statistic are converted into an equivalent value from the standard normal distribution.

The actual Z statistic calculation depends on the type of performance measure.

Mean Measure

$$Z_j = \Phi^{-1}(\alpha)$$

where α is determine by the following algorithm.

If $\min(n_{1j}, n_{2j}) > 6$, then determine α as

$$\alpha = P(t_{n_{1j}-1} \le T_j),$$

that is, α is the probability that a t random variable,

$$t_j = \frac{\bar{X}_{1j} - \bar{X}_{2j}}{s_{1j}\sqrt{\frac{1}{n_{1j}} + \frac{1}{n_{2j}}}}$$

with n_{1j} - 1 degrees of freedom, is less than T_j.

$$T_j = t_j + \frac{g}{6}\left(\frac{n_{1j} + 2n_{2j}}{\sqrt{n_{1j}\, n_{2j}(n_{1j} + n_{2j})}}\right)\left(t^2 + \frac{n_{2j} - n_{1j}}{2n_{1j} + n_{2j}}\right).$$

Here the coefficient g is an estimate of the skewness of the parent population, which we assume is the same in all cells. It can be estimated from the ILEC values in the largest cells. This needs to be done only once for each measure. We have found that attempting to estimate this skewness parameter for each cell separately leads to excessive variability in the "adjusted" t. We therefore use a single compromise value in all cells.

Note, that t_j is the "modified Z" statistic. The statistic T_j is a "modified Z" corrected for the skewness of the ILEC data.

Appendix B. Trunk Blocking

This Appendix describes how the trunk blocking data can be processed to apply the Truncated Z Statistic. Trunk blocking is defined as the proportion of blocked calls a trunk group experiences in a time interval. It is a ratio of two numbers—blocked and attempted calls, both of which can vary over time and across trunk groups. Since the measure is a proportion where the numerator is a subset of the denominator, the truncated Z statistic, modified for proportions, can be applied here (see Appendix A).

As with other performance measures, data are first assigned to like-to-like cells, and the Z statistic is then computed within each cell. For trunk blocking, cells are defined by three variables: hour, day, and trunk group size or capacity. The next sections will describe the data and the data processing steps in greater detail.

Data Sources

Two data files are processed for the trunk blocking measure. One is the Trunk Group Data File that contains the Trunk Group Serial Number (TGSN), Common Language Location Identifier (CLLI) , and other characteristics needed to categorize trunk groups and to identify them as BellSouth or CLEC.

The other file is the Blocking Data File (BDF), which contains the actual 24 hour blocking ratios for each weekday. There are 4 or 5 weeks in a monthly report cycle. The current system, however, allows the storage of daily blocking data by hour for a week only. Therefore, the data elements necessary to compute the Truncated Z must be extracted each week.

Two important data fields of interest on the Blocking Data File are the Blocking Ratio and Offered Load. The basic definition of Blocking Ratio is the proportion of all attempted calls that were blocked. For the simplest case of one way trunk groups, this is computed by dividing the number of blocked calls by the total call attempts, given that the data are valid. If they are not valid (e.g., actual usage exceeds capacity), blocking is estimated via the Neal Wilkinson algorithm.

Although the raw data--blocked calls (overflow) and peg counts (total call attempts)--are available, the calculation of the Blocking Ratio may be complicated for two-way trunk groups and trunk groups with invalid data. For this reason, we use the blocking ratios from the BDF instead of computing the ratios from the raw data. In order to reflect different call volumes processed through each trunk group, however, the blocking ratios need to be either weighted by call volume or converted to blocked and attempted calls before they are aggregated.

The measure of call traffic volume recommended for weighting is Offered Load. Offered Load is different from call counts in that it incorporates call duration as well. Since it is

not just the number of calls but the total usage—number of calls multiplied by average call duration--that determines the occurrence of any blocking, this pseudo measure, Offered Load, appears to be the best indicator of call volume.

Cells or comparison classes are determined by three factors—hour, day, and trunk group capacity (number of trunks in service). The first two factors represent natural classes because trunk blocking changes over time. The third factor is based on our finding that high blocking tends to occur in small trunk groups. A pattern was found not only in the magnitude of blocking but also in its variability. Both the magnitude and variability of blocking decrease as trunk group capacity increases. Additional work is needed to establish the appropriate number of capacity levels and the proper location of boundaries.

Data Processing

The data are processed using the five steps below:

1. Merge the two files by TGSN and select only trunk groups listed in both files.
2. Reset the blocking of all high use trunk groups to zero[1].
3. Assign trunk group categories to CLEC and BellSouth: Categories 1, 3, 4, 5, 10, and 16 for CLEC and 9 for BellSouth[2]. The categories used here for comparison are:

Category	Administrator	Point A	Point B
1	BellSouth	BellSouth End Office	BellSouth Access Tandem
3	BellSouth	BellSouth End Office	CLEC Switch
4	BellSouth	BellSouth Local Tandem	CLEC Switch
5	BellSouth	BellSouth Access Tandem	CLEC Switch
9	BellSouth	BellSouth End Office	BellSouth End Office
10	BellSouth	BellSouth End Office	BellSouth Local Tandem
16	BellSouth	BellSouth Tandem	BellSouth Tandem

4. Recode the missing data. The Blocking Data File assigns all missing data (no valid measurement data) zero blocking. To differentiate true zero blocking from zeroes due to missing data, invalid records were identified and the ratios reset to missing. The blocking value was invalid if both the number of Loaded Days and the Offered Load were 0 for a given hourly period.
5. Form comparison classes based either on the data (i.e., quartiles) or on a predetermined set of values.

[1] The high use trunk groups cannot have any blocking. These are set up such that all overflow calls are automatically routed to other trunk groups instead of being physically blocked.
[2] More detailed information on all categories is described in a report 'Trunk Performance Report Generation' by Ernst & Young (March 1999).

Calculation of the Proportion of Blocked Calls

Each cell is determined by day of the month, hour of the day, and trunk group capacity. To use the Truncated Z method, we generate summary information, to include the total number of blocked calls and the total number of attempted calls, for each cell.

For the details of each calculation step, the following notation is used. For a given hour of a day, let $\overline{X}_{1ij}$ be the proportion of BellSouth blocked calls for trunk group i in cell j and $\overline{X}_{2ij}$ be the corresponding proportion for CLEC. Then $\overline{X}_{1ij} = X_{1ij} / n_{1ij}$ where X_{1ij} denotes the number of BellSouth blocked calls and n_{1ij} denotes the number of BellSouth total call attempts (indicated by Offered Load) for trunk group i in cell j. Likewise, $\overline{X}_{2ij} = X_{2ij} / n_{2ij}$. For the steps outlined below, only the CLEC notation is provided.

1. Compute the number of blocked calls for trunk group i: $X_{2ij} = \overline{X}_{2ij} * n_{2ij}$
2. Compute total call attempts for all trunk groups in the cell: $n_{2j} = \sum_i n_{2ij}$
3. Compute mean blocking proportion for cell j: $\overline{X}_{2j} = \sum_i X_{2ij} \Big/ \sum_i n_{2ij}$
4. Compute the total number of BellSouth and CLEC blocked calls in cell j: $t_j = \sum_i X_{1ij} + \sum_i X_{2ij}$
5. Apply the Truncated Z Statistic for Proportion measures presented in Appendix A.

Appendix C
Balancing the Type I and Type II Error Probabilities of the Truncated Z Test Statistic

This appendix describes a the methodology for balancing the error probabilities when the Truncated Z statistic, described in Appendix A, is used for performance measure parity testing. There are four key elements of the statistical testing process:

1. the null hypothesis, H_0, that parity exists between ILEC and CLEC services
2. the alternative hypothesis, H_a, that the ILEC is giving better service to its own customers
3. the Truncated Z test statistic, Z^T, and
4. a critical value, c

The decision rule[1] is

- If $Z^T < c$ then accept H_a.
- If $Z^T \geq c$ then accept H_0.

There are two types of error possible when using such a decision rule:

Type I Error: Deciding favoritism exists when there is, in fact, no favoritism.
Type II Error: Deciding parity exists when there is, in fact, favoritism.

The probabilities of each type of each are:

Type I Error: $\alpha = P(Z^T < c \mid H_0)$.
Type II Error: $\beta = P(Z^T \geq c \mid H_a)$.

In what follows, we show how to find a balancing critical value, c_B, so that $\alpha = \beta$.

General Methodology

The general form of the test statistic that is being used is

[1] This decision rule assumes that the smaller a performance measure is, the better the service. If the opposite is true, then reverse the decision rule.

$$z_0 = \frac{\hat{T} - E(\hat{T}|H_0)}{SE(\hat{T}|H_0)}, \tag{C.1}$$

where

$\hat{T}$ is an estimator that is (approximately) normally distributed,

$E(\hat{T} \mid H_0)$ is the expected value (mean) of $\hat{T}$ under the null hypothesis, and

$SE(\hat{T} \mid H_0)$ is the standard error of $\hat{T}$ under the null hypothesis.

Thus, under the null hypothesis, z_0 follows a standard normal distribution. However, this is not true under the alternative hypothesis. In this case,

$$z_a = \frac{\hat{T} - E(\hat{T}|H_a)}{SE(\hat{T}|H_a)}$$

has a standard normal distribution. Here

$E(\hat{T} \mid H_a)$ is the expected value (mean) of $\hat{T}$ under the alternative hypothesis, and

$SE(\hat{T} \mid H_a)$ is the standard error of $\hat{T}$ under the alternative hypothesis.

Notice that

$$\begin{aligned} \beta &= P(z_0 > c \mid H_a) \\ &= P\left(z_a > \frac{cSE(\hat{T} \mid H_0) + E(\hat{T} \mid H_0) - E(\hat{T} \mid H_a)}{SE(\hat{T} \mid H_a)} \right) \end{aligned} \tag{C.2}$$

and recall that for a standard normal random variable z and a constant b, $P(z < b) = P(z > -b)$. Thus,

$$\alpha = P(z_0 < c) = P(z_0 > -c) \tag{C.3}$$

Since we want $\alpha = \beta$, the right hand sides of (C.2) and (C.3) represent the same area under the standard normal density. Therefore, it must be the case that

$$-c = \frac{cSE(\hat{T} \mid H_0) + E(\hat{T} \mid H_0) - E(\hat{T} \mid H_a)}{SE(\hat{T} \mid H_a)}.$$

Solving this for c give the general formula for a balancing critical value:

$$c_B = \frac{E(\hat{T}\,|\,H_a) - E(\hat{T}\,|\,H_0)}{SE(\hat{T}\,|\,H_a) + SE(\hat{T}\,|\,H_0)} \tag{C.4}$$

The Balancing Critical Value of the Truncated Z

In Appendix A, the Truncated Z statistic is defined as

$$Z^T = \frac{\sum_j W_j Z_j^* - \sum_j W_j E(Z_j^*|H_0)}{\sqrt{\sum_j W_j^2 \mathrm{Var}(Z_j^*|H_0)}}$$

In terms of equation (C.1) we have

$$\hat{T} = \sum_j W_j Z_j^*$$

$$E(\hat{T}\,|\,H_0) = \sum_j W_j E(Z_j^*|H_0)$$

$$SE(\hat{T}\,|\,H_0) = \sqrt{\sum_j W_j^2 \mathrm{Var}(Z_j^*|H_0)}$$

To compute the balancing critical value (C.4), we also need $E(\hat{T}|H_a)$ and $SE(\hat{T}|H_a)$. These values are determined by

$$E(\hat{T}|H_a) = \sum_j W_j E(Z_j^*|H_a), \text{ and}$$

$$SE(\hat{T}|H_a) = \sqrt{\sum_j W_j^2 \,\mathrm{var}(Z_j^*|H_a)}\,.$$

In which case equation (C.4) gives

$$c_B = \frac{\sum_j W_j E(Z_j^*|H_a) - \sum_j W_j E(Z_j^*|H_0)}{\sqrt{\sum_j W_j^2 \,\mathrm{var}(Z_j^*|H_a)} + \sqrt{\sum_j W_j^2 \,\mathrm{var}(Z_j^*|H_0)}}\,. \tag{C.5}$$

Thus, we need to determine how to calculate $E(Z_j^*|H_0)$, $\mathrm{Var}(Z_j^*|H_0)$, $E(Z_j^*|H_a)$, and $\mathrm{Var}(Z_j^*|H_a)$. These values depend on the distribution of Z_j (see Appendix A) under the null and alternative hypotheses.

One possible set of hypotheses, that take into account the assumption that transaction are identically distributed within cells, is:

H_0: $\mu_{1j} = \mu_{2j}$, $\sigma_{1j}^2 = \sigma_{2j}^2$

H_a: $\mu_{2j} = \mu_{1j} + \delta_j \cdot \sigma_{1j}$, $\sigma_{2j}^2 = \lambda_j \cdot \sigma_{1j}^2$ $\quad \delta_j > 0$, $\lambda_j \geq 1$ and $j = 1,\ldots,L$.

Under this null hypothesis, Z_j has a standard normal distribution within each cell j. In which case,

$$E(Z_j^* \mid H_0) = -\frac{1}{\sqrt{2\pi}}, \text{ and}$$

$$\text{var}(Z_j^* \mid H_0) = \frac{1}{2} - \frac{1}{2\pi}.$$

Under the alternative hypothesis, Z_j has a normal distribution with

$$E(Z_j \mid H_a) = m_j = \frac{-\delta_j}{\sqrt{\frac{1}{n_{1j}} + \frac{1}{n_{2j}}}}, \text{ and}$$

$$SE(Z_j \mid H_a) = se_j = \sqrt{\frac{\lambda_j n_{1j} + n_{2j}}{n_{1j} + n_{2j}}}$$

In general, the mean of a normal distribution truncated at 0 is

$$M(\mu,\sigma) = \int_{-\infty}^{0} \frac{x}{\sqrt{2\pi}\sigma} \exp(-\tfrac{1}{2}\left(\tfrac{x-\mu}{\sigma}\right)^2) dx,$$

and the variance is

$$V(\mu,\sigma) = \int_{-\infty}^{0} \frac{x^2}{\sqrt{2\pi}\sigma} \exp(-\tfrac{1}{2}\left(\tfrac{x-\mu}{\sigma}\right)^2) dx - M(\mu,\sigma)^2$$

It can be shown that

$$M(\mu,\sigma) = \mu\,\Phi(\tfrac{-\mu}{\sigma}) - \sigma\,\phi(\tfrac{-\mu}{\sigma})$$

and

$$V(\mu,\sigma) = (\mu^2 + \sigma^2)\Phi(\tfrac{-\mu}{\sigma}) - \mu\,\sigma\,\phi(\tfrac{-\mu}{\sigma}) - M(\mu,\sigma)^2$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function, and $\phi(\cdot)$ is the standard normal density function.

Using the above notation, and equation (C.5), we get the formula for the balancing critical of Z^T for the alternative hypothesis defined above.

$$c_B = \frac{\sum_j W_j M(m_j, se_j) - \sum_j W_j \frac{-1}{\sqrt{2\pi}}}{\sqrt{\sum_j W_j^2 V(m_j, se_j)} + \sqrt{\sum_j W_j^2 \left(\frac{1}{2} - \frac{1}{2\pi}\right)}}. \qquad \text{(C.6)}$$

This formula assumes that Z_j, is approximately normally distributed within cell j. When the cell sample sizes, n_{1j} and n_{2j}, are small this may not be true. It is possible to determine the cell mean and variance under the null hypothesis when the cell sample sizes are small. It is much more difficult to determine these values under the alternative hypothesis. Since the cell weight, W_j will also be small (see Appendix A) for a cell with small volume, the cell mean and variance will not contribute much to the weighted sum. Therefore, formula (C.6) should provide a reasonable approximation to the balancing critical value.

Determining the Parameters of the Alternative Hypothesis

In this appendix we have indexed the alternative hypothesis by two sets of parameters, λ_j and δ_j. 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. Specific choices are best left to telephony experts. Still, it is possible to comment on some aspects of these choices:

- Parameter Choices for λ_j. The set of parameters λ_j index alternatives to the null hypothesis that arise because there might be greater unpredictability or variability in the delivery of service to a CLEC customer over that which would be achieved for an otherwise comparable ILEC customer. While concerns about differences in the variability of service are important, it turns out that the truncated Z testing which is being recommended here is relatively insensitive to all but very large values of the λ_j. Put another way, reasonable differences in the values chosen here could make very little difference in the balancing points chosen.

- Parameter Choices for δ_j. The set of parameters δ_j are much more important in the choice of the balancing point than was true for the λ_j. The reason for this is that they directly index differences in average service. The truncated Z test is very sensitive to any such differences; hence, even small disagreements among experts in the choice of the δ_j could be very important. Sample size matters here too. For example, setting all the δ_j to a single value — $\delta_j = \delta$ — might be fine for tests across individual CLECs where currently in Louisiana the CLEC customer bases are not too different. Using the same

value of δ for the overall state testing does not seem sensible, however, since the state sample would be so much larger.

The bottom line here is that beyond a few general considerations, like those given above, a principled approach to the choice of the alternative hypotheses to guard against must come from elsewhere.

Appendix D. Trimming Outliers for Mean Measures

The arithmetic average is extremely sensitive to outliers; a single large value, possibly an erroneous value, can significantly distort the mean value. And by inflating the error variance, this also affects conclusions in the test of hypotheses. Extreme data values may be correct, but since they are rare measurements, they may be considered to be statistical outliers. Or they may be values that should not be in the analysis data set because of errors in the measurement or in selecting the data.

At this time, only two mean measures have been analyzed: Order Completion Interval and Maintenance Average Duration. Maintenance Average Duration data are truncated at 240 hours and therefore this measure was not trimmed further. For Order Completion Interval, the underlying distribution of the observations is clearly not normal, but rather skewed with a very long upper-tail.

A useful technique, coming from the field of robust statistical analysis, is to trim a very small proportion from the tails of the distribution before calculating the means. The resulting mean is referred to as a trimmed mean. Trimming is beneficial in that it speeds the convergence of the distribution of the means to a normal distribution. Only extreme values are trimmed, and in many cases the data being trimmed are, in fact, data that might not be used in the analysis on other grounds.

In the first analysis of the verified Order Completion Interval-Provisioning measure, after removing data that were clearly in error or were not applicable, we looked at the cases that represented the largest 0.01% of the BST distribution. In the August data, this corresponded to orders with completion intervals greater than 99 days. All of these were BellSouth orders. In examining the largest 11 individual examples that would be removed from analysis, we found that only 1 of the 11 cases was a valid case where the completion interval was unusually large. The other 10 cases were examples of cases that should not have been included in the analysis. This indicates that at least in preliminary analysis, it is both beneficial to examine the extreme outliers and reasonable to remove them.

A very slight trimming is needed in order to put the central limit theorem argument on firm ground. But finding a robust rule that can be used in a production setting is difficult. Also, any trimming rule should be fully explained and any observations that are trimmed from the data must be fully documented.

When it is determined that a measure should be trimmed, a trimming rule that is easy to implement in a production setting is:

> **Trim the ILEC observations to the largest CLEC value from all CLEC observations in the month under consideration.**

That is, no CLEC values are removed; all ILEC observations greater than the largest CLEC observation are trimmed.

While this method is simple, it does allow for extreme CLEC observations to be part of the analysis. For instance, suppose that the amount of time to complete an order was less than 40 days for all CLEC orders except one. Let's say that this extreme order took 100 days to complete. The trimming rule says that all ILEC orders above 100 days should be trimmed, but a closer look at the data might suggest trimming at 40 days instead.

Since we are operating in a production mode system, it is not possible to explore the data before the trimming takes place. Other automatic trimming rules present other problems, so our solution is to use the simple trimming rule above, and have the system automatically produce a trimming report that can be examined at a later point in time.

The trimming report should include:

- The value of the trim point.
- Summary statistics and graphics of the ILEC observations that were trimmed.
- A listing of the trimmed ILEC transaction for a random sample of 10 trimmed transactions. This listing should not disclose sensitive information.
- A listing of the 10 most extreme CLEC transactions. This listing should not disclose sensitive information.
- The number of ILEC and CLEC observations above some fixed point, so that changes in the upper tail can be better tracked over time.

The trimming report should be part of the overall report discussed in Appendix D. Examples of tables contained within the trimming report are shown below.

April XXXX
Performance Measure Extreme Values

Cutoff	26	Cutoff	26
# of Records	20,573	# of Records	367,065
10 Largest		Extreme Values	652
Minimum	19	Minimum	27
Median	23	Median	32
Maximum	26	Maximum	283
Subtotal	20,573	Subtotal	366,413

April XXXX
Performance Measure Weighting Report

# of Records	20,573	# of Records	366,413
No Matching BST		No Matching CLEC	
Classification (1)	47	Classification (2)	21,974
Subtotal	20,526	Subtotal	344,439

April XXXX
Perormance Measure Filtering Information

This table displays information about the size of the database files and the cases that were removed from the analysis.

Unfiltered Total	28,691	1999	Unfiltered Total	453,107
Records Removed for Business Reasons *(e.g. not N, T, C, or P orders, not resale and not UNE)*	7,242		Records Removed for Business Reasons *(e.g. not N, T, C, or P orders, not retail)*	78,613
Total Reported on Web Report	21,449		Total Reported on Web Report	374,494
Additional Records Removed for Business Reasons	876		Additional Records Removed for Business Reasons	7,429
Missing Appointment code is 'S'	844		Missing Appointment code is 'S'	7,172
General Class Service = 'O'	0		General Class Service = 'O'	279
UNE Cases	102			
Records Removed for Statistical Reasons			Records Removed for Statistical Reasons	
Extreme Values Removed	0		Extreme Values Removed	652
No Matching Classification Removals	47		No Matching Classification Removals	21,974
FILTERED TOTAL	20,526		FILTERED TOTAL	344,439

CLEC Extreme Values

Wire Center	Time	Dispatch	Residence	Circuits	Order Type	Order Interval
NWORLAMA	1	1	3	1	N	61
OPLSLATL	1	2	1	1	C	53
NWORLAMA	2	1	3	1	N	44
NWORLAMA	1	1	3	1	N	39
BTRGLAWN	1	1	2	1	C	38
LKCHLADT	1	1	1	1	T	37
NWORLAMA	1	1	3	1	N	32
NWORLAMA	2	1	3	1	N	32
SHPTLACL	1	1	2	1	N	28

Frequency of Extreme Values Removed from BST file (Top 10)

Wire Center	Time	Dispatch	Residence	Circuits	Order Type	Frequency
NWORLAMA	1	1	3	1	N	55
NWORLAMA	2	1	3	1	N	25
BTRGLASB	2	1	3	1	C	23
NWORLAMC	2	1	3	1	C	23
NWORLAMC	1	1	3	1	C	22
NWORLAMA	2	1	3	1	C	18
NWORLAMA	1	1	3	1	C	17
BTRGLASB	1	1	3	1	C	16
LEYTLAMA	1	1	3	1	C	15
NWORLAMA	2	2	3	1	C	14

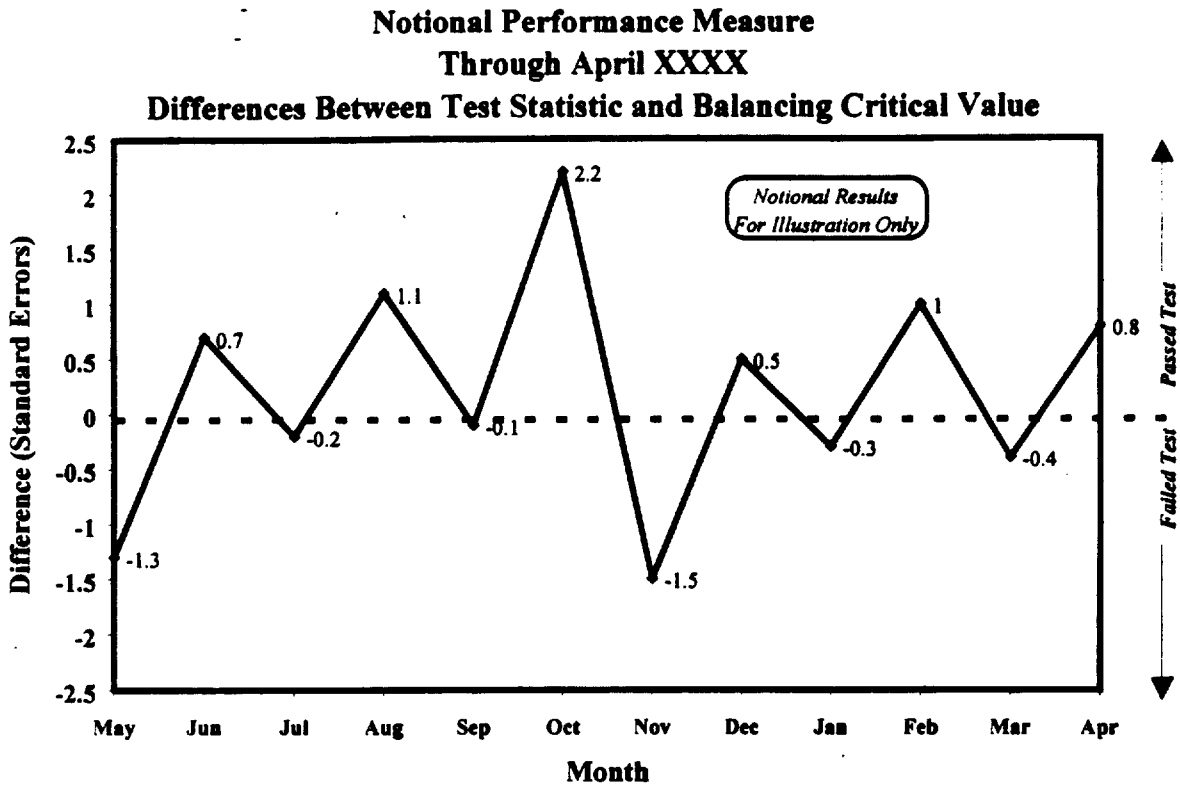
Reporting Guidelines

The general structure for reporting statistical results in a production environment will be the same for the different measures and will consist of three components. For each measure, we will first present the monthly test statistics over a period of time. Following this will be the results for the current month, with summary statistics, test statistics, and descriptive graphs. Finally, the third component of the reporting structure is a summary of any adjustments to the data made in the process of running the tests, mainly including a description of how many records were excluded from analysis and the reason for the exclusion (i.e., excluded due to business rules, or due to statistical/methodological rules pertaining to the measure). This component is important to assure that the reported results can be audited.

Selected components of the reporting structure are illustrated in the samples that follow. An outline of the report is shown below. Monthly results will be presented for each level of aggregation required.

I. Test Statistics Over Time
II. Monthly Results
   A. Summary Statistics
   B. Test Statistics
   C. Descriptive Graphs (Frequency Distributions, etc.)
III. Adjustments to Data
   A. Records Excluded Due to Business Rules
   B. Records Excluded Due to Statistical Rules

Test Statistic Over Time. The first component of the reporting structure is an illustration of the trend of the particular performance measure over time together with a tabular summary of results for the current month. We will show at a glance whether the tests consistently return non-statistically significant results; consistently indicate disparity (be that in favor of BellSouth or in favor of the CLECs); or vary month by month in their results. An example of this component follows.

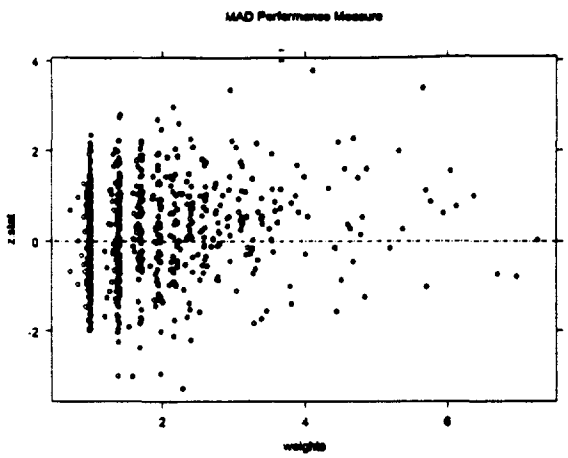


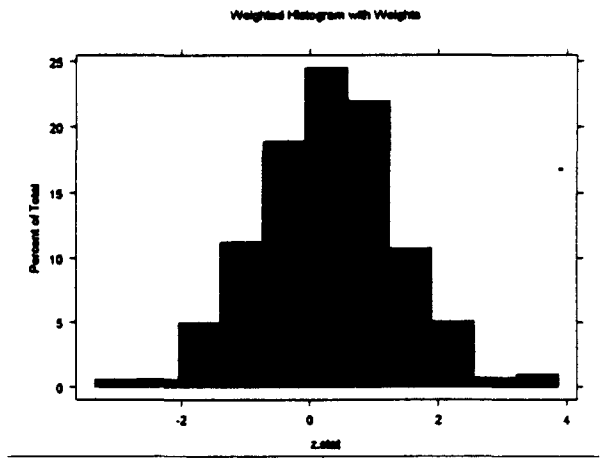
Result for Current Month	
Test Statistic	0.410
Balancing Critical Value	-1.210
Difference	0.800

Monthly Results. The most important component of the reporting structure is the part which presents results of the monthly statistical tests on the given performance measure. The essential aspects included in this component are the summary statistics; the test statistics and results; and descriptive graphs of the results.

It is important to present basic summary statistics to complete the comparison between BellSouth and the CLECs. At a minimum, these statistics will include the means, standard deviations, and population sizes. In addition to basic descriptive statistics, we also present the test statistic results. Examples of ways we have presented these statistics in the past can be found in BellSouth's February 25, 1999 filing before the Louisiana Public Service Commission.

Finally, the results will be presented in graphical format. Below is an example of how to graphically present the data behind the Truncated Z statistic. One graph shows a plot of cell Z score versus cell weights. The other is a histogram of the weighted cell Z scores.





Adjustments to Data. The third important component of the reporting structure is information on any adjustments performed on the data. This information is essential in order that the results may be verified and audited. The most prevalent examples of such modifications would be removal of observations and weighting of the data.

Records can be removed from analysis for both business reasons (these will likely be taken into account in the PMAP system) and for statistical reasons. All of the performance measures exclude certain records based on business rules underlying each measure's particular definitions and methodologies. The number of records excluded for each rule will be summarized. In addition, some of the measures will have observations excluded for statistical reasons, particularly in the case of "mean measures" (OCI and MAD); these exclusions will be summarized as well. The tables below show examples of the current method for summarizing this information:

April XXXX
Perormance Measure Filtering Information

This table displays information about the size of the database files and the cases that were removed from the analysis.

		1999	
Unfiltered Total	**28,691**	**Unfiltered Total**	**453,107**
Records Removed for Business Reasons *(e.g. not N, T, C, or P orders, not resale and not UNE)*	**7,242**	**Records Removed for Business Reasons** *(e.g. not N, T, C, or P orders, not retail)*	**78,613**
Total Reported on Web Report	**21,449**	**Total Reported on Web Report**	**374,494**
Additional Records Removed for Business Reasons	**876**	**Additional Records Removed for Business Reasons**	**7,429**
Missing Appointment code is 'S'	844	Missing Appointment code is 'S'	7,172
General Class Service = 'O'	0	General Class Service = 'O'	279
UNE Cases	102		
Records Removed for Statistical Reasons		**Records Removed for Statistical Reasons**	
Extreme Values Removed	**9**	**Extreme Values Removed**	**652**
No Matching Classification Removals	**47**	**No Matching Classification Removals**	**21,974**
FILTERED TOTAL	**20,517**	**FILTERED TOTAL**	**344,439**