BEFORE THE COMMONWEALTH OF KENTUCKY

PUBLIC SERVICE COMMISSION

REBUTTAL TESTIMONY OF ROBERT M. BELL

ON BEHALF OF

AT&T COMMUNICATIONS OF THE SOUTH CENTRAL STATES, INC.

AND TCG OHIO, INC.

CASE NO. 2001-105

JULY 9, 2001

Q. PLEASE STATE YOUR NAME AND ADDRESS.

A. My name is Robert M. Bell. My business address is AT&T Labs-Research, 180 Park Avenue, Florham Park, New Jersey 07932.

Q. PLEASE DESCRIBE YOUR EDUCATION AND PROFESSIONAL EXPERIENCE AS THEY RELATE TO ISSUES IN THIS PROCEEDING.

A. I received a Ph.D. in Statistics from Stanford University in 1980. From 1980 to 1998, I was promoted to Senior Statistician at RAND, a non-profit institution that conducts public policy analysis. While at RAND, I supervised the design and/or analysis of many projects, including large multi-site evaluations in the fields of preventive dentistry, drug prevention, and depression care. I also headed the RAND Statistics Group from 1993 to 1995 and taught statistics in the RAND Graduate School from 1992 to 1998.

In 1998, I joined the Statistics Research Department at AT&T Labs-Research, where I am a Principal Member of Technical Staff. I have authored or co-authored fifty articles on statistical analysis that have appeared in a variety of refereed, professional journals. I am a fellow of the American Statistical Association. I currently serve on the Panel to Review the 2000 Census organized by the National Academy of Sciences.

Q. HAVE YOU TESTIFIED IN OTHER REGULATORY PROCEEDINGS IN THE PAST?

A. The proceedings in which I have testified are listed on my Curriculum Vitae, which is attached as Exhibit RMB-1.

Q. WHAT IS THE PURPOSE OF YOUR TESTIMONY IN THIS PROCEEDING?

A. BellSouth urges the Commission to draw conclusions regarding the readiness of BellSouth's OSS to support local competition based upon the Georgia Performance Measurement plan and the *BellSouth Telecommunications, Inc. OSS Evaluation – Georgia, Master Test Plan Final Report* ("Final Report") on the third-party test. BellSouth has proposed that the Commission use the Georgia Performance Measurement plan as the basis for its review of BellSouth's performance in Kentucky. The purpose of this testimony is to discuss statistical issues related to the Georgia Performance Measurement plan and the third-party test. I describe several problems with the data, analysis, and conclusions reported in the third-party test Final Report. In particular, I explain that there is insufficient evidence to support many of the conclusions reached by KPMG Consulting, Inc. ("KCI"). For these reasons, I conclude that, to the extent

BellSouth urges this Commission to rely upon the third-party test and the Georgia data, such reliance is unwarranted and cannot serve as the basis for statistically sound conclusions.

I also comment on issues relating to BellSouth's proposed performance measures plan generally. I describe problems that can arise with the truncated-z methodology used to aggregate results of like-to-like comparisons in BellSouth's proposed Performance Measurements plan. I explain why AT&T believes that the value of delta used to compute balancing critical values in the plan should be 0.25 or lower for both Tier I and Tier II tests. I also explain why BellSouth's calculation of "affected volume," which is central to the remedy calculations, is inappropriate.

I. <u>KCI'S GEORGIA THIRD-PARTY TEST</u>

Q. PLEASE SUMMARIZE YOUR CONCLUSIONS REGARDING THE THIRD-PARTY TEST.

A. Based on my review of the Final Report, I have concluded the following:

- KCI should not have applied a statistical method known as "P-value" analysis to tests involving Commission-established benchmarks. This process rewarded BellSouth with a "satisfied" rating even though BellSouth failed to meet the benchmark.

- In cases where it may have been appropriate to apply some sort of statistical analysis to account for random variation in results, KCI applied a P-value analysis to account for random "bad" results (to reach a "satisfied" result) but did not use a statistical analysis to rule out random "good" results.

- KCI's "satisfied" determinations based on aggregated rather than disaggregated service types masked poor performance that otherwise would have led to a conclusion of "not satisfied."

- KCI's practice of using its "professional judgment" to change a finding of "not satisfied" to "satisfied" is unusual and questionable.

- KCI should have implemented military style testing in a manner that revealed BellSouth's true performance. When KCI retested following a "not satisfied" result, KCI often applied a less robust, less reliable retest using a smaller sample size.

- KCI should have taken steps to improve blindness to the test subject.

In short, the conclusions drawn by KCI are based on incomplete statistical analysis. More complete analysis shows that many of the conclusions that particular standards were satisfied are not justified by the data in the test. Unless KCI corrects its errors in carrying out its statistical analysis as I recommend, this Commission should not rely on the conclusions reached by KCI in its Final Report.

Q. PLEASE DESCRIBE THE STANDARDS USED IN THE THIRD-PARTY TEST.

A. Some background information regarding the standards used for the test is necessary to understand my concerns regarding the statistical analysis of test results. KCI used three types of standards: (1) parity with a retail analog; (2) benchmarks, *i.e.*, quantitative standards set by either the Georgia Public Service Commission ("GPSC"), BellSouth, or KCI; and (3) subjective, non-quantitative standards set by KCI. Because KCI applied statistical methods only to the quantitative standards (1) and (2) above, I will comment only on those.

Q. PLEASE PROVIDE ANY ADDITIONAL BACKGROUND INFORMATION NECESSARY FOR YOUR ANALYSIS.

A. When reviewing the analysis of test results against these two standards, it is useful to consider three quantities that share a common scale: (1) the observed service,

(2) the true service, and (3) the standard. The definition of each of these quantities differs slightly depending upon whether the standard is a benchmark or parity.

For benchmarks, the truth, an unknown quantity that characterizes the service process, is the mean or average that would occur if we could observe an unlimited number of pseudo-CLEC observations under the same conditions as the test. The observed service is simply the mean (average) measurement or the proportion of successes observed for all pseudo-CLEC cases in the test. The observed service estimates the true service based on a finite sample of observations. The benchmark standard is a specified quantity—*e.g.*, a 95% success rate.

Parity measures involve a comparison of the results or process available for BellSouth retail with the results or process provided for the pseudo-CLEC. For parity measures, the truth (or true service) is the difference between the values of the mean (or proportion) for unlimited pseudo-CLEC service and the mean (or proportion) for unlimited retail results. Similarly, the observed service for parity purposes is the difference between the mean (average) measurement or the proportion of successes observed for the pseudo-CLEC cases and the corresponding value for BellSouth's retail cases. For parity measures, the standard is always a difference of zero.

Q. PLEASE EXPLAIN WHY KCI SHOULD NOT HAVE APPLIED STATISTICAL ANALYSIS TO ANY TESTS WITH BENCHMARK MEASURES.

A. KCI's statistical analyses (P-values) for benchmarks assume that the standard (*e.g.*, 95% success) was intended as a standard for the true service. However, both BellSouth's V-SEEM and the AT&T Performance Incentive Plan proposed in Georgia (and more recently in Florida and North Carolina) treat benchmarks as strict cutoffs for all samples of 30 or larger. Based on this interpretation, it was improper for KCI to use statistical analysis to reclassify benchmark standards as satisfied.

Q. FOR WHICH TESTS DID KCI IMPROPERLY APPLY A STATISTICAL (P-VALUE) ANALYSIS?

A. The Commission set the benchmark measures for the following twenty-nine individual tests in the third-party test: PRE-1-1-1, PRE-4-1-1, PRE-5-1-1, O&P-1-1-1, O&P-1-3-2a, O&P-1-3-2b, O&P-1-3-3a, O&P-1-3-3b, O&P-1-3-5, O&P-2-1-1, O&P-2-3-2a, O&P-2-3-2b, O&P-2-3-3a, O&P-2-3-3b, O&P-3-1-1, O&P-3-1-2, O&P-3-3-3, O&P-3-3-4, O&P-4-1-1, O&P-4-1-2, O&P-4-2-1, O&P-4-2-2, O&P-4-3-3, O&P-4-3-4, O&P-5-2-3, O&P-10-1-1, O&P-10-1-2, O&P-10-3-3, and O&P-10-3-4. Consequently, statistical analysis (P-values) should not be applied to any of these benchmarks. Any benchmark for which the observed service does not meet or exceed the standard should be classified as not satisfied.

Q. WHAT DOES THE TERM "SATISFIED" MEAN FOR THE PURPOSES OF THE THIRD-PARTY TEST REPORT?

A. Based on the way KCI structured the test, "satisfied" in the Final Report does not mean that BellSouth met the specified standard. Instead, "satisfied" implies only that there was not enough evidence to conclude that the true service process was below standard. In light of the small sample sizes in the test, BellSouth could easily have received a score of "satisfied" despite seriously substandard performance.

Q. HOW DID KCI STRUCTURE ITS STATISTICAL ANALYSIS?

A. KCI's analysis of test results began with the null hypothesis.[1] The null hypothesis assumes that BellSouth is performing at the standard.

Q. PLEASE EXPLAIN KCI'S COMPUTATION OF P-VALUES.

A. If the observed results in a specific test met or exceeded the standard, KCI did no further analysis and classified the standard as "satisfied." Whenever the observed results for a measure failed to reach the specified standard, KCI applied one of four statistical tests to compute a P-value. The P-value compares the observed results with the standard and provides a quantitative measure of how likely a result as bad as the one observed would be under the null hypothesis that the true service process employed by BellSouth exactly met the standard. In other words, the P-value indicates how reasonable it is to conclude that the observed result is explained solely by bad luck as opposed to any deficiency in the true service. Small P-values provide evidence against the null hypothesis that the true process was meeting the standard. They represent a smaller chance that the observed result was due to chance.

[1] The null hypothesis takes one of two forms. For parity measures, the null hypothesis states that the distribution of outcomes is the same for CLEC and BellSouth customers. For benchmarks, KCI's null hypothesis is that the mean or proportion for CLEC customers is exactly at the standard. In my view, it is inappropriate to apply statistical analysis to benchmarks at all. If it is applied, however, the guidelines discussed in this section should be followed.

Q. HOW DID KCI USE THESE P-VALUES?

A. KCI applied the following rule: If BellSouth's observed performance did not meet the specified standard, BellSouth calculated a P-value. If the P-value was less than 0.05, KCI concluded that the standard was "not satisfied." If the P-value exceeded 0.05, KCI concluded that the standard was "satisfied."

Q. PLEASE PROVIDE AN EXAMPLE OF KCI'S P-VALUE ANALYSIS.

A. Test O&P-5-2-3 measured whether provisioning was completed on time for Coordinated Customer Conversion orders. The BellSouth Service Quality Measurements Plan applies a standard of 95% within 15 minutes of the scheduled start time for coordinated customer conversions. This means that, to meet the standard, BellSouth would have to start coordinated customer conversions within 15 minutes of the scheduled start time for 95% of the coordinated conversions it performs. Out of 63 observed conversions with Georgia CLECs, there were 57 successes, a success rate of 90.4%. Because the observed rate failed to meet the standard, KCI calculated a P-value. Based on a P-value of 0.0945, KCI concluded that the 95% standard was "satisfied." In essence, KCI based its conclusion of "satisfied" on a 0.09 probability that this bad a result could occur by chance if BellSouth is meeting the standard.

Q. PLEASE EXPLAIN THE CONCEPT OF "ERROR" IN CONNECTION WITH STATISTICAL TESTING.

A. Because statistical tests are based on finite amounts of data, they are subject to error. There is some chance that a measure will be classified as not satisfied when, in fact, the true service meets the standard (*i.e.*, the observed service was substandard purely due to random variation). Likewise, when the true service fails to meet the standard, there is a chance that the statistical test will fail to find the measure in violation, again due to random variation.

Q. WHAT IS A TYPE I ERROR?

A. A Type I error occurs if the statistical analysis indicates that BellSouth is not meeting the standard when, in fact, the true service does. Type I errors occur because of random variation.

Q. WHAT IS A TYPE II ERROR?

A. A Type II error occurs if the statistical analysis indicates that BellSouth is meeting the standard when, in fact, the true service falls short of the standard by a certain amount. Like Type I errors, Type II errors occur because of random variation. In contrast to Type I errors, a Type II error requires specification of an alternative hypothesis that quantifies the size of the shortfall relative to the standard.

Q. PLEASE EXPLAIN HOW KCI LIMITED THE LIKELIHOOD THAT BELLSOUTH WOULD BE ERRONEOUSLY FOUND TO BE BELOW THE STANDARD (TYPE I ERROR).

A. KCI's rule, that if the P-value exceeded 0.05 the standard was satisfied, explicitly controls the probability of Type I error to be 0.05 or less. In other words, under KCI's rule, for any measure where BellSouth's true service process meets the standard, there is at most a 1-in-20 chance that BellSouth's observed service would be found out of compliance.

Q. IN SIMPLE TERMS, WHAT IS WRONG WITH KCI'S P-VALUE ANALYSIS?

A. It is skewed in favor of BellSouth. KCI's procedure asks the question: "Given the observed results, how good could the true service be?" However, KCI does not address the corresponding question: "Given the observed results, how bad could the true service be?" A balanced, complete analysis needs to address both questions.

Q. WHY IS IT IMPORTANT TO CONSIDER HOW BAD THE TRUE SERVICE COULD BE?

A. The latter question – Given the observed results, how bad could the true service be? – must be considered. A Type II error occurs when an *alternative hypothesis* is true (*i.e.*, BellSouth's true process fails to meet the standard by a certain amount), but the conclusion is that the standard was satisfied. Type II errors are important because they are instances in which BellSouth's process is not up to the standard but KCI concludes that it is. Consequently, it is just as important to limit the probability of Type II errors as it is to limit Type I errors.

Q. UNDER AN ANALYSIS LIKE KCI'S, WHAT DETERMINES THE SIZE OF TYPE II ERROR?

A. With KCI's fixed cutoff of 0.05 for P-values, sample size determines the probability of a specific Type II error. With sufficient sample size for a measure, Type II errors do not pose much problem. For example, consider measure O&P-5-2-3 again, for which the standard is 95%. If we used a sample size of 250 and BellSouth's true performance is equal to 90.0%, the probability of a Type II error would be just 0.12. That is, with an adequate sample size, the odds would strongly favor seeing the failure in the observed service.

Q. WERE THE SAMPLE SIZES USED IN THE THIRD-PARTY TEST LARGE ENOUGH TO MITIGATE THE OCCURRENCE OF TYPE II ERRORS?

A. No. Type II error presents a serious problem for the sample sizes generally employed in this test. With the sample of 72 used for O&P-5-2-3, the probability of a Type II error when the true success rate is 90.0% equals 0.71. In other words, if BellSouth was failing at twice the rate specified in the standard (*i.e.*, 90.0% success, 10.0% failure), the odds are more than two-to-one that KCI would have judged the standard to be satisfied.

In fact, there would still be a 0.25 probability of a Type II error even if BellSouth's true performance rate had dropped to 85.0%. In other words, even if BellSouth were performing at 85.0%, one out of four times KCI would incorrectly classify the standard as "satisfied." The test should have been designed to avoid such large Type II error probabilities for serious violations of this sort.

Q. IN LIGHT OF THE HIGH LIKELIHOOD THAT THE THIRD-PARTY TEST MISSED SERIOUS VIOLATIONS, CAN THE RESULTS OF THE THIRD-PARTY TEST BE RELIED UPON?

A. No. Based on these large Type II error probabilities, it is clear that further analysis is required before concluding that any standard in the test is satisfied.

Q. WHAT FURTHER ANALYSIS SHOULD BE DONE?

A. The best way to address both Type I and Type II errors is to compute a two-sided confidence interval that summarizes the uncertainty associated with observed performance. Two-sided confidence intervals tell a balanced, complete story by providing information about both how good and how bad the true level of service might be. For example, the exact, two-sided 90-percent confidence interval for measure O&P-5-2-3 runs from an upper limit of 95.8% down to a lower limit of 82.1%.

Q. WHAT DOES THE TWO-SIDED CONFIDENCE INTERVAL MEAN?

A. This confidence interval states the range of true proportions that are consistent with the observed results—57 successes out of 63. In other words, having observed these results, BellSouth's true performance is likely to be anything from 82.1% to 95.8%. For proportions outside the confidence interval (above 95.8% or below 82.1%), the observed number of successes is either surprisingly low or surprisingly high. By surprising, I mean that we would expect such an extreme result in either direction less than 5 percent of the time. Because the confidence interval is two-sided, the coverage probability is 90 percent (100 percent minus 5 percent off each end).

Q. HOW DOES KCI'S P-VALUE RELATE TO THE TWO-SIDED CONFIDENCE INTERVAL YOU PROPOSE?

A. The upper confidence limit tells the same story as KCI's P-value. The fact that the upper confidence limit for O&P-5-2-3 exceeds 95% is exactly equivalent to the P-value exceeding 0.05.

Q. IF THE UPPER LIMIT OF THE CONFIDENCE INTERVAL ACCOMPLISHES THE SAME GOAL AS THE P-VALUE, WHAT DOES THE LOWER LIMIT TELL US?

A. The lower confidence limit provides important information that the P-value misses. Indeed, the lower confidence limit of 82.1% for O&P-5-2-3 tells a very different story from the P-value alone. The lower confidence limit means that the observed data cannot rule out the possibility that the true failure rate was as high as 17.9%—*i.e.*, that it exceeded the standard by a factor of 2.0, 3.0, or even 3.6. Consequently, KCI's conclusion of "satisfied" could easily have been a serious Type II error. Clearly, it is wrong to conclude that BellSouth's service process meets the standard for this measure.

Q. HOW SHOULD KCI HAVE CLASSIFIED YOUR EXAMPLE, MEASURE O&P-5-2-3?

A. Rather than classify O&P-5-2-3 as satisfied, KCI should have classified it, as well as all similar measures, as "inconclusive" because neither standard performance nor significantly substandard performance could be ruled out.

Q. HOW COULD KCI'S INCOMPLETE ANALYSIS BE CORRECTED?

A. Two-sided 90-percent confidence intervals should be computed for all measures. In addition, a threshold should be specified indicating a level of poor service that must be ruled out before the standard is classified as satisfied. BellSouth's performance on a measure would be classified as satisfied only if the upper confidence limit exceeded the standard *and* the lower confidence limit exceeded the threshold. If the confidence interval is large enough to cover both values, the result would be classified as *inconclusive*. For a measure with a standard of 95 percent, I believe that 87.5% would be an appropriate threshold. This threshold represents a failure rate that is 2.5 times that specified in the standard. Observed data that are consistent with an even greater lack of compliance should not be used as evidence that the standard was satisfied. If my proposed procedure and threshold are applied to O&P-5-2-3, this standard would be classified "inconclusive."

Q. SHOULD A TWO-SIDED 90%CONFIDENCE INTERVAL BE APPLIED TO TESTS IN WHICH KCI DID NOT USE A P-VALUE ANALYSIS AS WELL?

A. Yes. If the sample is small enough, the lower confidence limit may fail to meet the threshold even when the observed result exceeds the standard. Consequently, confidence intervals should be computed and compared with a threshold even when KCI did not compute P-values.

Q. WHAT SHOULD BE DONE WITH THE INCONCLUSIVE TESTS?

A. For any measure with inconclusive results, more data should be collected. Sufficient additional data should be collected and combined with the existing data to narrow the confidence limits substantially before recomputing a P-value and confidence limits. If the new confidence limits meet the conditions outlined above, the measure can be determined to be satisfied or not satisfied.

Q. IN ADDITION TO YOUR EXAMPLE, O&P-5-2-3, HOW MANY OTHER TESTS EMPLOYED INADEQUATE SAMPLE SIZES?

A. The problem illustrated by the inadequate sample size of O&P-5-2-3 is not an isolated instance. Similar sample sizes occurred for the following thirteen measures: PRE-1-3-1 (n=57), PRE-1-3-2 (n=68), PRE-1-3-3 (n=73), PRE-1-3-5 (n=51), PRE-1-3-9 (n=83), O&P-1-3-2b (n=70), O&P-1-3-3a (n=50), O&P-1-3-3b (n=50), O&P-2-3-2a (n=89), O&P-2-3-3b (n=74), O&P-5-2-1 (n=89), O&P-5-2-2 (n=72), and O&P-5-2-5 (n=55).[2]

Q. DO ANY OF THESE TESTS WARRANT SPECIAL CONCERN?

A. Yes. Small sample size is especially a concern for five of these measures that were classified as satisfied on the basis of the statistical tests: PRE-1-3-1, PRE-1-3-2, O&P-2-3-2a, O&P-5-2-2, and O&P-5-2-5. For example, the confidence intervals for O&P-5-2-2 (n = 72) and O&P-5-2-5 (n = 55), both 95% benchmarks, look similar to that for O&P-5-2-3 and would also lead to classifications of "inconclusive." Measure O&P-2-3-2a, timeliness of response for fully mechanized order error notices, had a GPSC-set benchmark of 97% within one hour. The observed success rate was 94% (84 of 89). KCI classifies this standard as satisfied even though the 90-percent confidence interval ranges from 97.8% all the way down to 88.6%. This lower limit would imply almost four times the failure rate specified by the GPSC.

Clearly, each of these measures should lead to a classification of "inconclusive," not a classification of "satisfied." The other measures listed above, both those that were satisfied with use of the P-values and some that were satisfied without a P-value, are also suspect because of the small sample sizes tested.

---

[2] The reference to "n" refers to the sample size for the test.

Q. WHAT ABOUT TESTS WITH EXTREMELY SMALL SAMPLE SIZES?

A. An even greater problem exists with some of the very small sample sizes, for example O&P-1-3-5 (n=7), O&P-1-3-6 (n=15), O&P-2-3-6 (n=15), and O&P-5-1-1 (n=7). For these sample sizes, even 100% observed performance at the standard would be inconclusive. For example, on measure O&P-2-3-6, a 95% benchmark, BellSouth achieved 15 successes in 15 tries. Nevertheless, the lower confidence limit is only 81.9%. A sample size of 15 is simply inadequate to rule out the real possibility that true service is far below the standard.

Q. WOULD YOUR PROCEDURE LEAD TO A CLASSIFICATION OF "INCONCLUSIVE" FOR ALL TESTS?

A. No. Despite the generally inadequate sample sizes, my proposed procedure would substantiate KCI's conclusions of "satisfied" whenever the observed service is good enough to rule out true service as bad as the threshold.

Q. PLEASE SUMMARIZE THE PROBLEMS WITH KCI'S USE OF AGGREGATED DATA.

A. Evaluation of BellSouth's provisioning of individual service and activity types requires analysis of data for each individual service and activity. When it ordered this test, the Georgia Commission required KCI to perform disaggregated testing that would provide such data. *See* June 6, 2000 Georgia Public Service Commission Order, which is attached as Exhibit RMB-2. Attachment A to the Georgia Commission's June 6, 2000 Order lists the individual service and activity types that were to be tested. *See id.* Although KCI's Final Report provides test results in tables broken down by the listed service and activity types, examination of the tables reveals two problems with KCI basing its decisions on aggregated results. First, in certain instances, KCI reached a conclusion of satisfied when BellSouth's performance at the service/activity level was well below the standard for certain service types. Second, due to the small sample sizes for the individual service/activity types, most results would be statistically inconclusive.

Q. PLEASE PROVIDE EXAMPLES ILLUSTRATING KCI'S IMPROPER USE OF DISAGGREGATED DATA.

A. KCI determined that BellSouth satisfied test O&P-1-3-2b (BellSouth's EDI interface provides timely partially mechanized order clarifications), even though BellSouth did not meet the specified standard for 2-wire loops with local number portability. *See* Final Report, p. V-A-34. In the first retest, KCI evaluated 34 orders for 2-wire loops with LNP. On those transactions, BellSouth failed to meet the Georgia Commission's standards for order clarification and error notices for either fully mechanized or partially mechanized orders. BellSouth completed just 8 of 14 partially mechanized orders in less than 24 hours—significantly less than the GPSC-approved standard of 85% (P-value = 0.012). *Id.* Nonetheless, based on the summary data for partially mechanized orders across all service types, KCI determined that BellSouth has satisfied the Georgia Commission's standard for timely error and clarification notices for partially mechanized orders. Accordingly, KCI concluded that BellSouth had satisfied the test even though the test results reveal that BellSouth did not satisfy the Georgia Commission's standard for timely order and clarification notices for orders that allow a customer to keep his or her own phone number when switching carriers.[3]

In other situations, KCI determined that BellSouth satisfied the standard without testing any of certain important service/activity types. Test 1-3-3a and 1-3-3b tested firm order confirmation ("FOC") timeliness. As with the previous example, KCI did not perform its evaluation at the required levels of disaggregation, nor did it set its sample sizes to ensure adequate evaluation of each service/activity type. When evaluating FOC timeliness, KCI initially tested fifteen 2-wire loops with LNP and three LNP standalone orders. *See id.* at V-A-41-42. In the first retest, KCI evaluated twenty-six 2-wire loops with LNP and fourteen standalone LNP orders. *See id.* at V-A-43-46. After this first retest, based on summary data aggregated across all service/activity types, KCI determined that BellSouth had met the Georgia Commission standard of 85% of FOCs returned within thirty-six hours for orders that did not flow through. The disaggregated view, however, reveals that no orders for 2-wire loops with LNP and no orders for standalone LNP were included in this evaluation of non-flow-through orders. Thus, again, this test does not reveal that BellSouth has satisfied the Georgia Commission's standards for FOC timeliness for non-flow-through orders when the orders are ones that permit customers to keep their own phone numbers when switching carriers.

[3] KCI tested only one standalone LNP on this retest.

Q. WERE ANY OF KCI'S DISAGGREGATED SAMPLE SIZES SUFFICIENT TO SUPPORT KCI'S CONCLUSIONS?

A. While the two samples above present the most extreme examples, virtually none of BellSouth's sample sizes were adequate to test BellSouth's performance at the disaggregated level. For example, on tests O&P-1-3-2a and O&P-1-3-2b, which tested whether BellSouth provided timely order error notices for both fully mechanized (O&P-1-3-2a) and partially-mechanized (O&P-1-3-2b) orders, BellSouth tested small numbers of each disaggregated service type. In the first retest for fully mechanized orders, KCI tested only twenty-five 2-wire loops-design, only twenty 2-wire loops-nondesign, only thirteen 2-wire loops with LNP-design, only seven 2-wire loops with LNP-nondesign, only three switch ports, and only eleven loop port combinations. As explained above, sample sizes like these are not sufficient to support the conclusion that the standard has been met even when there are few or no observed failures.

For partially mechanized orders, BellSouth tested only twenty-three 2-wire loops-design, only six 2-wire loops-nondesign, nine 2-wire loops with LNP-design, five 2-wire loops with LNP-nondesign, one LNP stand-alone, five switch ports, and seven loop port combinations. These sample sizes are similarly too small to have any statistical power. Anyone analyzing the observed results would not have nearly enough evidence to predict BellSouth's true performance.

Q. PLEASE DESCRIBE ANY ADDITIONAL PROBLEMS WITH KCI'S AGGREGATION OF DATA.

A. KCI also based its determinations upon aggregated test results for other tests, including O&P-1-3-3a, O&P-1-3-3b, O&P-2-3-2a, O&P-2-3-2b, and

O&P-2-3-3b. Each of the problems addressed above applies to each of these tests. If BellSouth is failing to meet the standard for some services or activities, an assessment based on aggregated data could easily miss the failure. Moreover, certain activity types were not tested at all. The sample sizes for other activity types are too small to support any valid conclusions about the individual service types. Consequently, the Commission lacks the information it needs to determine whether BellSouth is meeting the standard for individual services and activities.

Q. WHY IS THE BLINDNESS OF THE THIRD-PARTY TEST IMPORTANT?

A. The validity of the conclusions of a statistical analysis can be no better than the data that go into the analysis. A key question underlying the validity of the third-party test is whether the service received by the pseudo-CLEC established by KCI is representative of service that would have been received by a real CLEC. If, for some reason, this service were substantially better, the conclusions reached by KCI would be completely invalid. KCI devotes just two short paragraphs of the Final Report to this important concern under the heading *"Blindness*" (Section II-6.5).

Q. WHAT DOES BLINDNESS MEAN IN GENERAL?

A. Blindness means that the subject of an experiment does not know whether he or she is in the treatment or control condition. The most common example of this occurs in clinical trials. Patients who believe that they are receiving an experimental treatment may tend to improve simply because of that belief. This is why control-group patients routinely receive placebos.

Q. WHAT DOES BLINDNESS MEAN SPECIFICALLY FOR THE THIRD-PARTY TEST?

A. In this test, blindness refers to BellSouth. It means that BellSouth should have been unable to distinguish service requests of the pseudo CLEC from requests of any other CLEC. That is, BellSouth should not have known which service requests it was being evaluated on. In the absence of blindness, it is impossible to establish that the results observed by KCI are representative of the service that real CLECs were receiving at the same point in time.

Q. WAS THE GEORGIA THIRD-PARTY TEST BLIND?

A. No. The test was not designed in a way to blind BellSouth to the identity of KCI orders. The report acknowledged, "Yet, it was virtually impossible for the KCI/HP test to be truly blind to BellSouth." While complete blindness was probably impossible to achieve, every effort should have been made to minimize the opportunity for BellSouth to discover the source of the service requests. Instead, as KCI reports, "Each CLEC has a unique set of IDs assigned by BellSouth that must be included in every transaction."

Q. WHAT IS THE LIKELY RESULT OF THIS DESIGN FLAW?

A. I am not alleging that BellSouth purposely gave preferential treatment to KCI requests. Rather, it is basic human nature for a person to try harder when that person knows he or she is being evaluated. This implies that lack of blindness may not be a concern for services that are fully mechanized, *e.g.*, orders that flow through BellSouth's electronic systems. However, this problem may be very important for services that depend completely, or in part, on human intervention, *e.g.*, partially mechanized or non-flow-through orders. According to BellSouth's October Flow Through information evaluated by KPMG, ten percent of orders are designed to be handled manually and another ten percent fell out for manual processing due to BellSouth-caused errors. For such measures, there remains a real risk that the test results are completely invalid.

Q. COULD THE BLINDNESS OF THE THIRD-PARTY TEST HAVE BEEN IMPROVED?

A. Yes. KCI could have established processes to minimize BellSouth's knowledge regarding impending and ongoing tests. Additionally, KCI could have reported to the Commission directly rather than to BellSouth, and KCI could have ensured that all information made available to KCI was available to all CLECs.

Q. WHAT IS THE RELEVANCE OF THE "MILITARY STYLE" TEST FORMAT?

A. Several of the concerns raised above are exacerbated by the manner in which KCI implemented military style testing. In a military style test, a mindset of "test until you pass" is adopted. *See* Final Report at II-6. This approach can impact results in several ways.

Q. HOW DOES THE MILITARY STYLE TEST FORMAT IMPACT THE CHANCE OF TYPE II ERRORS (*I.E.*, FINDING BELLSOUTH PASSED WHEN BELLSOUTH'S TRUE PERFORMANCE IS SUBSTANDARD)?

A. Military style testing greatly increases the chance of Type II errors. In other words, the military style third-party test is much more likely to conclude that BellSouth's process meets a test standard when, in fact, it does not. Suppose that BellSouth provides chronically substandard service on a particular measure. As long as there is a possibility of Type II error on a single test, successive retests will eventually lead to a pass, resulting in a finding that BellSouth satisfied the standard when the result was merely the product of chance.

Q. HOW DOES THE MILITARY STYLE TEST FORMAT KCI EMPLOYED HANDLE RETESTS?

A. KCI's military style test structure did not account for the increased scrutiny that should be applied in a retest. The fact that a measure failed one or more previous tests (a repeat offender) makes it more important to conduct a balanced analysis. This balanced analysis should include, among other things, increased sample sizes for retests. Logically, sample sizes should be larger for retests because the initial tests provided hard evidence of a problem that warrants close scrutiny. In a number of retests, however, KCI used sample sizes that were smaller than the original test where BellSouth failed. Examples include the following sixteen tests: PRE-1-3-1, PRE-1-3-2, PRE-1-3-3, PRE-1-3-4, PRE-1-3-5, PRE-1-3-6, PRE-1-3-7, PRE-1-3-8, PRE-1-3-9, O&P-2-3-2a, O&P-2-3-2b, O&P-2-3-3b, O&P-5-1-1, O&P-5-2-1, O&P-5-2-2, and O&P-5-2-5.

For six of these measures—PRE-1-3-1, PRE-1-3-2, O&P-2-3-2a, O&P-2-3-2b, O&P-5-2-2, and O&P-5-2-5—KCI reached a conclusion of "satisfied" based only on the computed P-value. As noted earlier, the inadequate sample sizes for these six measures dictate against conclusions of satisfied. When also viewed in the context of BellSouth's initial test failures, these retest results clearly do not support the conclusion that BellSouth's true performance meets the standards.

Q. HOW COULD KCI HAVE AVOIDED THESE COMPROMISING EFFECTS OF THE MILITARY STYLE TEST FORMAT?

A. KCI could have implemented military style testing differently to avoid this concern. KCI could have retested using substantially larger sample sizes.

Q. PLEASE DESCRIBE YOUR CONCERNS REGARDING KCI'S USE OF ITS "PROFESSIONAL JUDGMENT" TO DETERMINE THAT BELLSOUTH'S PERFORMANCE SATISFIED THE TESTS.

A. I am concerned that so many measures that failed on the basis of the statistical analysis were reclassified as satisfied based on "professional judgment." For example, measure PRE-1-3-8, mean time for Service Availability Queries (SAQs), had a GPSC-approved standard of parity with retail performance, which was 1.3 seconds. The mean for pseudo-CLEC cases was 11.6 seconds, which was statistically significant in comparison to the BellSouth retail mean (*i.e.*, P-value < 0.05). However, KCI reversed this classification, concluding, "it is KCI's professional judgment that the average response interval for Test-CLEC-submitted SAQ pre-orders is within a reasonable timeframe." Similar reversals occurred for another twenty tests: PRE-1-3-3, PRE-4-3-1, PRE-4-3-2, PRE-4-3-3, PRE-4-3-4, PRE-4-3-5, PRE-4-3-8, PRE-5-3-1, PRE-5-3-2, PRE-5-3-3, PRE-5-3-4, PRE-5-3-5, PRE-5-3-8, O&P-5-1-1, O&P-10-3-5, O&P-10-3-6, O&P-10-3-7, O&P-10-3-8, O&P-10-3-9, and O&P-10-3-12.

Q. WAS KCI'S DECISION TO APPLY A SUBJECTIVE STANDARD AFTER BELLSOUTH FAILED TO MEET THE OBJECTIVE STANDARDS, EVEN WITH THE AID OF STATISTICAL ANALYSIS, A VALID STATISTICAL PRACTICE?

A. Although I lack the business knowledge to comment on the validity of these judgments, I find it an unusual and questionable statistical practice to change evaluation criteria after seeing the data. If KCI thought that parity was too high a standard for certain measures, it should have specified revised criteria before doing the statistical analysis, not after it saw the results.

Q. DID KCI'S USE OF "PROFESSIONAL JUDGMENT" APPEAR CONSISTENT?

A. No. It is curious that KCI, after seeing the results, judged 11.6 seconds a reasonable average time for SAQs while, in its "professional judgment," while setting 8.0 seconds as the standard for a long series of other measures. *See, e.g.*, PRE-1-3-6, PRE-1-3-7, PRE-1-3-9, PRE-4-3-6, PRE-4-3-7, PRE-4-3-9, PRE-5-3-6, PRE-5-3-7, PRE-5-3-9, O&P-10-3-10, O&P-10-3-11, and O&P-10-3-13.

Q. HAS KCI RESPONDED TO ANY ISSUES YOU HAVE RAISED?

A. KCI filed a Motion for Leave to Basis of its Statistical Analysis in the Final Report in the proceeding before the Louisiana Public Service Commission, Docket No. U-22252-E, attached as Exhibit RMB-3. KCI stated, "[a]s the author of the Georgia 271 Test Final Reports, no other party can adequately represent and articulate the basis for KPMG Consulting's use of statistical analysis in such reports."

Q. DO KCI'S COMMENTS IMPACT YOUR ANALYSIS?

A. No. Indeed, KCI's discussion of the general design of the test acknowledges its highly subjective nature. KCI also admits that "the sample sizes for each specific service or transaction type were not designed for statistical precision."

Further, KCI's filing fails to address a number of my points. For example, regarding the use of statistical analyses, KCI does not dispute my statement that the tests provide no information regarding how poor BellSouth's true performance is likely to be in light of the observed data. While KCI's comments are technically correct on this issue, they do not address or refute my criticisms that the KCI statistical analysis is incomplete in this respect. Similarly, their defense of using statistical analysis for benchmark measures completely ignores the basis of my criticism. Both BellSouth's V-SEEM and the AT&T Performance Incentive Plan treat benchmarks as strict cutoffs for all samples of 30 or larger—implying that sufficient allowance has already been made for random variation.

Still further, KCI's challenge to my concerns regarding inadequate retest sample sizes is simply illogical. The results from the smaller sample sizes used on retest cannot, as KCI claims, have been more focused and therefore more powerful than the original test. The data are what they are. They do not become more powerful by virtue of the purpose for which they were collected. As discussed above, samples of the sizes KCI used on retest are simply too small to rule out meaningful Type II errors; therefore they cannot form the basis for reliable conclusions.

II. STATISTICAL METHODOLOGY FOR THE PERFORMANCE MEASUREMENT PLAN

Q. PLEASE SUMMARIZE THE CHANGES AT&T BELIEVES ARE NECESSARY TO IMPROVE THE GEORGIA PERFORMANCE MEASURES PLAN.

A. This Commission should not adopt the BellSouth Performance Measurements plan that was adopted in Georgia. Instead the Commission should make the following improvements to allow a more accurate assessment of whether BellSouth provides nondiscriminatory access to its network. First, AT&T recommends use of the modified z statistic to make parity determinations. Second, the parameter delta value should be lower to ensure parity. Third, the remedy plan should calculate the parity gap in a way that penalizes BellSouth based on how far it strays from providing parity of performance.

Q. WHY ARE STATISTICAL TESTS USEFUL AS PART OF A PERFORMANCE MEASURES PLAN?

A. Merely reporting averages of performance measurements, without further analysis, does not indicate whether differences in performance results for CLEC customers versus a retail analog reflect actual discrimination or simply random variation. Once appropriate measures and comparison samples have been established, statistical tests compare the size of observed differences with the amount that could be expected to occur by chance under conditions of true parity of service. These comparisons help to determine quantitatively whether BellSouth has provided nondiscriminatory treatment to CLECs for measures with a retail analog. The FCC supported the use of statistical comparisons in its Bell Atlantic Order for New York. *See In the Matter of Application of Bell Atlantic for*

Provision of In-Region InterLATA Services In New York, CC Docket No. 99-295 (December 23, 1999), Appendix B, Para. 2&4. In that Order, the FCC stated:

> When making a parity comparison, statistical analysis is a useful tool to take into account random variations in the metrics. In the Second BellSouth Louisiana Order, we encouraged BOCs to submit data allowing us to determine if any detected difference between the wholesale and retail metrics is statistically significant.

Q. WHAT STATISTICAL METHODOLOGY DOES AT&T RECOMMEND?

A. AT&T recommends use of the modified z statistic and the balancing critical value method to make parity determinations. The modified z statistic is described in a paper attached to this testimony as Exhibit RMB-4.[4] For each parity submeasurement (a disaggregated measure), BellSouth's performance for its retail operation (or that of its affiliates) is compared with the performance it provides to a given CLEC to create a z score (the modified z statistic), which then can be used to determine whether BellSouth's performance for the CLEC is in parity with its performance for its retail operation. For small sample sizes (30 or fewer observations in either of the data sets to be compared), permutation analysis is used to compute the z score. Permutation analysis is a computer-intensive method that compares the observed results for the CLEC customers with the distribution of results that would be observed if the CLEC customers had been drawn at random from the pool of CLEC and BellSouth customers (*see* Exhibit RMB-4, "Permutation Analysis Procedural Steps").

---

[4] *See* Exhibit RMB-4, "Statistical Tests for Local Service Parity," Version 1.0, February 6, 1998, Local Competition Users Group.

Out-of-parity performance occurs when the z score falls below a pre-specified critical value that depends on the two sample sizes. Values of z that fall below the critical value are taken as indications of discrimination. AT&T uses a principle called "balancing" to determine the critical value.

Q. IS MODIFIED Z AN APPROPRIATE COMPONENT OF THE STATISTICAL METHODOLOGY FOR MAKING PARITY DETERMINATIONS?

A. Yes. Experience with BellSouth's raw data confirms that the modified z statistic is an appropriate and effective component of the methodology for parity determinations. In its August 31, 1998 order in Docket No. U-22252-C, the Louisiana Public Service Commission required BellSouth to give CLECs access to raw data that underlies BellSouth's reports.[5] In that proceeding, Dr. Colin Mallows, an AT&T statistician, was able to receive and work with at least some of BellSouth's performance data in order to assess the performance of the statistical test.[6] The Louisiana Public Service Commission's order provided the opportunity for Dr. Mallows to actually see raw data and, thereby, confirm and refine the statistical methodology. Dr. Mallows' analysis of the raw data confirmed that the modified z statistic is an effective component of the methodology for parity determinations.

---

[5] Order, *In re: BellSouth Telecommunications, Inc., Service Quality Performance Measurements*, Docket No. U-22252, Subdocket C, August 31, 1998.

[6] Pursuant to a protective agreement, BellSouth provided some of its raw data associated with four measures it includes in its SQM. The measures for which Dr. Mallows received some raw data were: Order Completion Interval, Maintenance Average Duration, Missed Repair Appointments, and Missed Installation Appointments.

Q. DOES AT&T AGREE WITH THE REMEDY PLAN'S IMPLEMENTATION OF TRUNCATED Z?

A. No. Although truncated z is a valid method for aggregating cells, AT&T believes that the remedy plan implements truncated z improperly by aggregating cells in a way that could conceal discrimination.

Q. COULD TRUNCATED Z ALLOW PARITY SERVICE IN SOME CELLS TO CONCEAL DISCRIMINATION IN OTHER CELLS?

A. Yes, it could. The truncation step, setting $Z_j^* = \min(0, Z_j)$, is designed to keep a single cell where the CLEC's customers receive much better than parity service from canceling out poor service in other cells. However, it does not prevent parity, or better, service in a large number of cells from concealing very poor service in other cells. Suppose that in cells being aggregated BellSouth provides very poor service in a few cells (*e.g.*, modified z scores extreme enough to rule out random variation as the explanation) and parity service in other cells. The more parity cells that are included, the greater the chance is that truncated z will not be significant. The reason is that each cell that is found to be in parity increases the value of the truncated z statistic (high values are taken as evidence of parity). In addition, each new cell (whether in parity, or not) decreases the balancing critical value that truncated z must fall below to be judged significant. Similarly, parity service in just a few large cells can conceal very poor service in much smaller cells because truncated z weights the modified z scores according to sample sizes in the cells.

Q. CAN YOU PROVIDE A SIMPLE ILLUSTRATION OF HOW THIS WORKS?

A. Consider a simple example with just two cells, using delta equal to 1.0. Assume that BellSouth provides a very large number of DS3 and POTS loops to itself with means and standard deviations of 5 days for each product. Now suppose that BellSouth provides a CLEC 30 DS3 loops in an average of 10 days and 250 POTS loops in an average of 5.1 days. The modified z for DS3 is –5.48, overwhelming evidence of discrimination, and easily significant compared with the balancing critical value (BCV) of -2.74. The modified z for POTS is –0.32, which is not significant compared with a BCV of -7.90. If the two cells are aggregated using truncated z, the resulting truncated z score of –2.71 is much less extreme than the modified z for DS3 alone and is not close to significant when compared with the BCV of -10.24 for the aggregated test. Consequently, no remedy would be paid despite the clear evidence of large discrimination for DS3. Similar examples could easily be given for other values of delta.

Q. DOES THIS MEAN THAT TRUNCATED Z SHOULD NOT BE USED?

A. Not necessarily. I support truncated z as a method for aggregation of homogeneous cells. However, aggregation methods—including truncated z—should not be used to aggregate heterogeneous cells, for example, for services that involve distinct delivery processes.

Q. PLEASE EXPLAIN THE BALANCING CRITICAL VALUE AND WHY IT IS IMPORTANT.

A. The balancing critical value is used, along with the modified z, to determine whether the performance for a particular measure is considered to be in violation.

As the modified z statistic is defined in the AT&T plan, negative values of modified z provide evidence that a CLEC's customers are receiving worse service than the corresponding BellSouth customers, with large negative numbers providing the most evidence. The value of the modified z statistic is compared with a pre-specified negative number, called the critical value. If modified z is more negative than the critical value, then the measure is determined to be in violation. Otherwise, the measure is not determined to be in violation, even though service for the CLEC customers may have been worse than service received by the retail customers.

Q. HOW DOES THE CHOICE OF THE CRITICAL VALUE AFFECT TYPE I AND TYPE II ERRORS?

A. The critical value balances the probabilities of Type I and Type II errors. A large negative critical value holds down the probability of a Type I error, but allows the probability of a Type II error to grow larger. A less negative critical value keeps down the probability of a Type II error but allows the probability of a Type I error to grow. Put simply, a large negative critical value reduces the possibility of determining noncompliance when BellSouth is in fact providing parity service, while less negative values reduce the possibility of determining BellSouth is compliant when in fact they are providing noncompliant support.

Q. HOW SHOULD THE ALTERNATIVE HYPOTHESIS FOR THE BALANCING METHOD BE DETERMINED?

A. The alternative hypothesis should describe the minimum degree of disparity that constitutes a "material impact" on competition. The balancing method recognizes that small degrees of disparity may not significantly hinder competition, and thereby do not require protection for the CLECs. However, the degree of disparity specified by the alternative hypothesis should not exceed the minimum amount that would constitute a material impact on competition because doing so would deny the CLECs adequate protection against that degree of discrimination.

Q. PLEASE DESCRIBE THE PARAMETER "DELTA" AND EXPLAIN WHY IT IS IMPORTANT.

A. The parameter delta can be used to define the degree of violation of parity (*i.e.*, the alternative hypothesis) for which the probability of Type II error is balanced against the probability of Type I error under parity. Delta specifies the difference between the CLEC mean and the BellSouth mean. To account for the fact that performance measures do not share a common scale, the difference between the CLEC and BellSouth means is stated as delta times the standard deviation for BellSouth customers. For example, suppose that the measure Order Completion Interval has a mean of 5.0 days and a standard deviation of 5.0 days for BellSouth customers. Then a delta of 0.25 would yield an alternative hypothesis that the true mean for CLEC customers is 6.25 days (5.0 + 0.25 x 5.0).

Q. HAS A VALUE OF THE DELTA PARAMETER BEEN AGREED ON?

A. No. AT&T's and BellSouth's statisticians agree on the principle of balancing Type I and Type II errors, but they have not agreed on a value for the delta parameter. The balancing critical value development is incomplete until the value of the delta parameter is specified.

Q. WHY HAS THE DETERMINATION OF THE DELTA PARAMETER NOT BEEN RESOLVED?

A. Resolution of this question cannot be based solely on theoretical statistical analysis. Ideally, this decision should be based on business judgment, namely by determining the smallest violation of parity that is "material." The parameter delta measures the size of this violation. Once delta is chosen, the formula makes proper allowance for the effect of the sample size. When delta is large, the balancing occurs at a more extreme degree of observed disparity. BellSouth wants a large delta because this means a smaller probability of Type I error and hence, larger probability of Type II errors for any given degree of true disparity. The CLECs want a value of delta that protects them against any degree of disparity that would pose a material obstacle to competition. If the parameter delta is set too high—such that some smaller violation would present a material obstacle to competition—then the balancing principle would be violated.

Q. AT WHAT VALUE DOES AT&T BELIEVE DELTA SHOULD BE SET?

A. Delta should be set at the minimum value that represents a material impact on competition for a particular measure. AT&T believes that any value larger than 0.25 would not adequately protect CLECs against Type II errors. Accordingly, the Commission should adopt 0.25 or less as the parameter delta value for all submeasures in both Tier I and Tier II.

Q. PLEASE EXPLAIN THE IMPLICATIONS OF DELTA.

A. To understand the implications of delta = 0.25 and alternative values of delta, consider what they imply for an interval measure. Suppose that Order

Completion Interval for BellSouth customers has a mean of 5.0 days and a standard deviation of 5.0 days. Specifying delta sets the alternative hypothesis for which Type II error is balanced against Type I error. This alternative hypothesis states that the CLEC mean equals the BellSouth mean (5.0 days) plus a disparity of delta times the BellSouth standard deviation (delta x 5.0 days). Table 1 shows what this implies for three values of delta: 0.25, 0.50, and 1.00. A value of delta equal to 0.50 would be justified only if any disparity of less than 2.5 days is judged *not* to pose a material impact on competition. A delta of 1.0 would be justified only if any disparity of less than 5 days is judged *not* to pose a material impact on competition—*i.e.*, only if a 100 percent increase in the order completion interval was judged to be immaterial.

Table 1

Implied Disparity for Order Completion Interval, by Value of Delta

	Delta		
Item	0.25	0.50	1.00
	(Days)		
Disparity [a]	1.25	2.50	5.00
CLEC mean under alternative hypothesis [b]	6.25	7.50	10.00

Table assumes the BellSouth mean and standard deviation are both 5.0 days.
[a] Disparity = delta x BellSouth standard deviation
[b] CLEC mean = BellSouth mean + disparity

Next, consider a counted measure indicating a particular service problem that is triggered for 1 percent of BellSouth's own customers. Column 1 of Table 2 shows that the degree of disparity quantified by delta equal to 0.25 implies that

5.0% of CLEC customers would encounter the same problem; that is, the CLEC rate is five times the BellSouth rate.[7] Subsequent rows of the same column show the problem rates for CLEC customers implied by a delta of 0.25 for problems that affect 5, 10, or 20 percent of BellSouth customers. AT&T judges that disparities of this size pose material obstacles to competition. Therefore, delta should be no more than 0.25. Any larger value of delta would require even greater disparities before balancing takes place. For example, for a problem that occurs for 1 percent of BellSouth customers, a delta value of 0.50 would not balance the two types of error until the CLEC rate reached 11.8%, nearly a twelve-fold increase. These disparities are highlighted in Table 2.

Table 2

Percentage of CLEC Customers Receiving Bad Service, by BellSouth Percent and Delta

	Delta		
BellSouth Percent	0.25	0.50	1.00
1.0	5.0	11.8	31.9
5.0	11.8	21.0	44.0
10.0	18.7	29.3	53.6
20.0	30.8	42.8	67.4

[7] The table assumes use of the arcsine square root transformation to stabilize the variance of observed proportions. Using this function, transformed proportions have a nearly constant variance across the range of possible true proportions.

Q. MR. VARNER URGES THIS COMMISSION TO USE THE SAME VALUES OF DELTA AS THOSE CHOSEN BY THE LOUISIANA PUBLIC SERVICE COMMISSION. DO YOU AGREE?

A. No. Mr. Varner's only justification for BellSouth's proposed values is that Louisiana used them. Although the Louisiana workshops led to the development of the balancing critical value methodology, the value of delta remained as an open issue at the conclusion of those workshops. Mr. Varner is correct that "there is no absolutely 'right' delta," but that does not mean that the Commission needs to pick a value out of thin air. As shown in Table 2 above, a delta value of 1.0 for Tier I measures means that if 5% of BellSouth's customers receive poor service for a particular proportion measure, up to 44% of CLEC customers could receive poor service before the disparity would be determined to be material.

Q. DR. MULROW (PP. 16-17) PRESENTED A HYPOTHETICAL EXAMPLE TO ILLUSTRATE THE EFFECTS OF ALTERNATIVE VALUES OF DELTA FOR A SPECIFIC MEASURE. PLEASE SUMMARIZE THE SIMILARITIES WITH YOUR EXAMPLE PRESENTED IN TABLE 1.

A. The examples are similar in several ways. Each uses a completion time as the measure: Order Completion Interval in my example and Time to Provision a Dispatched Residential Order in Dr. Mulrow's example. We each assume the same mean time for BellSouth customers (5 days) and the same standard deviation (5 days). Finally, we both present calculations purporting to show the implications of delta on what constitutes a material obstacle to competition for delta values of 0.50 and 1.0 (I also do so for delta = 0.25).

Q. HOW DO THE TWO EXAMPLES DIFFER?

A. The most important difference is between our criteria for evaluating a specific value of delta, for example, 0.50. My criteria are based upon the principle that the degree of disparity specified by the alternative hypothesis should not exceed the minimum amount that would constitute a material impact on competition. The degree of disparity between the BellSouth and CLEC means under the null hypothesis equals delta x BellSouth's standard deviation, which is 2.5 days for delta = 0.50 and a standard deviation of 5 days. In contrast, Dr. Mulrow's criteria assume that as long as the disparity did not exceed 0.5 x delta x BellSouth's standard deviation (1.25 days for delta = 0.50 and a standard deviation of 5.0 days), the difference would not be material. Dr. Mulrow's definition of materiality differs from mine by a factor of one-half.

Q. ISN'T THIS DISAGREEMENT ABOUT THE DEFINITION OF MATERIALITY JUST A MATTER OF SEMANTICS?

A. No. If the Commission decides that a disparity of 1.25 days (6.25 days for CLEC customers versus 5.0 days for BellSouth customers) is material, it needs to know what that implies about delta for measures like this. My definition would imply that delta should be no larger than 0.25, while Dr. Mulrow's definition would imply a value around 0.50.

Q. WHY IS YOUR DEFINITION THE APPROPRIATE ONE FOR THE COMMISSION TO USE?

A. Including Dr. Mulrow's factor of one-half violates the basic principle of the balancing critical value methodology. Balancing occurs when the true difference in means equals delta x BellSouth's standard deviation. The Louisiana joint statistician's report implicitly defines materiality in terms of the alternative hypothesis, "If a standard of materiality is set by stating a specific alternative hypothesis for the test, ...then a critical value can be determined so that the two error probabilities are equal." (Exhibit EJM-1, p. 8). That is, a material difference must be defined as delta x BellSouth's standard deviation (the difference between the BellSouth mean and the CLEC mean under the alternative hypothesis). If delta is set incorrectly, so that a difference of one-half that size is material, then proper balancing does not occur. The probability of a Type II error when there is a difference corresponding to one-half delta remains at 50 percent, no matter how low the Type I error falls.

Q. WHAT DOES THIS MEAN FOR SETTING DELTA?

A. If the Commission makes a judgment about how large a disparity in the Order Completion Interval poses a material obstacle to competition, that judgment implies an upper bound for how large delta should be for that measure. If a disparity of 1.25 days is material (6.25 versus 5.0), then delta should be no larger than 0.25 (based on a BellSouth standard deviation of 5.0 days). In contrast, a value of delta equal to 1.0 would be justified only if any disparity less than 5.0 days is judged not to pose a material risk to competition. As Table 1 illustrates, selection of a delta value of 1.0 would allow BellSouth to take up to twice as long to complete orders for CLEC customers than for BellSouth customers before the disparity in service is determined to be material. Similarly, selection of a delta value of 0.50 would allow BellSouth to take up to 2.5 days longer to complete orders for CLEC customers before the disparity in service is determined to be material.

Q. WHAT ARE THE CONSEQUENCES OF SETTING DELTA TOO LARGE?

A. Suppose that delta is set substantially above the minimum value that represents material impact on competition for a particular measure. Then the CLECs will face greater risk of a Type II error in the face of disparity constituting material impact than BellSouth would face of a Type I error under parity. In other words, proper balancing would not occur. This problem would be magnified for large sample sizes, because balancing can produce unconventionally large, negative critical values. For example, with samples sizes of 2,500 and 250 for BellSouth and a CLEC, respectively, a delta equal to 0.50 yields a balancing critical value of –3.77, corresponding to a Type I error probability of 0.00008 (*i.e.*, 1 in 12,000), far below any conventional significance level used in statistical testing. Consequently, compelling statistical evidence of discrimination, *e.g.*, a z score of -3.6, might be ignored. Such an outcome would be justified only if one could be certain that delta had not been set too large.

Q. HOW CAN THIS PROBLEM BE SOLVED?

A. If there is concern that delta is set too large, one solution would be to place a lower limit on the size of the critical value. That is, for a given delta, the balancing approach is employed as sample size increases until the BCV reaches a specified limit, and then balancing is stopped. As sample size increases beyond this point, the critical value remains at the specified limit and the probability of a

Type I error would remain fixed. For example, a floor of –3.73 for the critical value would still produce an extremely conservative 0.0001 level of significance (probability of Type I error). With a delta of 0.25 or less, as recommended by AT&T, a floor value should be unnecessary.

Q. WHAT DOES AT&T RECOMMEND THAT THIS COMMISSION ORDER CONCERNING THE STATISTICAL METHODOLOGY?

A. There are two things that should be included in the Commission's order. First, AT&T proposes that the modified z be the statistic used for making parity determinations. Second, AT&T proposes that this Commission order the parameter delta value be set no higher than 0.25 for all submeasures.

Q. WHEN THE MODIFIED Z AND A DELTA VALUE OF 0.25 OR LESS ARE ESTABLISHED, WILL AT&T BE SATISFIED THAT THE RECOMMENDED STATISTICAL METHODOLOGY WILL ACCURATELY EVALUATE BELLSOUTH'S PERFORMANCE?

A. Yes. Although no perfect methodology for this purpose can be created, the methodology proposed by AT&T will be fair to both sides. We expect to monitor how the methodology works in "production mode," when very large amounts of data are being analyzed. AT&T's statistician will monitor how the methodology works after implementation and will make recommendations for improvements, if necessary.

Q. PLEASE SUMMARIZE YOUR CONCERN REGARDING THE REMEDY CALCULATIONS.

A. The remedy calculations should be improved. Under the SEEM remedy procedure BellSouth advocates, BellSouth may stray far from providing parity with only limited consequences. Absent meaningful consequences, BellSouth has little incentive to provide parity.

Q. ARE THE CALCULATIONS ILLUSTRATING THE SEEM REMEDY PROCEDURE, ON PAGES 39-41 OF EXHIBIT AJV-3, CORRECT?

A. No. Although the ILEC sample sizes for cells 1-10, which are not provided, would be required to validate the modified z and truncated z values, there is enough information available to prove that the balancing critical values shown in the tables are wrong by as much as a factor of 70. The tables all report balancing critical values of –0.21. However, for Order Completion Interval (p. 5), if the total ILEC sample size of 50,000 is divided equally among the ten cells, the correct balancing critical value (BCV) is -14.58. If, instead, the ILEC sample is divided in proportion to the CLEC sample, the correct BCV is -14.67. Even if each ILEC cell size were only 10 (for a total of ILEC sample of 100), the correct BCV would be -4.75. Under any of these three scenarios for the correct BCV, a truncated z of –1.92 would not even approach the BCV, and no payout would be made. Consequently, all three tables give a distorted impression of the SEEM remedy procedure.

Q. THE SEEM REMEDY CALCULATION MULTIPLIES "PER AFFECTED ITEM" DOLLAR AMOUNTS BY A CALCULATED "AFFECTED VOLUME." DO YOU BELIEVE THAT THE AFFECTED VOLUME CALCULATION FOR RETAIL-ANALOG MEASURES ESTIMATES A TRUE AFFECTED VOLUME?

A. No. BellSouth presents no evidence that its calculation produces any semblance of a true affected volume, nor does anything in the formulas suggest that it would. Indeed, the so-called "parity gap," which is a direct factor in the affected volume, is clearly intended to calculate something else. Instead of computing how far BellSouth was from providing parity service, the parity gap computes how far BellSouth was from not being found in violation. Consider this analogy. Suppose that the police patrol a stretch of highway with a 65 MPH speed limit, but that they only stop drivers who exceed 75 MPH. Also, suppose that state law calls for a fine of $10 per MPH in excess of the limit. If I am caught going 77 MPH, can I expect only a $20 fine because I was going just 2 MPH too fast to get caught? Unlikely. But that is analogous to how the BellSouth plan computes remedies. Although the statistical tests need to allow some leeway for random variation, we should not forget the goal is parity service.

Q. DOES THIS CONCLUDE YOUR TESTIMONY?

A. Yes.

Exhibit RMB-1
Curriculum Vitae

ROBERT M. BELL

EDUCATION

Ph.D., Statistics, 1980, Stanford University
M.S., Statistics, 1973, University of Chicago
B.S., Mathematics, 1972, Harvey Mudd College

PROFESSIONAL EXPERIENCE

1998-Present – Principal Member Technical Staff, Statistics Research Department, AT&T Labs - Research, Florham Park, NJ

1991-1999 -- Senior Statistician, RAND, Santa Monica, California; Head, RAND Statistics Group (1993-1995); Member, RAND Graduate School Faculty (1991-1998)

1988-1991 -- Statistician, Social Policy Department, RAND, Santa Monica, California

1980-1988 -- Associate Statistician, Economics and Statistics Department, RAND, Santa Monica, California

1975-1979 -- Teaching Assistant/Research Assistant, Department of Statistics, Stanford University.

1973-1975 -- Consultant and Mathematical Assistant, Economics Department, The RAND Corporation, (also intermittently during educational leave).

RESEARCH AREAS

Experimental Design and Survey Development. Dr. Bell supervised statistical design of Project Alert, an experiment of drug abuse prevention in thirty California and Oregon junior high schools. This work has involved data collection and analysis for sample selection/assignment, development of a series of 30 page questionnaires, and design of sampling procedures for several secondary analyses.

Data Analysis. Dr. Bell supervised the main data analysis in Project ALERT. He previously supervised analysis of clinical data from the National Preventive Dentistry Demonstration Program, a study of school-based preventive treatments. Data from that study included one to five annual examinations of 30,000 children in 10 communities, over 10,000 replicate examinations, and 20,000 surveys.

Statistical Methodology. Dr. Bell's methodological interests include survey research methods, analysis of data from complex samples, record linkage methods, analysis of missing data, measurement and scaling, robust procedures, empirical Bayes estimation, and sample reuse methods.

PROFESSIONAL ORGANIZATIONS/HONORS

Elected Fellow, American Statistical Association, 1998.
Member, Panel to Review the 2000 Census, National Academy of Sciences, 1998-present.
Chair, American Statistical Association Subcommittee, Census Advisory Committee of Professional Associations, 1997-1998; Member, 1995-2000.

Member, Panel on Alternative Census Methodologies, National Academy of Sciences, 1995-1999.
Member, Committee on Minorities in Statistics, American Statistical Association, 1995-2000.
Member, Panel to Evaluate Alternative Census Methods, National Academy of Sciences, 1992-1994.
Visiting Lecturer for American Statistical Association, 1984-1986.
Program Chairman, Applied Statistics Workshop, Southern California Section of American Statistical Association, 1984.
Institute of Mathematical Statistics, since 1979.
American Statistical Association, since 1974.

PUBLICATIONS

Published Articles

"School-Based Drug Prevention: Challenges in Designing and Analyzing Social Experiments," in *Public Policy and Statistics: Case Studies from RAND*, eds. S.C. Morton and J.E. Rolph, Springer-Verlag, New York, 2000.

"Appropriateness of the Decision to Transfer Nursing Facility Residents to the Hospital," *Journal of the American Geriatric Society*, Vol. 48, 2000, 154-163 (Saliba, Kington, Buchanan, Bell, et al.).

"Cross-Lagged Relationships among Adolescent Problem Drug Use, Delinquent Behavior, and Emotional Distress," *Journal of Drug Issues*, Vol., 30, 2000, 283-304 (Bui, Ellickson, and Bell).

"Adolescent Use of Illicit Drugs Other Than Marijuana: How Important is Social Bonding ant for Which Ethnic Groups?" *Substance Use and Misuse*, Vol. 34, 1999, 317-346 (Ellickson, Collins, and Bell).

"Simultaneous Polydrug Use among Teens: Prevalence and Predictors," *Journal of Substance Use*, Vol. 10, 1999, 233-253 (Collins, Ellickson, and Bell).

"Physician Response to Prenatal Substance Exposure," *Maternal and Child Health Journal*, 1999, 29-38 (Zellman, Bell, Archie, DuPlessis, Hoube, and Miu).

"Underuse and Overuse of Diagnostic Testing for Coronary Artery Disease in Patients Presenting with New-Onset Chest Pain," *American Journal of Medicine*, 1999, 391-398, (Carlisle, Leape, Bickel, Bell, et al.).

"Underuse of Cardiac Procedures: Do Women, Ethnic Minorities, and the Uninsured Fail to Receive Needed Revascularization?," *Annals of Internal Medicine*, Vol. 130, 1999, 183-192 (Leape, Hilborne, Bell, Kamberg, and Brook).

"The Sexual Practices of Asian and Pacific Islander High School Students," *Journal of Adolescent Health*, Vol. 23, 1998, 221-231 (Schuster, Bell, Nakajima, and Kanouse).

"Does Early Drug Use Increase the Risk of Dropping out of High School?," *Journal of Drug Issues*, Vol. 28, 1998, 357-380 (Ellickson, Bui, Bell, and McGuigan).

"Impact of a High School Condom Availability Program on Sexual Attitudes and Behaviors," *Family Planning Perspectives*, Vol. 30, 1998, 67-72 & 88 (Schuster, Bell, Berry, and Kanouse).

"Analytic Versus Holistic Scoring of Science Performance Tasks," *Applied Measurement in Education*, Vol. 11, 1998, 121-137 (Klein, Stecher, Shavelson, McCaffrey, Ormseth, Bell, Comfort, and Othman).

"Influencing Physician Response to Prenatal Substance Exposure Through State Legislation and Work-Place Policies," *Addiction*, Vol. 92, 1997, 1123-1131 (Zellman, Jacobson, and Bell).

"Adjusting Cesarean Delivery Rates for Case Mix," *Health Services Research*, Vol. 32, 1997, 509-526. (Keeler, Park, Bell, Gifford, and Keesey).

"Students' Acquisition and Use of School Condoms in a High School Condom Availability Program," *Pediatrics*, Vol. 100, October 1997, 689-694 (Schuster, Bell, Berry, and Kanouse).

"Impact Of Response Options And Feedback About Response Inconsistencies On Alcohol Use Self-Reports By Microcomputer," *Journal of Alcohol and Drug Education*, Vol. 42, 1997, 1-18 (Hays, Bell, Gillogly, Hill, Giroux, Davis, Lewis, Damush, and Nicholas).

"Adjusting for Attrition in School-Based Samples: Bias, Precision, and Cost Trade-Offs of Three Methods," *Evaluation Review*, Vol. 21, October 1997, 554-567 (McGuigan, Ellickson, Hays, and Bell).

"Teenagers and Alcohol Misuse in the United States: By any Definition, it's a Big Problem," *Addiction*, Vol. 91, 1996, 1489-1506 (Ellickson, McGuigan, Adams, Bell, and Hays).

"Communication Between Adolescents and Physicians About Sexual Behavior and Risk Prevention," *Archives of Pediatrics and Adolescent Medicine*, Vol. 150, 1996, 906-913 (Schuster, Bell, Petersen, and Kanouse).

"The Sexual Practices of Adolescent Virgins: Genital Sexual Activities of High School Students Who Have Never Had Vaginal Intercourse," *American Journal of Public Health*, Vol. 86, 1996, 1570-1576 (Schuster, Bell, and Kanouse).

"How Will the NCAA's New Standards Affect Minority Student-Athletes?," *Chance*, Vol. 8, 18-21, Summer 1995 (Klein and Bell).

"Discussion of Census 2000: Statistical Issues in Reengineering the Decennial Census," *Proceedings of the Social Statistics Section, American Statistical Association*, 1995, 17-18 (Bell).

"Effects of Reporting Methods on Infant Mortality Rate Estimates for Racial and Ethnic Subgroups," *Journal of Health Care for the Poor and Underserved*, Vol. 6, 1995, 60-75 (Farley, Richards, and Bell).

"Do Response Options Influence Self-Reports of Alcohol Use?," *The International Journal of the Addictions*, Vol. 29, 1994, 1909-1920 (Hays, Bell, Damush, Hill, DiMatteo, and Marshall).

"The Utility of Multiple Raters and Tasks in Science Performance Assessments," *Educational Assessment*, Vol. 2, 1994, 257-272 (Saner, Klein, Bell, and Comfort).

"Sampling and Statistical Estimation in the Decennial Census," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1994, 71-79 (Bell).

"The Impact of Response Options and Location in a Microcomputer Interview on Drinking Drivers' Alcohol Use Self-Reports," *Alcohol and Alcoholism*, Vol. 29, 1994, 203-209 (Hays, Bell, Hill, Gillogly, Lewis, Marshall, Nicholas, and Marlatt).

"The Urge to Merge: Linking Vital Statistics Records and Medicaid Claims," *Medical Care*, Vol. 32, 1994, 1004-1018, reprinted by invitation in *Yearbook of Medical Informatics*, 1995, 366-380 (Bell, Keesey, and Richards).

"The 1966 Enactment of Medicare: Its Effect on Discharges from Los Angeles County-Operated Hospitals," *American Journal of Public Health*, Vol. 84, 1994, 1325-1327 (Glassman, Bell, and Tranquada).

"The Urge to Merge: A Computational Method for Linking Datasets with No Unique Identifier," *Proceedings of the 18th Annual SAS Users' Group International Conference, 1993* (Bell, Keesey, and Richards).

"Using Response Agreement to Evaluate Suspect Links on a Longitudinal Survey," *Proceedings of Section on Survey Research Methods, American Statistical Association*, 1993, 286-291 (Bell).

"Changing Adolescent Propensities to Use Drugs: Results from Project ALERT," *Health Education Quarterly*, Vol. 20, 1993, 227-242 (Ellickson, Bell, and Harrison).

"Response Times for the CAGE, Short-MAST, AUDIT, and JELLINEK Alcohol Scales," *Behavior Research Methods, Instruments, & Computers*, Vol. 25, 1993, 304-307 (Hays, Hill, Gillogly, Lewis, Bell, and Nicholas).

"Do Drug Prevention Effects Persist into High School? How Project ALERT Did with Ninth Graders," *Preventive Medicine*, Vol. 22, 1993, 463-483 (Bell, Ellickson, and Harrison).

"Preventing Adolescent Drug Use: Long Term Results of a Junior High Program," *American Journal of Public Health*, Vol. 83, 1993, 856-861 (Ellickson, Bell, and McGuigan).

"Stepping Through the Drug Use Sequence: Longitudinal Scalogram Analysis of Initiation and Regular Use," *Journal of Abnormal Psychology*, Vol. 101, 1992, 441-451 (Ellickson, Hays, and Bell).

"New DEALEs: Other Approximations of Life Expectancy," *Medical Decision Making*, Vol. 12, 1992, 307-311 (Keeler and Bell).

"A Microcomputer Assessment System (MAS) for Administering Computer-Based Surveys: Preliminary Results from Administration to Clients at an Impaired-Driver Treatment Program," *Behavior Research Methods, Instruments, & Computers*, Vol. 24, 1992, 358-365 (Hays, Gillogly, Hill, Lewis, Bell, and Nicholas).

"Challenges to Social Experiments: A Drug Prevention Example," *J. Res. in Crime and Delinquency*, Vol. 29, 1992, 79-101 (Ellickson and Bell).

"Preventing Drug Use among Young Adolescents," *The Education Digest*, Vol. 56, 1990, 63-67 (Ellickson and Bell).

"Assessing Cost Effects of Nursing-Home-based Geriatric Nurse Practitioners," *Health Care Financing Review*, Vol. 11, No. 3, 1990, 67-78 (Buchanan, Bell, Arnold, Witsberger, Kane, and Garrard).

"Drug Prevention in Junior High: A Multi-Site Longitudinal Test," *Science*, Vol. 247, 1990, 1299-1305 (Ellickson and Bell).

"A Case Study in Contesting the Conventional Wisdom: School Based Fluoride Mouthrinse Programs in the USA," *Community Dentistry and Oral Epidemiology*, Vol. 18, 1990, 46-54 (Disney, Bohannan, Klein, and Bell).

"Does Pooling Saliva for Cotinine Testing Save Money Without Losing Information?," *Journal of Behavioral Medicine*, Vol. 12, October 1989, 503-507 (Bell and Ellickson).

"Affirmative Action in Medical Education and its Effect on Howard and Meharry: A Study of the Class of 1975," *Journal of the National Medical Association*, Vol. 80, 1988, 153-158 (Klein, Bell, and Williams).

"Game-Theoretic Optimal Portfolios," *Management Science*, Vol. 34, 1988, 724-733 (Bell and Cover).

"Value Preferences for Nursing Home Outcomes," *The Gerontologist*, Vol. 26, 1986, 303-308 (Kane, Bell, and Riegler).

"Conjecture Versus Empirical Data: A Response to Concerns Raised about the National Preventive Dentistry Demonstration Program (Different Views)," *Am J. Public Health*, Vol. 76, 1986, 448-452 (Klein, Bohannan, Bell, Disney, and Graves).

"Effects of Affirmative Action in Medical Schools, a Study of the Class of 1975," *New England Journal of Medicine*, Vol. 313 (Special Article), 1985, 519-525 (Keith, Bell, Swanson, and Williams).

"The Cost and Effectiveness of School-Based Preventive Dental Care," *American Journal of Public Health*, Vol. 75, 1985, 382-391 (Klein, Bohannan, Bell, Disney, Foch, and Graves).

"Management and Evaluation of the Effects of Misclassification in a Controlled Clinical Trial," *Journal of Dental Research*, Vol. 63 (Special Issue), 1984, 731-734 (Bell and Klein).

"Predicting the Course of Nursing Home Patients: A Progress Report," *The Gerontologist*, Vol. 23, 1983, 200-206 (Kane, Bell, Riegler, Wilson, and Keeler).

"Assessing the Outcomes of Nursing-Home Patients," *Journal of Gerontology*, Vol. 38, 1983, 385-393 (Kane, Bell, Riegler, Wilson, and Kane).

"An Adaptive Choice of the Scale Parameter for M-Estimators of Location," Ph.D. thesis, Stanford University, 1980 (Bell).

"Competitive Optimality of Logarithmic Investment," *Mathematics of Operations Research*, Vol. 5, 1980, 161-166 (Bell and Cover).

National Academy of Sciences Panel Reports

Measuring a Changing Nation: Modern Methods for the 2000 Census, Panel to Evaluate Alternative Census Methodologies, National Research Council, Committee on National Statistics, Michael L. Cohen, Andrew A. White, and Keith F. Rust (Eds.), National Academy Press, Washington, D.C., 1999.

Preparing for the 2000 Census: Interim Report II, Panel to Evaluate Alternative Census Methodologies, National Research Council, Committee on National Statistics, Andrew A. White and Keith F. Rust (Eds.), National Academy Press, Washington, D.C., 1997.

Sampling in the 2000 Census: Interim Report I, Panel to Evaluate Alternative Census Methodologies, National Research Council, Committee on National Statistics, Andrew A. White and Keith F. Rust (Eds.), National Academy Press, Washington, D.C., 1996.

Counting People in the Information Age, Final Report, Panel to Evaluate Alternative Census Methods, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council, National Academy Press, Washington, D.C., 1994.

A Census that Mirrors America, Interim Report, Panel to Evaluate Alternative Census Methods, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council, National Academy Press, Washington, D.C. 1993.

RAND Publications

The Sexual Practices of Asian and Pacific Islander High School Students, RP-744, RAND, 1998 (Schuster, Bell, Nakajima, and Kanouse).

Analysis of Data from Complex Surveys (videorecording), Statistics Short Course Series, V-092, RAND, 1997 (McCaffrey and Bell).

Graphical Methods for Data Analysis, (videorecording), Statistics Short Course Series, V-022 through V-025, RAND 1996 (Bell and McCaffrey).

Defining Infants' Race and Ethnicity in a Study of Very Low Birthweight Infants, MR-191-AHCPR, RAND, 1993 (Farley, Richards, and Bell).

Do Teens Tell the Truth? The Validity of Self-Reported Tobacco Use in Adolescents, N-3291-CHF, RAND, July 1991 (Freier, Bell, and Ellickson).

How Accurate Are Adolescent Reports of Drug Use?, N-3189-CHF, RAND, May 1991 (Reinisch, Bell, and Ellickson).

Multiplying Inequalities, The Effects of Race, Social Class, and Tracking on Opportunities to Learn Mathematics and Science, R-3928-NSF, RAND, July 1990 (Oakes, Ormseth, Bell, and Camp).

Baseline Nonresponse in Project ALERT: Does it Matter?, N-2933-CHF, The RAND Corporation, Santa Monica, California, April 1990 (Bell, Gareleck, and Ellickson).

Prospects for Preventing Drug Use Among Young Adolescents, R-3896-CHF, The RAND Corporation, Santa Monica, California, March 1990 (Ellickson and Bell).

The Role of Professional Background, Case Characteristics, and Protective Agency Response in Mandated Child Abuse Reporting, R-3825-HHS, The RAND Corporation, Santa Monica, California, January 1990 (Zellman and Bell).

Results from the Evaluation of the Massachusetts Nursing Home Connection Program, JR-01, The RAND Corporation, Santa Monica, California, October 1989 (Buchanan, Kane, Garrard, Bell, Witsberger, Rosenfeld, Skay, and Gifford).

A Matched Sampling Algorithm for the Nursing Home Connection Demonstration, N-2823-HCFA, The RAND Corporation, Santa Monica, California, July 1989 (Buchanan, Bell, Witsberger, Kane, Garrard, Rosenfeld, and McDermott).

Provider Visit Patterns to Nursing Home Patients, N-2824-HCFA, The RAND Corporation, Santa Monica, California, June 1989 (Buchanan, Witsberger, Bell, Kane, Garrard, and Rosenfeld).

The Financial Impact of Nursing Home-Based Geriatric Nurse Practitioners, An Evaluation of the Mountain States Health Corporation GNP Project, R-3694-HCFA/RWJ, The RAND Corporation, Santa Monica, California, May 1989 (Buchanan, Arnold, Bell, Witsberger, Kane, Garrard).

Designing and Implementing Project ALERT, A Smoking and Drug Prevention Experiment, R-3754-CHF, The RAND Corporation, Santa Monica, California, December 1988 (Ellickson, Bell, Thomas, Robyn, and Zellman).

Assessing the Outcome of Affirmative Action in Medical Schools, A Study of the Class of 1975, R-3481-CWF, The RAND Corporation, Santa Monica, California, August 1987 (Keith, Bell, and Williams).

The Cost and Effectiveness of School-Based Preventive Dental Care, R-3203-RWJ, The RAND Corporation, Santa Monica, California, April 1985 (Klein, Bohannan, Bell, Disney, Foch, and Graves).

The Dynamic Retention Model, N-2141-MIL, The RAND Corporation, Santa Monica, California, April 1985 (Fernandez, Gotz, and Bell).

The Reliability of Clinical and Radiographic Examinations in the National Preventive Dentistry Demonstration Program, R-3138-RWJ, The RAND Corporation, Santa Monica, California, June 1984 (Klein, Bell, Bohannan, Disney, and Wilson).

Treatment Effects in the National Preventive Dentistry Demonstration Program, R-3072-RWJ, The RAND Corporation, Santa Monica, California, February 1984 (Bell, Klein, Bohannan, Disney, Graves, and Madison).

Outcome-Based Reimbursement for Nursing-Home Care, R-3092-NCHSR, The RAND Corporation, Santa Monica, California, December 1983 (Kane, Bell, Hosek, Riegler, and Kane).

The Military Application Process: What Happens and Can it be Improved?, R-2986-MRAL, The RAND Corporation, Santa Monica, California May 1983 (Berryman, Bell, and Lisowski).

Predicting the Course of Nursing Home Patients: A Progress Report, N-1786-NCHSR, The RAND Corporation, Santa Monica, California, January 1982 (Kane, Riegler, Bell, Potter, and Koshland).

Results of Baseline Dental Examinations in the National Preventive Dentistry Demonstration Program, R-2862-RWJ, The RAND Corporation, Santa Monica, California, April 1982 (Bell, Klein, Bohannan, Graves, and Disney).

CETA: Is it Equitable to Women?, N-1683-DOL, The RAND Corporation, Santa Monica, California, May 1981 (Berryman, Chow, and Bell).

Plan for the Analysis of Dental Examination Data in the National Preventive Dentistry Demonstration Program, N-1658-RWJ, The RAND Corporation, Santa Monica, California, April 1981 (Klein and Bell).

Medical School and Physician Performance: Predicting Scores on the American Board of Internal Medicine Written Examination, R-1723-HEW, The RAND Corporation, Santa Monica, CA, August 1977 (Bell).

May 2001

Exhibit RMB-2

Georgia Public Service Commission Order, June 6, 2000

COMMISSIONERS:

JB DURDEN, CHAIRMAN
ROBERT B. (BOBBY) BAKER, JR.
DAVID L. BURGESS
LAUREN "BUBBA" MCDONALD, JR.
STAN WISE

DEBORAH K. FLANNAGAN
EXECUTIVE DIRECTOR

HELEN O'LEARY
EXECUTIVE SECRETARY

Georgia Public Service Commission

(404) 656-4501
1 (800) 282-5813

47 TRINITY AVENUE, S.W.
ATLANTA, GEORGIA 30334-5701

FAX: (404) 656-2341
www.psc.state.ga.us

Docket No. 8354-U

ORDER ADOPTING STANDARDS AND BENCHMARKS

IN RE: **Investigation into Development of Electronic Interfaces for BellSouth's Operational Support System**

On March 22, 2000, KPMG filed with the Commission BellSouth Telecommunications, Inc.'s ("BellSouth") proposed set of standards and benchmarks to be used in the above referenced docket. A copy of these standards and benchmarks is attached hereto as Attachment A. On March 23, 2000, AT&T Communications of the Southern States, Inc. ("AT&T"), Media One, Inc. and Sprint Communications Company, LP ("Sprint") sent letters to this Commission requesting an extension to file comments on April 12, 2000 to enable Competitive Local Exchange Carriers to adequately review the standards and benchmarks and to prepare comments. The Commission, by letter from Leon Bowles, Director of Telecommunications, granted CLECs an extension until April 5, 2000.

On April 5, 2000, AT&T and Covad Communications Company ("Covad") filed comments expressing concerns with BellSouth's standards and benchmarks to be used for the Georgia OSS evaluation. Additionally, on April 5, 2000, BellSouth filed comments to highlight that the proposed standards and benchmarks are appropriate for purposes of the third-party OSS Evaluation. On May 10, 2000, and May 23, 2000, BellSouth and AT&T respectively filed additional comments. After reviewing the proposed standards and benchmarks, and after reviewing the comments filed with the Commission, the Commission finds that it is appropriate to approve the proposed standards and benchmarks for the limited purpose of use during the third-party OSS evaluation. The approval of these standards and benchmarks for use during the third-party OSS evaluation is not intended to constitute an approval of these standards and benchmarks for any other purpose.

WHEREFORE, it is

ORDERED, that the standards and benchmarks set forth in Attachment A are hereby approved by this Commission for the limited purpose of use during the third-party OSS evaluation.

ORDERED FURTHER, that a motion for reconsideration, rehearing, or oral argument or any other motions shall not stay the effective date of this Order, unless otherwise ordered by the Commission.

ORDERED FURTHER, that jurisdiction over these matters is expressly retained for the purpose of entering such further Order or Orders as this Commission may deem just and proper.

The above action of the Commission in Administrative Session on June 6, 2000.

Helen O'Leary
Executive Secretary

06/29/00
Date

Bob Durden
Chairman

06/29/00
Date

HOL/BD/LEB

Encl.:

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

A. RESALE

	MEASUREMENTS	Analog or Benchmark	PRODUCT	MECHANIZATION
1. ORDERING	1. % Rejected Service Requests	Diagnostic	1. Residence	1. Electronic
		Diagnostic	2. Business	2. Partial Electronic
		Diagnostic	3. Design	3. Manual
		Diagnostic	4. PBX	
		Diagnostic	5. ISDN	
	2. Reject Interval	97% < 1 hr	1. Residence	1. Electronic
		97% < 1 hr	2. Business	
		97% < 1 hr	3. Design	
		97% < 1 hr	4. PBX	
		97% < 1 hr	5. ISDN	
	2. Reject Interval	85% < 24 hrs	1. Residence	2. Partial Electronic
		85% < 24 hrs	2. Business	3. Manual
		85% < 24 hrs	3. Design	
		85% < 24 hrs	4. PBX	
		85% < 24 hrs	5. ISDN	
	3. FOC Timeliness	95% < 3 Hrs	1. Residence	1. Electronic
		95% < 3Hrs	2. Business	
		95% < 3 Hrs	3. Design	
		95% < 3Hrs	4. PBX	
		95% < 3 Hrs	5. ISDN	
	3. FOC Timeliness	85% < 36 hrs	1. Residence	2. Partial Electronic
		85% < 36 hrs	2. Business	3. Manual
		85% < 36 hrs	3. Design	
		85% < 36 hrs	4. PBX	
		85% < 36 hrs	5. ISDN	

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

	MEASUREMENTS	Analog or Benchmark	PRODUCT	CIRCUIT INTERVAL	DISPATCH IDENTITY
2. PROVISIONING	1. Order Completion Interval	Parity with retail Res.	1. Residence	1. < 10 circuits	1. Dispatch
		Parity with retail Bus	2. Business	2. >= 10 circuits	2. Non Dispatch
		Parity with retail Design	3. Design		
		Parity with retail PBX	4. PBX		
		Parity with retail Centrex	5. Centrex		
		Parity with retail ISDN	6. ISDN		
	2. Held Orders	Parity with retail Res.	1. Residence	1. < 10 circuits	
		Parity with retail Bus	2. Business	2. >= 10 circuits	
		Parity with retail Design	3. Design		
		Parity with retail PBX	4. PBX		
		Parity with retail Centrex	5. Centrex		
		Parity with retail ISDN	6. ISDN		
	3. % Jeopardies (Mechanized)	Parity with retail Res.	1. Residence		
		Parity with retail Bus	2. Business		
		Parity with retail Design	3. Design		
		Parity with retail PBX	4. PBX		
		Parity with retail Centrex	5. Centrex		
		Parity with retail ISDN	6. ISDN		
	4. Average Jeopardy Notice Interval (Mechanized)	95% >=48 hrs	1. Residence		
		95% >=48 hrs	2. Business		
		95% >=48 hrs	3. Design		
		95% >=48 hrs	4. PBX		
		95% >=48 hrs	5. Centrex		
		95% >=48 hrs	6. ISDN		

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

A. RESALE

	MEASUREMENTS	*Analog or Benchmark*	*PRODUCT*	*CIRCUIT INTERVAL*	*DISPATCH IDENTITY*
2. PROVISIONING (continued)	5. % Missed Installation Appointments	Parity with retail Res. Parity with retail Bus Parity with retail Design Parity with retail PBX Parity with retail Centrex Parity with retail ISDN	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. < 10 circuits 2. >= 10 circuits	1. Dispatch 2. Non Dispatch
	6. % Provisioning Troubles within 30 Days	Parity with retail Res. Parity with retail Bus Parity with retail Design Parity with retail PBX Parity with retail Centrex Parity with retail ISDN	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. < 10 circuits 2. >= 10 circuits	1. Dispatch 2. Non Dispatch
	7. Avg. Completion Notice Interval (Mechanized Orders Only)	Parity with retail Res. Parity with retail Bus Parity with retail Design Parity with retail PBX Parity with retail Centrex Parity with retail ISDN	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. < 10 circuits 2. >= 10 circuits	1. Dispatch 2. Non Dispatch
	8. Total Svc. Ord. Cycle Time	Diagnostic Diagnostic Diagnostic Diagnostic Diagnostic Diagnostic	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. < 10 circuits 2. >= 10 circuits	1. Dispatch 2. Non Dispatch

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

A. RESALE

	MEASUREMENTS	*Analog or Benchmark*	*PRODUCT*	*DISPATCH IDENTITY*
3. MAINTENANCE & REPAIR	1. Missed Repair Appointments	Parity with retail Res. Parity with retail Bus Parity with retail Design Parity with retail PBX Parity with retail Centrex Parity with retail ISDN	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. Dispatch 2. Non Dispatch
	2. Customer Trouble Report Rate	Parity with retail Res. Parity with retail Bus Parity with retail Design Parity with retail PBX Parity with retail Centrex Parity with retail ISDN	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. Dispatch 2. Non Dispatch
	3. Maintenance Average Duration	Parity with retail Res. Parity with retail Bus Parity with retail Design Parity with retail PBX Parity with retail Centrex Parity with retail ISDN	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. Dispatch 2. Non Dispatch
	4. % Repeat Trbles within 30 Days	Parity with retail Res. Parity with retail Bus Parity with retail Design Parity with retail PBX Parity with retail Centrex Parity with retail ISDN	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. Dispatch 2. Non Dispatch
	5. Out of Service > 24 Hours	Parity with retail Res. Parity with retail Bus Parity with retail Design Parity with retail PBX Parity with retail Centrex Parity with retail ISDN	1. Residence 2. Business 3. Design 4. PBX 5. Centrex 6. ISDN	1. Dispatch 2. Non Dispatch

	MEASUREMENTS	
4. BILLING	1. Invoice Accuracy	Parity with retail
	2. Mean Time to Deliver Invoices - CRIS	Parity with retail

B. UNBUNDLED NETWORK ELEMENTS (UNEs)

	MEASUREMENTS	Analog or Benchmark	PRODUCT	MECHANIZATION
1 ORDERING	1 % Rejected Service Requests	Diagnostic	1 2-W Analog Loop Design	1 Electronic
		Diagnostic	2 2-W Analog Loop Non-Design	2 Partial Electronic
		Diagnostic	3 2-W Analog Loop w/INP Design	
		Diagnostic	4 2-W Analog Loop w/INP Non Design	
		Diagnostic	5 2-W Analog Loop w/LNP Design	
		Diagnostic	6 2-W Analog Loop w/LNP Non Design	
		Diagnostic	7 INP (Standalone)	
		Diagnostic	8 LNP (Standalone)	
		Diagnostic	9 Switch Port	
		Diagnostic	10 Loop + Port Combination	
		Diagnostic	11 Local Interconnection Trunks	
		Diagnostic	12 Local Transport	
		Diagnostic	13 UNE Other Non-Design	
		Diagnostic	14 UNE Other Design	
	2 % Rejected Service Requests	Diagnostic	1 2-W Analog Loop Design	1 Manual
		Diagnostic	2 2-W Analog Loop Non Design	
		Diagnostic	3 2-W Analog Loop w/INP Design	
		Diagnostic	4 2-W Analog Loop w/INP Non Design	
		Diagnostic	5 2-W Analog Loop w/LNP Design	
		Diagnostic	6 2-W Analog Loop w/LNP Non Design	
		Diagnostic	7 INP (Standalone)	
		Diagnostic	8 LNP (Standalone)	
		Diagnostic	9 Switch Ports	
		Diagnostic	10 Loop+Port Combination	
		Diagnostic	11 Local Interconnection Trunks	
		Diagnostic	12 Local Transport	
		Diagnostic	13 UNE Other Non Design	
		Diagnostic	14 UNE Other Design	
	3 Reject Interval	97% < 1 hr	1 2-W Analog Loop Design	1 Electronic
		97% < 1 hr	2 2-W Analog Loop Non Design	
		97% < 1 hr	3 2-W Analog Loop w/INP Design	
		97% < 1 hr	4 2-W Analog Loop w/INP Non Design	
		97% < 1 hr	5 2-W Analog Loop w/LNP Design	
		97% < 1 hr	6 2-W Analog Loop w/LNP Non Design	
		97% < 1 hr	7 INP (Standalone)	
		97% < 1 hr	8 LNP (Standalone)	
		97% < 1 hr	9 Switch Ports	
		97% < 1 hr	10 Loop+Port Combination	
		97% < 1 hr	11 Local Interconnection Trunks	
		97% < 1 hr	12 Local Transport	
		97% < 1 hr	13 UNE Other Non-Design	
		97% < 1 hr	14 UNE Other Design	
	3 Reject Interval	85% < 24 hrs	1 2-W Analog Loop Design	2 Partial Electronic
		85% < 24 hrs	2 2-W Analog Loop Non-Design	
		85% < 24 hrs	3 2-W Analog Loop w/INP Design	
		85% < 24 hrs	4 2-W Analog Loop w/INP Non Design	
		85% < 24 hrs	5 2-W Analog Loop w/LNP Design	
		85% < 24 hrs	6 2-W Analog Loop w/LNP Non Design	
		85% < 24 hrs	7 INP (Standalone)	
		85% < 24 hrs	8 LNP (Standalone)	
		85% < 24 hrs	9 Switch Ports	
		85% < 24 hrs	10 Loop+Port Combination	
		85% < 24 hrs	11 Local Interconnection Trunks	
		85% < 24 hrs	12 Local Transport	
		85% < 24 hrs	13 UNE Other Non Design	
		85% < 24 hrs	14 UNE Other Design	
	4 Reject Interval	85% < 24 hrs	1 2-W Analog Loop Design	1 Manual
		85% < 24 hrs	2 2-W Analog Loop Non-Design	
		85% < 24 hrs	3 2-W Analog Loop w/INP Design	
		85% < 24 hrs	4 2-W Analog Loop w/INP Non Design	
		85% < 24 hrs	5 2-W Analog Loop w/LNP Design	
		85% < 24 hrs	6 2-W Analog Loop w/LNP Non Design	
		85% < 24 hrs	7 INP (Standalone)	
		85% < 24 hrs	8 LNP (Standalone)	
		85% < 24 hrs	9 Switch Ports	
		85% < 24 hrs	10 Loop+Port Combination	
		85% < 24 hrs	11 Local Interconnection Trunks	
		85% < 24 hrs	12 Local Transport	
		85% < 24 hrs	13 UNE Other Non Design	
		85% < 24 hrs	14 UNE Other Design	

B. UNBUNDLED NETWORK ELEMENTS (UNEs)

	MEASUREMENTS	*Analog or Benchmark*	*PRODUCT*	*MECHANIZATION*
1. *ORDERING*	5 FOC Timeliness	95% < 3 hrs	1 2-W Analog Loop Design	1. Electronic
		95% < 3 hrs	2. 2-W Analog Loop Non-Design	
		95% < 3 hrs	3. 2-W Analog Loop w/INP Design	
		95% < 3 hrs	4. 2-W Analog Loop w/INP Non Design	
		95% < 3 hrs	5. 2-W Analog Loop w/LNP Design	
		95% < 3 hrs	6. 2-W Analog Loop w/LNP Non Design	
		95% < 3 hrs	7. INP (Standalone)	
		95% < 3 hrs	8. LNP (Standalone)	
		95% < 3 hrs	9 Switch Ports	
		95% < 3 hrs	10. Loop+Port Combination	
		95% < 3 hrs	11. Local Interconnection Trunks	
		95% < 3 hrs	12. Local Transport	
		95% < 3 hrs	13. UNE Other Non-Design	
		95% < 3 hrs	14 UNE Other Design	
	5 FOC Timeliness	85% < 36 hrs	1. 2-W Analog Loop Design	2. Partial Electronic
		85% < 36 hrs	2. 2-W Analog Loop Non-Design	
		85% < 36 hrs	3. 2-W Analog Loop w/INP Design	
		85% < 36 hrs	4. 2-W Analog Loop w/INP Non Design	
		85% < 36 hrs	5. 2-W Analog Loop w/LNP Design	
		85% < 36 hrs	6. 2-W Analog Loop w/LNP Non Design	
		85% < 36 hrs	7 INP (Standalone)	
		85% < 36 hrs	8. LNP (Standalone)	
		85% < 36 hrs	9. Switch Ports	
		85% < 36 hrs	10. Loop+Port Combination	
		85% < 36 hrs	11. Local Interconnection Trunks	
		85% < 36 hrs	12. Local Transport	
		85% < 36 hrs	13. UNE Other Non-Design	
		85% < 36 hrs	14. UNE Other Design	
	6 FOC Timeliness	85% < 36 hrs	1. 2-W Analog Loop Design	1. Manual
		85% < 36 hrs	2. 2-W Analog Loop Non-Design	
		85% < 36 hrs	3. 2-W Analog Loop w/INP Design	
		85% < 36 hrs	4. 2-W Analog Loop w/INP Non Design	
		85% < 36 hrs	5. 2-W Analog Loop w/LNP Design	
		85% < 36 hrs	6. 2-W Analog Loop w/LNP Non Design	
		85% < 36 hrs	7. INP (Standalone)	
		85% < 36 hrs	8. LNP (Standalone)	
		85% < 36 hrs	9 Switch Ports	
		85% < 36 hrs	10. Loop+Port Combination	
		85% < 36 hrs	11. Local Interconnection Trunks	
		85% < 36 hrs	12. Local Transport	
		85% < 36 hrs	13. UNE Other Non-Design	
		85% < 36 hrs	14. UNE Other Design	

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

B. UNBUNDLED NETWORK ELEMENTS (UNEs)

	MEASUREMENTS	*Analog or Benchmark*	*PRODUCT*	*CIRCUIT INTERVAL*	*DISPATCH IDENTITY*
2. PROVISIONING	1. Order Completion Interval	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. < 10 circuits	1. Dispatch
		Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. >= 10 circuits	2 Non Dispatch
		Retail Residence and Business Dispatch	3. 2-W Analog Loop w/INP Design		
		Retail Residence and Business Dispatch	4. 2-W Analog Loop w/INP Non Design		
		Retail Residence and Business Dispatch	5. 2-W Analog Loop w/LNP Design		
		Retail Residence and Business Dispatch	6. 2-W Analog Loop w/LNP Non Design		
		Retail Residence and Business Dispatch	7. INP (Standalone)		
		Retail Residence and Business Dispatch	8. LNP (Standalone)		
		Retail POTS	9. Switch Ports		
		Retail Residence and Business	10. Loop+Port Combination		
		Parity with Retail	11. Local Interconnection Trunks		
		Retail DS1 or DS3	12. Local Transport		
		Retail Residence and Business Dispatch	13. UNE Other Non-Design		
		Retail Design	14. UNE Other Design		
	2 Held Orders	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. < 10 circuits	
		Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. >= 10 circuits	
		Retail Residence and Business Dispatch	3. 2-W Analog Loop w/INP Design		
		Retail Residence and Business Dispatch	4. 2-W Analog Loop w/INP Non Design		
		Retail Residence and Business Dispatch	5. 2-W Analog Loop w/LNP Design		
		Retail Residence and Business Dispatch	6 2-W Analog Loop w/LNP Non Design		
		Retail Residence and Business Dispatch	7 INP (Standalone)		
		Retail Residence and Business Dispatch	8. LNP (Standalone)		
		Retail POTS	9. Switch Ports		
		Retail Residence and Business	10. Loop+Port Combination		
		Parity with Retail	11. Local Interconnection Trunks		
		Retail DS1 or DS3	12. Local Transport		
		Retail Residence and Business Dispatch	13. UNE Other Non-Design		
		Retail Design	14. UNE Other Design		
	3. % Jeopardies (Mechanized)	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design		
		Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design		
		Retail Residence and Business Dispatch	3 2-W Analog Loop w/INP Design		
		Retail Residence and Business Dispatch	4. 2-W Analog Loop w/INP Non Design		
		Retail Residence and Business Dispatch	5. 2-W Analog Loop w/LNP Design		
		Retail Residence and Business Dispatch	6. 2-W Analog Loop w/LNP Non Design		
		Retail Residence and Business Dispatch	7 INP (Standalone)		
		Retail Residence and Business Dispatch	8. LNP (Standalone)		
		Retail POTS	9. Switch Ports		
		Retail Residence and Business	10. Loop+Port Combination		
		Parity with Retail	11. Local Interconnection Trunks		
		Retail DS1 or DS3	12. Local Transport		
		Retail Residence and Business Dispatch	13. UNE Other Non-Design		
		Retail Design	14 UNE Other Design		

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

B. UNBUNDLED NETWORK ELEMENTS (UNEs)

	MEASUREMENTS	Analog or Benchmark	PRODUCT	CIRCUIT INTERVAL	DISPATCH IDENTITY
2. PROVISIONING	4. Average Jeopardy	95% >= 48 hours	1. 2-W Analog Loop Design		
(continued)	Notice Interval	95% >= 48 hours	2. 2-W Analog Loop Non-Design		
	(Mechanized)	95% >= 48 hours	3. 2-W Analog Loop w/INP Design		
		95% >= 48 hours	4. 2-W Analog Loop w/INP Non Design		
		95% >= 48 hours	5. 2-W Analog Loop w/LNP Design		
		95% >= 48 hours	6. 2-W Analog Loop w/LNP Non Design		
		95% >= 48 hours	7. INP (Standalone)		
		95% >= 48 hours	8. LNP (Standalone)		
		95% >= 48 hours	9. Switch Ports		
		95% >= 48 hours	10. Loop+Port Combination		
		95% >= 48 hours	11. Local Interconnection Trunks		
		95% >= 48 hours	12. Local Transport		
		95% >= 48 hours	13. UNE Other Non-Design		
		95% >= 48 hours	14. UNE Other Design		
	5. Coordinated	95% <= 15 mins	1. UNE Loops w NP		
	Customer	95% <= 15 mins	2. UNE Loops w/o NP		
	Conversions				
	6. % Missed	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. < 10 circuits	1. Dispatch
	Installation	Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. >= 10 circuits	2. Non Dispatch
	Appointments	Retail Residence and Business Dispatch	3. 2-W Analog Loop w/INP Design		
		Retail Residence and Business Dispatch	4. 2-W Analog Loop w/INP Non Design		
		Retail Residence and Business Dispatch	5. 2-W Analog Loop w/LNP Design		
		Retail Residence and Business Dispatch	6. 2-W Analog Loop w/LNP Non Design		
		Retail Residence and Business Dispatch	7. INP (Standalone)		
		Retail Residence and Business Dispatch	8. LNP (Standalone)		
		Retail POTS	9. Switch Ports		
		Retail Residence and Business	10. Loop+Port Combination		
		Parity with Retail	11. Local Interconnection Trunks		
		Retail DS1 or DS3	12. Local Transport		
		Retail Residence and Business Dispatch	13. UNE Other Non-Design		
		Retail Design	14. UNE Other Design		
	7. % Provisioning	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. < 10 circuits	1. Dispatch
	Troubles within	Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. >= 10 circuits	2. Non Dispatch
	30 Days	Retail Residence and Business Dispatch	3. 2-W Analog Loop w/INP Design		
		Retail Residence and Business Dispatch	4. 2-W Analog Loop w/INP Non Design		
		Retail Residence and Business Dispatch	5. 2-W Analog Loop w/LNP Design		
		Retail Residence and Business Dispatch	6. 2-W Analog Loop w/LNP Non Design		
		Retail Residence and Business Dispatch	7. INP (Standalone)		
		Retail Residence and Business Dispatch	8. LNP (Standalone)		
		Retail POTS	9. Switch Ports		
		Retail Residence and Business	10. Loop+Port Combination		
		Parity with Retail	11. Local Interconnection Trunks		
		Retail DS1 or DS3	12. Local Transport		
		Retail Residence and Business Dispatch	13. UNE Other Non-Design		
		Retail Design	14. UNE Other Design		
	8. Avg. Completion	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. < 10 circuits	1. Dispatch
	Notice Interval	Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. >= 10 circuits	2. Non Dispatch
	(Mechanized Orders Only)	Retail Residence and Business Dispatch	3. 2-W Analog Loop w/INP Design		
		Retail Residence and Business Dispatch	4. 2-W Analog Loop w/INP Non Design		
		Retail Residence and Business Dispatch	5. 2-W Analog Loop w/LNP Design		
		Retail Residence and Business Dispatch	6. 2-W Analog Loop w/LNP Non Design		
		Retail Residence and Business Dispatch	7. INP (Standalone)		
		Retail Residence and Business Dispatch	8. LNP (Standalone)		
		Retail POTS	9. Switch Ports		
		Retail Residence and Business	10. Loop+Port Combination		
		Parity with Retail	11. Local Interconnection Trunks		
		Retail DS1 or DS3	12. Local Transport		
		Retail Residence and Business Dispatch	13. UNE Other Non-Design		
		Retail Design	14. UNE Other Design		
	9. Total Svc. Ord.	Diagnostic	1. 2-W Analog Loop Design	1. < 10 circuits	1. Dispatch
	Cycle Time	Diagnostic	2. 2-W Analog Loop Non-Design	2. >= 10 circuits	2. Non Dispatch
		Diagnostic	3. 2-W Analog Loop w/INP Design		
		Diagnostic	4. 2-W Analog Loop w/INP Non Design		
		Diagnostic	5. 2-W Analog Loop w/LNP Design		
		Diagnostic	6. 2-W Analog Loop w/LNP Non Design		
		Diagnostic	7. INP (Standalone)		
		Diagnostic	8. LNP (Standalone)		
		Diagnostic	9. Switch Ports		
		Diagnostic	10. Loop+Port Combination		
		Diagnostic	11. Local Interconnection Trunks		
		Diagnostic	12. Local Transport		
		Diagnostic	13. UNE Other Non-Design		
		Diagnostic	14. UNE Other Design		

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

B. UNBUNDLED NETWORK ELEMENTS (UNEs)

	MEASUREMENTS	*Analog or Benchmark*	*PRODUCT*	*CIRCUIT INTERVAL*	*DISPATCH IDENTITY*
	10. Disconnect Timeliness	95% <15 min	1. LNP	1. < 10 circuits	
				2. >= 10 circuits	
	11. % Missed Installation Appointments	Retail Res + Bus Dispatch	1. LNP (Standalone)	1. < 10 circuits	
		Retail Res + Bus Dispatch	2. UNE Loops w/ LNP	2. >= 10 circuits	
	12. Total Svc. Ord Cycle Time	Diagnostic	1. LNP (Standalone)	1. < 10 circuits	
		Diagnostic	2. UNE Loops w/ LNP	2. >= 10 circuits	
	13. Total Svc. Ord Cycle Time - Offered	Diagnostic	1. LNP (Standalone)	1. < 10 circuits	
		Diagnostic	2. UNE Loops w/ LNP	2. >= 10 circuits	

	MEASUREMENTS	*Analog or Benchmark*	*PRODUCT*	*DISPATCH IDENTITY*
3. MAINTENANCE & REPAIR	1. Missed Repair Appointments	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. Dispatch
		Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. Non Dispatch
		Retail Residence and Business Dispatch	3. INP (Standalone)	
		Retail Residence and Business Dispatch	4. LNP (Standalone)	
		Retail POTS	5. Switch Ports	
		Retail Residence and Business	6. Loop+Port Combination	
		Parity with Retail	7. Local Interconnection Trunks	
		Retail DS1 or DS3	8. Local Transport	
		Retail Residence and Business Dispatch	9. UNE Other Non-Design	
		Retail Design	10. UNE Other Design	
	2. Customer Trouble Report Rate	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. Dispatch
		Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. Non Dispatch
		Retail Residence and Business Dispatch	3. INP (Standalone)	
		Retail Residence and Business Dispatch	4. LNP (Standalone)	
		Retail POTS	5. Switch Ports	
		Retail Residence and Business	6. Loop+Port Combination	
		Parity with Retail	7. Local Interconnection Trunks	
		Retail DS1 or DS3	8. Local Transport	
		Retail Residence and Business Dispatch	9. UNE Other Non-Design	
		Retail Design	10. UNE Other Design	
	3. Maintenance Average Duration	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. Dispatch
		Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. Non Dispatch
		Retail Residence and Business Dispatch	3. INP (Standalone)	
		Retail Residence and Business Dispatch	4. LNP (Standalone)	
		Retail POTS	5. Switch Ports	
		Retail Residence and Business	6. Loop+Port Combination	
		Parity with Retail	7. Local Interconnection Trunks	
		Retail DS1 or DS3	8. Local Transport	
		Retail Residence and Business Dispatch	9. UNE Other Non-Design	
		Retail Design	10. UNE Other Design	

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

B. UNBUNDLED NETWORK ELEMENTS (UNEs)

	MEASUREMENTS	Analog or Benchmark	PRODUCT	DISPATCH IDENTITY
3. MAINTENANCE & REPAIR (continued)	4. % Repeat Trbles within 30 Days	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1 Dispatch
		Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2 Non Dispatch
		Retail Residence and Business Dispatch	3. INP (Standalone)	
		Retail Residence and Business Dispatch	4. LNP (Standalone)	
		Retail POTS	5. Switch Ports	
		Retail Residence and Business	6. Loop+Port Combination	
		Parity with Retail	7. Local Interconnection Trunks	
		Retail DS1 or DS3	8. Local Transport	
		Retail Residence and Business Dispatch	9. UNE Other Non-Design	
		Retail Design	10. UNE Other Design	
	5. Out of Service > 24 Hours	Retail Residence and Business Dispatch	1. 2-W Analog Loop Design	1. Dispatch
		Retail Residence and Business Dispatch	2. 2-W Analog Loop Non-Design	2. Non Dispatch
		Retail Residence and Business Dispatch	3. INP (Standalone)	
		Retail Residence and Business Dispatch	4. LNP (Standalone)	
		Retail POTS	5. Switch Ports	
		Retail Residence and Business	6. Loop+Port Combination	
		Parity with Retail	7. Local Interconnection Trunks	
		Retail DS1 or DS3	8. Local Transport	
		Retail Residence and Business Dispatch	9. UNE Other Non-Design	
		Retail Design	10. UNE Other Design	

	MEASUREMENTS	Analog or Benchmark
4. BILLING	1. Invoice Accuracy	Parity w retail
	2. Mean Time to Deliver Invoices - CRIS	Parity w retail

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

D. OPERATIONS SUPPORT SYSTEMS (OSS)

	MEASUREMENTS	SYSTEM IDENTITY	Analog or Benchmark
1. PRE - ORDERING	1. % Interface Availability - CLEC	1. EDI	99.5%
		2. HAL	99.5%
		3. LENS	99.5%
		4. LEO MAINFRAME	99.5%
		5. LEO UNIX	99.5%
		6. LESOG	99.5%
		7. TAG	99.5%
	2. % Interface Availability - BST & CLEC	1. ATLAS/COFFI	99.5%
		2. BOCRIS	99.5%
		3. DSAP	99.5%
		4. RSAG	99.5%
		5. SOCS	99.5%
	3. Average Response Interval - CLEC(TA	1. RSAG, BY TN	Parity with retail
		2. RSAG, BY ADDR	Parity with retail
		3. ATLAS	Parity with retail
		4. DSAP	Parity with retail
		5. CRSECSR	Parity with retail
		6. CRSEINIT	Parity with retail
2. MAINTENANCE	1. % Interface Availability - BST	1. TAFI	99.5%
	2. % Interface Availability - CLEC	1. CLEC TAFI	99.5%
	3. % Interface Availability - BST & CLEC	1. CRIS	99.5%
		2. LMOS HOST	99.5%
		3. LNP	99.5%
		4. MARCH	99.5%
		5. OSPCM	99.5%
		6. Predictor	99.5%
		7. SOCS	99.5%
	4. Average Response Interval	1. CRIS	Parity by design
		2. DLETH	Parity by design
		3. DLR	Parity by design
		4. LMOS	Parity by design
		5. LMOSupd	Parity by design
		6. LNP	Parity by design
		7. MARCH	Parity by design
		8. OSPCM	Parity by design
		9. Predictor	Parity by design
		10. SOCS	Parity by design

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

E. COLLOCATION

	MEASUREMENTS	*COLLOCATION TYPE*	*Analog or Benchmark*
1. COLLOCATION	1. Average Response Time	1. Virtual	20 calendar days
		2. Physical	30 calendar days
	2. Average Arrangement Time	1. Virtual	alendar days(ordinary) : 75 (extr
		2. Physical	90 calendar days (ord) ; 130 (ex
	3. % Due Dates Missed	1. Virtual	=> 90% <= commited dates
		2. Physical	=> 90% <= commited dates

F. GENERAL

	EXTERNAL REPORT	*Analog or Benchmark*
1. FLOW THROUGH	Residence	95%
	Business	90%
	UNE	85%

	MEASUREMENTS	*Analog or Benchmark*
2. ORDERING CENTER	1. Average Speed of Answer	Parity with retail
3. MAINTENANCE CENTER	1. Average Answer Time	Parity with retail
4. OPERATOR SERVICES (TOLL)	1. Average Speed to Answer	Parity by design
	2. % Answered in 10 secs.	Parity by design
5. DIRECTORY ASSISTANCE	1. Average Speed to Answer	Parity by design
	2. % Answered in 12 secs.	Parity by design

SERVICE QUALITY MEASUREMENTS INDEX, ANALOGS and BENCHMARKS

	MEASUREMENTS	Analog or Benchmark
6. E911	1. Mean Interval	Parity by design
	2. % Accuracy	Parity by design
	3. % Timeliness	Parity by design

	MEASUREMENTS	Analog or Benchmark
7. BILLING	1. Usage Data Delivery Accuracy	Parity with retail
	2. Usage Data Delivery Timeliness	Parity with retail
	3. Usage Data Delivery Completeness	Parity with retail
	4. Mean Time to Deliver Usage	Parity with retail

G. XDSL

Exhibit RMB-3

KPMG Consulting, Inc.'s Motion for Leave to Articulate Basis for Statistical Analysis in the Georgia 271 Test Final Reports

Local Competition Users Group

Statistical Tests for Local Service Parity

February 6, 1998
Membership: AT&T, Sprint, MCI, LCI, WorldCom

Version 1.0

Executive Summary

The Local Competition Users Group has drafted 27 Service Quality Measurements (SQMs) that will be used to measure parity of service provided by incumbent local exchange carriers (ILECs) to competitive local exchange carriers (CLECs). This set of measures includes means, proportions, and rates of various indicators of service quality. This document proposes statistical tests that are appropriate for determining if parity is being provided with respect to these measurements.

Each month, a specified report of the 27 SQMs will be provided by the ILEC, broken down by the requested reporting dimensions. The SQMs are to be systematically developed and provided by the ILECs as specified. Test parameters will be calculated so that the overall probability of declaring the ILEC to be out of parity purely by chance is very small. For each SQM and reporting dimension reported, the difference between the ILEC and CLEC results is converted to a z-value. Non-parity is determined if a z-value exceeds a selected critical value.

Introduction

Purpose

The Local Competition Users Group (LCUG) is a cooperative effort of AT&T, MCI, Sprint, LCI and WorldCom for establishing standards for the entry of new companies (competitive local exchange carriers, or CLECs) into the local telecommunications market. A key initiative of the LCUG is to establish measures of parity for services provided by incumbent local exchange carriers (ILECs). In short, parity means that the support ILECs provide on behalf of the CLECs is no lesser in quality than the service provided by the ILECs to their own customers.

The LCUG has drafted a document listing service quality measurements (SQMs) that must be reported by the ILECs to insure that CLECs are given parity of support. The SQM document has been submitted to the FCC and made available to PUCs in all 50 states and is pending approval by many of these regulatory agencies. This document has been drafted to describe statistical methodology for determining if parity exists based on the measurements defined in the SQM document.

Service Quality Measurements

The LCUG has identified 27 service quality measurements for testing parity of service. These are:

Category	ID	Description
Pre-Ordering	PO-1	Average Response Interval for Pre-Ordering Information
Ordering and Provisioning	OP-1	Average Completion Interval
	OP-2	Percent Orders Completed on Time
	OP-3	Percent Order Accuracy
	OP-4	Mean Reject Interval
	OP-5	Mean FOC Interval
	OP-6	Mean Jeopardy Interval
	OP-7	Mean Completion Interval
	OP-8	Percent Jeopardies Returned
	OP-9	Mean Held Order Interval
	OP-10	Percent Orders Held > = 90 Days
	OP-11	Percent Orders Held > = 15 Days
Maintenance and Repair	MR-1	Mean Time to Restore
	MR-2	Repeat Trouble Rate

	MR-3	Trouble Rate
	MR-4	Percentage of Customer Troubles Resolved Within Estimate
General	GE-1	Percent System Availability
	GE-2	Mean Time to Answer Calls
	GE-3	Call Abandonment Rate
Billing	BI-1	Mean Time to Provide Recorded Usage Records
	BI-2	Mean Time to Deliver Invoices
	BI-3	Percent Invoice Accuracy
	BI-4	Percent Usage Accuracy
Operator Services and Directory Assistance	OSDA-1	Mean Time to Answer
Network Performance	NP-1	Network Performance Parity
Interconnect / Unbundled Elements and Combos	IUE-1	Function Availability
	IUE-2	Timeliness of Element Performance

The Service Quality Measurements document describes the importance of each measure as an indicator of service parity. The SQM document also describes reporting dimensions that will be used to break each measure out by like factors (*e.g.*, major service group).

Why We Need to Use Statistical Tests

The Telecommunications Act of 1996 requires that ILECs provide nondiscriminatory support regardless of whether the CLEC elects to employ interconnection, services resale, or unbundled network elements as the market entry method. It is essential that CLECs and regulators be able to determine whether ILECs are meeting these parity and nondiscriminatory obligations. In order to make such a determination, the ILEC's performance for itself must be compared to the ILEC's performance in support of CLEC operations; and the results of this comparison must demonstrate that the CLEC receives no less than equal treatment compared to that the ILEC provides to its own operations. Where a direct comparison to analogous ILEC performance is not possible, the comparative standard is the level of performance that offers an efficient CLEC a meaningful opportunity to compete.

When making the comparison of ILEC results to CLEC results, it is necessary to employ comparative procedures that are based upon generally accepted statistical procedures. It is important to use statistical procedures because all of the ILEC-CLEC processes that will be measured are processes that contain some degree of randomness. Statistical procedures recognize that there is measurement variability, and assist in translating results data into useful decision-making information. A statistical approach allows for measurement variability while controlling the risk of drawing an inappropriate conclusion (*i.e,* a "type 1" or "type 2" error, discussed in the next section).

Basic Concepts and Terms

Populations and Samples

Statistical procedures will permit a determination whether the support that the ILECs provide to CLECs is indistinguishable from the support provided by the ILECs to their own customers. In statistical terms, we will determine whether two "samples", the ILEC sample and the CLEC sample, come from the same "population" of measurements.

The procedures described in this paper are based on the following assumption: *When parity is provided, the ILEC data and CLEC data can both be regarded as samples from a common population of possible outcomes.* In other words, if parity exists, the measured results for a CLEC should not be distinguishable from the measured results for the ILEC, once random variability is taken into account. Figure 1 illustrates this concept. On the right side of the figure are histograms of two samples. In this illustration, the ILEC sample contains 200 observations (data values) and the CLEC sample contains 50. Note that the two histograms are not exactly alike. This is due to sampling variation. The assumption that parity exists implies that both samples were drawn from the same population of values. If it were possible to observe this population completely, the population histogram might appear as shown on the left of the Figure. If the samples were indeed taken from this population, histograms drawn for larger and larger samples would look more and more like the population histogram. Figure 1 shows that even when parity is being provided, there will be differences between the samples due to sampling variability. Statistical tests quantify the differences between the two samples and make proper allowance for sampling variability. They assess the chance that the differences that are observed are due simply to sampling variability, if parity is being provided.

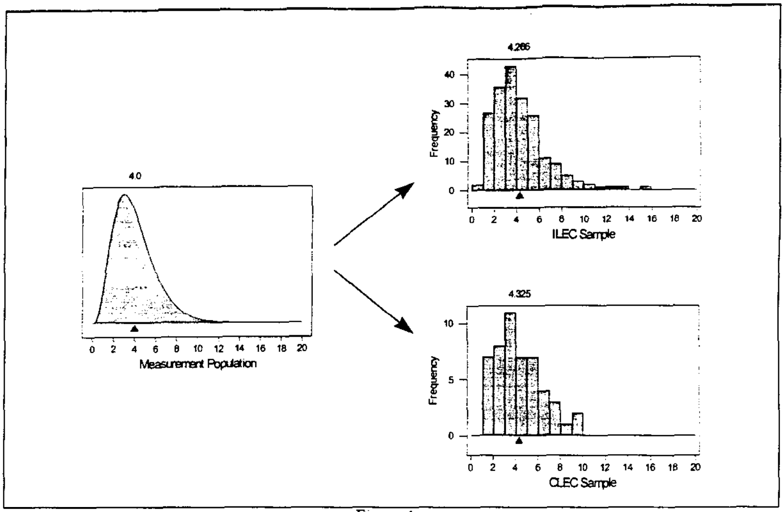

Figure 1.

Measures of Central Tendency and Spread

Often, distributions are summarized using "statistics." For the purpose of this paper, a "statistic" is simply a calculation performed on a sample set of data. Two common types of statistics are known as measures of "central tendency" and "spread."

A measure of central tendency is a summary calculation that describes the middle of the distribution in some way. The most common measure of central tendency is called the "mean" or "average" of the distribution. The mean of a sample is simply the sum of the data values divided by the sample size (number of observations). Algebraically, this calculation is expressed as

$$\bar{x} = \frac{\sum x}{n},$$

where x denotes a value in the sample and n denotes the sample size. The mean describes the center of the distribution in the following way: *If the histogram for a sample were a set of weights stacked on top of a flat board placed on top of a fulcrum (a "see-saw"), the mean would be the position along the board at which the board would balance.* (See Figure 1.) The mean in Figure 1 is indicated by the small triangle at approximately the value "4" on the horizontal axis.

A measure of spread is a summary calculation that describes the amount of variation in a sample. A common measure of spread is a called the "standard deviation" of the sample. The standard deviation is the typical size of a deviation of the observations in the sample from their mean value. The standard deviation is calculated by subtracting the mean value from each observation in the sample, squaring the resulting differences (so that negative and positive differences don't offset), summing the squared differences, dividing the sum by one less than the sample size, then taking the square root of the result. Algebraically, this calculation is expressed as

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}.$$

While the notion of mean and standard deviation exists for populations as well as samples, the mathematical definition for the mean and standard deviation for populations is beyond the scope of this paper. However, their interpretation is generally the same as for samples. In fact, for very large samples, the sample mean and sample standard deviation will be very close to the mean and standard deviation of the population from which the sample was taken.

Sampling Distribution of the Sample Mean

In Figure 1 we showed the positions of the means of the population and the two samples with triangular symbols beneath the distributions. If we sample over successive months, we will get new ILEC samples and new CLEC samples each and every month. These samples will not be exactly like the one for the first month; each will be influenced by sampling variability in a different way. In Figure 2, we show how sets of 100 successive ILEC means and 100 successive CLEC means might appear. The ILEC means can be thought of as being drawn from a population of sample means; this population is called the "sampling distribution" of these ILEC means. This sampling distribution is completely determined by the basic population of measurements that we start with, and the number of observations in each sample. The sampling distribution has the same mean as the population.

Figure 2 illustrates two important statistical concepts:

1. The histogram of successive sample means resembles a bell-shaped curve known as the Normal Distribution. This is true even though the individual observations came from a skewed distribution.

2. The standard deviation of the distribution of sample means is much smaller than the standard deviation of the observations themselves. In fact, statistical theory establishes the fact that the standard deviation on the population of means is smaller by a factor $\sqrt{n}$, where n is the sample size. This effect can be seen in our example: the distribution of the CLEC means is twice as broad as the distribution of the ILEC means, since the ILEC sample size (200) is four times as large as the CLEC sample size (50).

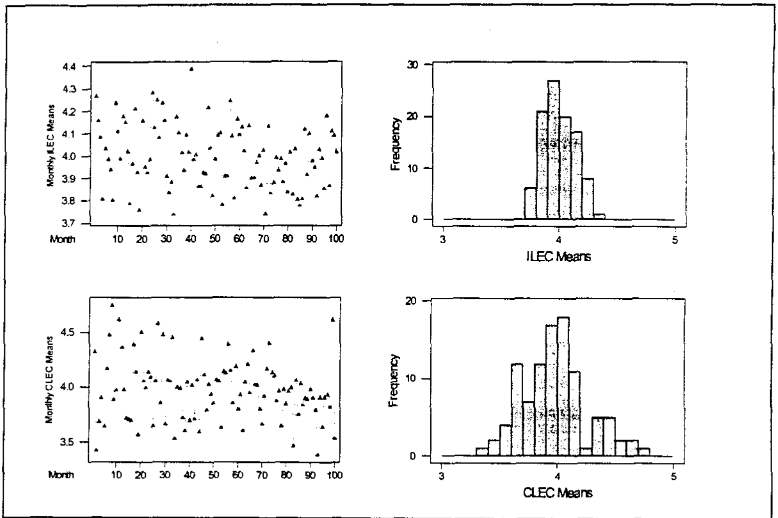

Figure 2.

It is common to call the standard deviation of the sampling distribution of a statistic the "standard error" for the statistic. We shall adopt this convention to avoid confusion between the standard deviation of the individual observations and the standard deviation (standard error) of the statistic. The latter is generally much smaller than the former. In the case of sample means, the standard error of the mean is smaller than the standard deviation of the individual observations by a factor of $\sqrt{n}$.

The Z-test

Our objective is to compare the mean of a sample of ILEC measurements with the mean of a sample of CLEC measurements. Suppose both samples were drawn from the same population; then the difference between these two sample means (*i.e.*, $DIFF = \bar{x}_{\text{CLEC}} - \bar{x}_{\text{ILEC}}$) will have a sampling distribution which will

(i) have a mean of zero; and
(ii) have a standard error that depends on the population standard deviation and the sizes of the two samples.

Statisticians utilize an index for comparing measurement results for different samples. The index employed is a ratio of the difference in the two sample means (being compared) and the standard deviation estimated for the overall population. This ratio is known as a z-score. The z-score compares the two samples on a standard scale, making proper allowance for the sample sizes.

The computation of the difference in the two sample means is straightforward.

$$DIFF = \bar{x}_{\text{CLEC}} - \bar{x}_{\text{ILEC}}$$

The standard deviation is less intuitive. Nevertheless, statistical theory establishes the fact that

$$\sigma^2_{\text{DIFF}} = \frac{\sigma^2}{n_{\text{CLEC}}} + \frac{\sigma^2}{n_{\text{ILEC}}},$$

where is the standard deviation of the population from which both samples are drawn. That is, the squared standard error of the difference is the sum of the squared standard errors of the two means being compared.[1]

We do not know the true value of the population because the population cannot be fully observed. However, we can estimate given the standard deviation of the ILEC sample ($_{\text{ILEC}}$).[2] Hence, we may estimate the standard error of the difference with

$$\sigma_{\text{DIFF}} = \sqrt{\frac{\sigma^2_{\text{ILEC}}}{n_{\text{CLEC}}} + \frac{\sigma^2_{\text{ILEC}}}{n_{\text{ILEC}}}} = \sqrt{\sigma^2_{\text{ILEC}}\left[\frac{1}{n_{\text{CLEC}}} + \frac{1}{n_{\text{ILEC}}}\right]}$$

If we then divide the difference between the two sample means by this estimate of the standard deviation of this difference, we get what is called a "*z*-score".

---

1 Winkler and Hays, *Probability, Inference, and Decision*. (Holt, Rinehart and Winston: New York), p. 370.

2 Winkler and Hays, *Probability, Inference, and Decision*. (Holt, Rinehart and Winston: New York), p. 338.

$$z = \frac{DIFF}{\sigma_{\text{DIFF}}}$$

Because we assumed that both samples were in fact drawn from the same population, this *z*-score has a sampling distribution that is very nearly Standard Normal, *i.e.*, having a mean of zero and a standard error of one. Thus, the *z*-score will lie between ± 1 in about 68% of cases, will lie between ± 2 in about 95% of cases, and will lie between ± 3 in about 99.7% of cases, always assuming that both samples come from the same population. Therefore, one possible procedure for checking whether both samples come from the same population is to compare the *z*-score with some cut-off value, perhaps +3. For comparisons where the values of z exceed the cutoff value, you reject the assumption of parity as not proven by the measured results. This is an example of a statistical test procedure. It is a formal rule of procedure, where we start with raw data (here two samples, ILEC measurements and CLEC measurements), and arrive at a decision, either "conformity" or" violation".

Type 1 Errors and Type 2 Errors

Each statistical test has two important properties. The first is the probability that the test will determine that a problem exists when in fact there is none. Such a mistaken conclusion is called a type one error. In the case of testing for parity, a type one error is the mistake of charging the ILEC with a parity violation when they may not be acting in a discriminatory manner. The second property is the probability that the test procedure will not identify a parity violation when one does exist. The mistake of not identifying parity violation when the ILEC is providing discriminatory service is called a type two error. A balanced test is, therefore, required.

From the ILEC perspective, the statistical test procedure will be unacceptable if it has a high probability of type one errors. From the CLEC perspective, the test procedure will be unacceptable if it has a high probability of type two errors.

Very many test procedures are available, all having the same probability of type one error. However the probability of a type two error depends on the particular kind of violation that occurs. For small departures from parity, the probability of detecting the violation will be small. However, different test procedures will have different type two error probabilities. Some test procedures will have small type two error when the CLEC mean is larger than the ILEC mean, even if the CLEC standard deviation is the same as the ILEC standard deviation, while other procedures will be sensitive to differences in standard deviation, even if the means are equal. Our proposals below are designed to have small type two error when the CLEC mean exceeds the ILEC mean, whether or not the two variances are equal.

Tests of Proportions and Rates

When our measurements are proportions (*e.g.* percent orders completed on time) rather than measurements on a scale, there are some simplifications. We can think of the "population" as being analogous to an urn filled with balls, each labeled either 0(failure) or 1(success). In this population, the fraction of 1's is some "population proportion". Making an observation corresponds to drawing a single ball from this urn. Each month, the ILEC makes some number of observations, and reports the ratio of failures or successes to the total number of observations; the ILEC does the same does the same for the CLEC. The situation is very similar to that discussed above; however, rather than a wide range of possible result values, we simply have 0's (failures) and 1's (successes). The "sample mean" becomes the "observed proportion", and this will have a sampling distribution just as before. The novelty of the situation is that now the population standard deviation is a known function of the population proportion[3]; if the population proportion is p, the population standard deviation is $\sqrt{p(1-p)}$, with similar simplifications in all the other formulas.

There is a similar simplification when the observations are of rates, *e.g.*, number of troubles per 100 lines. The formulas appear below.

Proposed Test Procedures

Applying the Appropriate Test

Three z-tests will be described in this section: the "Test for Parity in Means", the "Test for Parity in Rates", and the "Test for Parity in Proportions". For each LCUG Service Quality Measurement (SQM), one or more of these parity tests will apply. The following chart is a guide that matches each SQM with the appropriate test.

Measurement (Corresponding LCUG Number)	*Test*
Preordering Response Interval (PO-1)	**Mean**
Avg. Order Completion Interval (OP-1)	**Mean**
% Orders Completed On Time (OP-2)	**Proportion**
% Order (Provisioning) Accuracy (OP-3)	**Proportion**

[3] Winkler and Hays, *Probability, Inference, and Decision*. (Holt, Rinehart and Winston: New York), p. 212.

Order Reject Interval (OP-4)	**Mean**
Firm Order Confirmation Interval (OP-5)	**Mean**
Mean Jeopardy Interval (OP-6)	**Mean**
Completion Notice Interval (OP-7)	**Mean**
Percent Jeopardies Returned (OP-8)	**Proportion**
Held Order Interval (OP-9)	**Mean**
% Orders Held ≥ 90 Days (OP-10)	**Proportion**
% Orders Held ≥ 15 Days (OP-11)	**Proportion**
Time To Restore (MR-1)	**Mean**
Repeat Trouble Rate (MR-2)	**Proportion**
Frequency of Troubles (MR-3)	**Rate**
Estimated Time To Restore (MR-4)	**Proportion**
System Availability (GE-1)	**Proportion**
Center Speed of Answer (GE-2)	**Mean**
Call Abandonment Rate (GE-3)	**Proportion**
Mean Time to Deliver Usage Records (BI-1)	**Mean**
Mean Time to Deliver Invoices (BI-2)	**Mean**
Percent Invoice Accuracy (BI-3)	**Proportion**
Percent Usage Accuracy (BI-4)	**Proportion**
OS/DA Speed of Answer (OS/DA-1)	**Mean**
Network Performance (NP-1)	**Mean, Proportion**
Availability of Network Elements (IUE-1)	**Mean, Proportion**
Performance of Network Elements (IUE-2)	**Mean, Proportion**

Test for Parity in Means

Several of the measurements in the LCUG SQM document are averages (*i.e.*, means) of certain process results. The statistical procedure for testing for parity in ILEC and CLEC means is described below:

1. Calculate for each sample the number of measurements (n_{ILEC} and n_{CLEC}), the sample means ($\bar{x}_{ILEC}$ and $\bar{x}_{CLEC}$), and the sample standard deviations ($_{ILEC}$ and $_{CLEC}$).

2. Calculate the difference between the two sample means; if *larger* CLEC mean indicates possible violation of parity, use *DIFF* = $\bar{x}_{CLEC}$ - $\bar{x}_{ILEC}$, otherwise reverse the order of the CLEC mean and the ILEC mean.

3. To determine a suitable scale on which to measure this difference, we use an estimate of the population variance based on the ILEC sample, adjusted for the sized of the two samples: this gives the standard error of the difference between the means as

$$\sigma_{DIFF} = \sqrt{\sigma^2_{ILEC}\left[\frac{1}{n_{CLEC}} + \frac{1}{n_{ILEC}}\right]}$$

4. Compute the test statistic

$$z = \frac{DIFF}{\sigma_{DIFF}}$$

5. Determine a critical value c so that the type one error is suitably small.

6. Declare the means to be in violation of parity if $z > c$.

Example:

c:	3.58	Critical value for the test

ILEC			CLEC			Test	
n	mean	variance	n	mean	variance	z	Violation
250	4.038	1.9547	50	5.154	23.2035	5.15	YES!

Test for Parity in Proportions

Several of the measurements in the LCUG SQM document are proportions derived from certain counts. The statistical procedure for testing for parity in ILEC and CLEC proportions is described below. It is the same as that for means, except that we do not need to estimate the ILEC variance separately.

1. Calculate for each sample sample sizes (n_{ILEC} and n_{CLEC}), and the sample proportions (p_{ILEC} and p_{CLEC}).

2. Calculate the difference between the two sample means; if *larger* CLEC proportion indicates worse performance, use $DIFF = p_{CLEC} - p_{ILEC}$, otherwise reverse the order of the ILEC and CLEC proportions.

3. Calculate an estimate of the *standard error for the difference* in the two proportions according to the formula

$$\sigma_{DIFF} = \sqrt{p_{ILEC}(1 - p_{ILEC})\left[\frac{1}{n_{CLEC}} + \frac{1}{n_{ILEC}}\right]}$$

4. Hence compute the test statistic

$$z = \frac{DIFF}{\sigma_{DIFF}}$$

5. Determine a critical value c so that the type one error is suitably small.

6. Declare the means to be in violation of parity if $z > c$.

Example:

c:	3.58	Critical value for the test

ILEC			CLEC			Test	
num	den	p	num	den	p	z	Violation
5	250	2.00%	7	40	17.50%	6.50	YES!

Test for Parity in Rates

A rate is a ratio of two counts, *num/denom*. An example of this is the trouble rate experience for POTS. The procedure for analyzing measurements results that are rates is very similar to that for proportions.

1. Calculate the numerator and the denominator counts for both ILEC and CLEC, and hence the two rates $r_{ILEC} = num_{ILEC}/denom_{ILEC}$ and $r_{CLEC} = num_{CLEC}/denom_{CLEC}$.

2. Calculate the difference between the two sample rates; if *larger* CLEC rate indicates worse performance, use $DIFF = r_{CLEC} - r_{ILEC}$, otherwise take the negative of this.

3. Calculate an estimate of the *standard error for the difference* in the two rates according to the formula

$$\sigma_{DIFF} = \sqrt{r_{ILEC}\left[\frac{1}{denom_{CLEC}} + \frac{1}{denom_{ILEC}}\right]}$$

4. Compute the test statistic

$$z = \frac{DIFF}{\sigma_{DIFF}}$$

5. Determine a critical value c so that the type one error is suitably small.

6. Declare the means to be in violation of parity if $z > c$.

Example:

c:	3.58	Critical value for the test

ILEC			CLEC			Test	
num	den	rate	num	den	rate	z	Violation
250	610	0.409836	34	30	1.133333	6.04	YES!

Exhibit RMB-4

"Statistical Tests for Local Service Parity," Version 1.0, February 6, 1998, Local Competition Users Group

Permutation Analysis Procedural Steps

Permutation analysis is applied to calculate the z-statistic using the following logic:

1. Choose a sufficiently large number T.

2. Pool and mix the CLEC and ILEC data sets

3. Randomly subdivide the pooled data sets into two pools, one the same size as the original CLEC data set (n_{CLEC}) and one reflecting the remaining data points, (which is equal to the size of the original ILEC data set or n_{ILEC}).

4. Compute and store the Z-test score (Z_S) for this sample.

5. Repeat steps 3 and 4 for the remaining T-1 sample pairs to be analyzed. (If the number of possibilities is less than 1 million, include a programmatic check to prevent drawing the same pair of samples more than once).

6. Order the Z_S results computed and stored in step 4 from lowest to highest.

7. Compute the Z-test score for the original two data sets and find its rank in the ordering determined in step 6.

8. Repeat the steps 2-7 ten times and combine the results to determine P = (Summation of ranks in each of the 10 runs divided by 10T)

9. Using a cumulative standard normal distribution table, find the value Z_A such that the probability (or cumulative area under the standard normal curve) is equal to P calculated in step 8.

10. Compare Z_A with the desired critical value as determined from the critical Z table. If $Z_A >$ the designated critical Z-value in the table, then the performance is non-compliant.