

ORIGINAL



0000007420

BEFORE THE ARIZONA CORPORATION COMMISSION

CARL J. KUNASEK
Chairman
JAMES M. IRVIN
Commissioner
WILLIAM A. MUNDELL
Commissioner

RECEIVED
AZ CORP COMMISSION

JAN 18 2 20 PM '00
Arizona Corporation Commission
DOCKETED

DOCUMENT CONTROL
JAN 18 2000

IN THE MATTER OF U S WEST
COMMUNICATIONS, INC.'S
COMPLIANCE WITH § 271 OF THE
TELECOMMUNICATIONS ACT OF 1996

Docket No. T-00000A-97-0238
COMMENTS OF AT&T AND TCG
PHOENIX ON CAP GEMINI'S
STATISTICAL PLAN

On December 30, 1999, Cap Gemini Telecom ("CGT") distributed by e-mail its Proposed Statistical Approach and Contrast of Different Statistical Approaches to 271 Parity/Compliance. AT&T Communications of the Mountain States, Inc. and TCG Phoenix (collectively, "AT&T") responded by e-mail on January 6, 2000. Attached are AT&T's comments to the documents received from CGT (attached as Exhibits A & B respectively). AT&T's comments are incorporated in the CGT documents and are underlined for ease of reference.

RESPECTFULLY SUBMITTED this 17th day of January, 2000.

AT&T COMMUNICATIONS OF
THE MOUNTAIN STATES, INC.
AND TCG PHOENIX

By:

Thomas C. Pelto
Mary B. Tribby
Richard S. Wolters
1875 Lawrence Street
Suite 1575
Denver, Colorado 80202
Telephone: 303-298-6471
Facsimile: 303-298-6301
E-mail: rwolters@att.com

Statistical Validity of OSS Compliance / Parity Results

CGT Proposed Statistical Approach

1. Relevant Statistical Issues

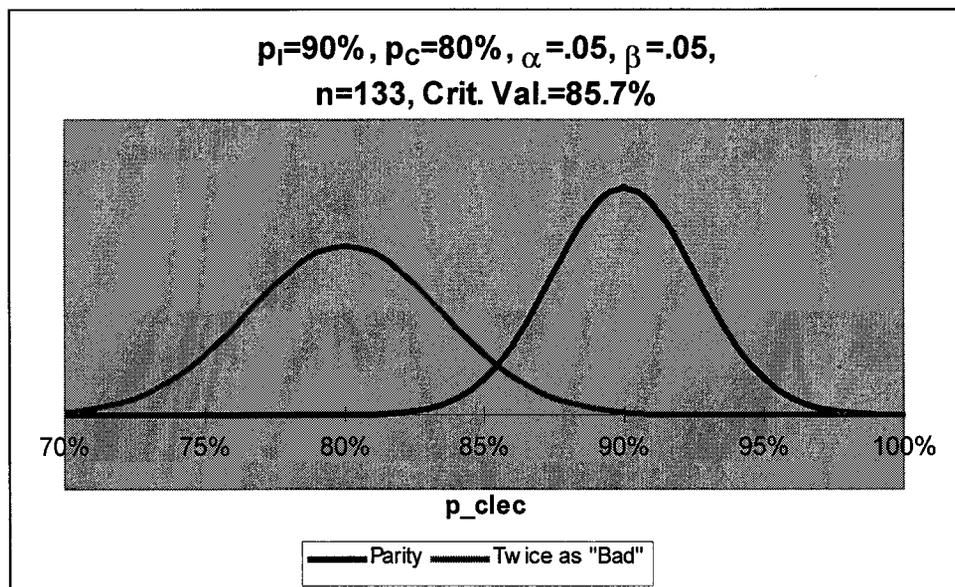
1.1. Proving Beyond Some Reasonable Level of Doubt

For evidence to constitute establishing something, there should only be a small probability of the evidence having occurred under the converse hypothesis. Typically, this probability level is called the significance level of a test and a value of .05 is used. In various discussions between the ILECs and CLECs, a value of .05 or .15 is typically suggested, depending on which side is at risk for the question at hand.

1.2. Direction of Hypotheses

When the level of significance is the smaller risk involved in a hypothesis test, the hypotheses are constructed such that the alternate hypothesis is that which we are trying to establish. Its converse, the null hypothesis, is never really established, it is rather assumed as what we will “conclude” by default in the absence of significant evidence to the contrary.

CGT maintains that by specifying that, in OSS testing, the burden of proof is on the ILECs to establish non-discrimination, the FCC has essentially said that without such tests, the ILECs are not given the benefit of the doubt. As such, parity / compliance must not be the default assumption we fall back on absent significant evidence to the contrary. Rather parity/compliance must have significant evidence supporting it, and therefore ought to be the alternate hypothesis. AT&T strongly supports the concept of disparity/noncompliance as the default assumption for statistical testing. As CGT appropriately recognizes, that



statistical concept aligns quite nicely with the FCC's statement that, "[w]e emphasize, however, that the BOC applicant retains at all times the ultimate burden of proof that its application satisfies section 271." (Ameritech Michigan Order, ¶ 44) In other words, U S WEST must prove parity/compliance. The statistical concept proposed by CGT of making U S WEST statistically prove parity/compliance is a sound concept that is supported by FCC precedent.

1.3. Statistical Power (Sensitivity) of Test – at a Materially Different Alternative, and Relationship to Sample Size (Benchmark Case)

The power of the test is the probability of rejecting the null hypothesis and thereby establishing the alternative hypothesis, evaluated at some materially different value (within

the alternate hypothesis region). Suppose for instance that we wish to establish that performance on CLEC orders for a specific measure is compliant with its stated benchmark success rate of 90%. Our hypotheses then ought to be as follows:

$$H_0: p_c < 90\%$$
$$H_A: p_c \geq 90\%$$

AT&T believes that given CGT's intent, the hypotheses could be stated in a manner that communicates more information. AT&T proposes that to communicate more information, the illustrative hypotheses should be stated as:

$$H_0 : p_c < 90\%$$
$$H_A : p_c \geq 95\%$$

This format communicates clearly to all the parties how the test will be defined without the additional step of describing how to calculate β . A test that fails to reject H_0 in favor of H_A would indicate significantly positive evidence that CLEC performance is worse than 95% ("twice as good" as 90%). This is exactly how CGT interprets the same result in the last sentence of the penultimate paragraph of section 1.4. In other words, CGT interpret results as if it were using AT&T's proposed hypotheses rather than their own.

Absent specific practical information on what is a practically meaningful materially different value, CGT proposes using "twice as good", i.e., half as many failures, as the materially different value at which to evaluate the power of the test. This approach views 90% as equally better than 80% as is 98% than 96%, and is an attempt to balance a desire for making the alternate value as close to the benchmark while recognizing the practical implications upon sample size. It also is chosen in the interests of generality. However it has its limitations and the CGT Statistics Team is proposing it out of ignorance of material difference requirements for specific measures which may yet be negotiated between the CLECs and ILEC, and could supersede the "twice as good" value standard proposed here (with implications on required sample size).

In the above paragraph, CGT discusses the "twice as good" concept as a means of defining practically meaningful material differences from specified benchmarks. Admittedly, CGT does not know whether "twice as good" does indeed imply a "meaningful material difference." The "meaningful material difference" concept is the correct one to apply. Rather than universally using the "twice as good" concept, AT&T proposes that a meaningful material difference be determined on a measure by measure basis through the collaborative process.

Within this framework we now have two possible errors to consider:

Type I Error: Rejecting H_0 , Non-Compliance, and declaring that Compliance has been established, when in fact the ILEC is Non-Compliant with the acceptable performance level indicated by the agreed-upon benchmark for the measure.

Type II Error: Failing to Reject H_0 , Non-Compliance, and not declaring that Compliance has been established, when in fact the ILEC is Compliant and even surpasses the acceptable performance level indicated by the agreed-upon benchmark to the extent of providing performance "twice as good" as the benchmark.

Making an error of Type I is of concern to the CLECs in that (a) they are not being provided with competitively acceptable service, and (b) the ILEC will thereby be allowed to enter a market which has been de-monopolized and compete with the CLECs (many of whom

provide long distance service) there on a level playing field, while maintaining some aspect of their own monopoly in local service and taking advantage of it.

Making an error of Type II is of concern to the ILEC in that they may be providing a level playing field to their competitors in local service and have spent substantial money on OSS testing to establish this, yet since the test failed to establish compliance, they are not allowed to compete in the long-distance market without further OSS testing.

While CGT agrees with the CLECs that the probability of Type I Error in this framework be held to no greater than 1 in 20, i.e., $\alpha=.05$, CGT also agrees with the ILEC that β , the probability of Type II Error also ought to be held low, especially as multiple measures will be tested and the burden of retesting would otherwise be severe.

It therefore seems reasonable to require sufficient sample size where feasible to ensure that both α and β be no greater than .05.

In this situation then the rejection region of the test would have the form:

$$\text{CLEC ontime orders / total CLEC orders} > .90 + 1.645 * \text{sqrt}(.9 * (1-.9) / n)$$

At the “twice as good” alternative value of .95, achieving power of .95 with the same critical value yields:

$$.95 - 1.645 * \text{sqrt}(.95 * (1-.95) / n).$$

These two expressions for the same critical value yield the following equation for the required sample size:

$$N = ((1.645 * \text{sqrt}(.9 * (1-.9)) + 1.645 * \text{sqrt}(.95 * (1-.95))) / (.95-.90)) ** 2 = 291.$$

Plugging this back into either equation for the rejection region yields a critical value of 92.9%. A CLEC sample result at least this high would then indicate significantly positive evidence of performance at least as good as the agreed upon benchmark. A result below this critical value would indicate significantly positive evidence that CLEC performance is worse than “twice as good” as the benchmark.

The critical value being higher than the benchmark is required in order to positively establish that CLEC performance is tuned to a level which meets or surpasses the benchmark. It is not onerous to the ILEC to require this in an OSS Test, because the benchmarks have been developed and negotiated with the concept that they are not process targets, but rather values of such minimally acceptable performance that performance any worse calls for immediate corrective action. Undoubtedly then the process target means are themselves far beyond the benchmark level of performance.

1.4. Critical Value which is Better than Parity Overly Onerous to ILEC

If we were to use the above approach to attempt to establish parity of CLEC performance with ILEC retail analog, we would then be requiring that in our sample, the CLEC results substantially surpass the ILEC’s results when serving their own retail customers. Such a requirement would be overly onerous to the ILEC as it would fail to establish parity 95% of the time when exact parity exists. As such, we would forever be retesting and it would take a tremendously long time until parity would be established.

In the above paragraph, CGT flip-flops the hypotheses making parity the null and disparity the alternative hypothesis. AT&T believes that this is completely unnecessary and contrary to the burden of proof legal standard identified by the FCC. CGT acknowledges that an exactly equivalent set of hypotheses could be defined in which disparity is the null

hypothesis and parity is the alternative hypothesis. Since there are several policy and statistical issues that AT&T believes need to be negotiated for each measure, it will be less confusing for all parties if hypotheses are defined in a consistent manner across all measures. This way each party is always concerned about α or β , and not concerned about α on some tests and β on other tests. Furthermore, test results should always be interpreted in the same way. That is, parity is established when the null hypothesis is rejected and not when rejected in some cases and failed to be rejected in other cases. Additionally, CGT made a strong argument that the burden of proof is on U S WEST, so AT&T fails to understand why the flip-flop is necessary. To retain consistent hypotheses and to maintain compliance with the FCC's burden of proof standard, AT&T proposes that the parity hypotheses be redefined as follows:

$H_0 : p_c \leq p_I - (1 - p_I)$ Disparity ("twice as bad")

$H_A : p_c \geq p_I$ Parity

A simpler, but equivalent parameterized version could be:

$H_0 : p_c \leq f \times p_I$ Disparity ("only fraction f as good")

$\neg H_A : p_c \geq p_I$ Parity

In support of AT&T's earlier proposal to negotiate the level that represents a meaningful material difference, the parties can negotiate over what value for the fraction f represents a meaningful material difference.

Therefore, we are forced to use hypotheses in the direction opposite to that which we specified earlier:

$H_0: p_c \geq p_i$ (Parity)
 $H_A: p_c < p_i$ (Dis-parity)

Now in order to withstand our previously stated objection to this approach, namely that we are by default assuming that which it is our objective to prove, and non-evidence of dis-parity is not significant evidence of parity, we can do the following:

Restrict β , the probability of Type II Error to .05 at a materially different alternative – analogously to above, in the absence of measure-specific values negotiated between the ILEC and CLECs, CGT would propose “twice as bad” as the materially different alternative. For instance, if ILEC performance for the retail analog was a 90% success rate, the materially disparate CLEC performance for which we would require no greater than .05 probability of accepting the null hypothesis of parity would be an 80% success rate.

By keeping the probability of Type II Error this low, we can then at least say that non-evidence of dis-parity is significantly positive evidence. While it is not exactly significantly positive evidence of parity, it is significantly positive evidence of “performance which is at least no worse than twice as bad as ILEC performance”.

Furthermore, we can still restrict α , the rate of false declarations of dis-parity to .05, keeping the risk to the ILEC at a similarly acceptable level.

(As Dr. Mount-Campbell observed, the above formulation can be considered essentially equivalent to using a null hypothesis value of “twice as bad as parity”, and an alternative hypothesis of “better than twice as bad as parity”, and evaluating β , the probability of Type II Error at the materially different alternative of “parity”).

Carrying through the above example, if retail performance is at 90% orders completed on-time, β would be restricted to .05 at the materially different alternative of 80% orders completed on-time. Together with restricting α to .05, we can solve for the required sample size, which in this case is 133, and the critical value, 85.7%.

1.5. Distributional Assumptions

In the above discussion, we have focused on binomial measures, percentages of success. In the case of interval measures, such as time to firm order confirmation, or mean time to restore service, there are two additional issues. The first of these is that not only the performance level, but also the variability of performance needs to be taken into account. Secondly, the standard approach of using the t -test requires that the underlying distribution be normal, and that the variances of the distributions be equal. While certain departures from normality are tolerable as long as the underlying distributions are symmetric and variance does not change with the mean, even these minimal requirements tend not to be satisfied with interval-type measures. Instead, interval measures tend to be prominently skewed, and their standard deviation tends to be proportional to the mean. In some cases with our measures, they also exhibit multi-modality (several peaks instead of one at the center).

An approach which would be reasonably robust to all (except perhaps the last) of these assumptions violations would be to assume that the underlying distributions are lognormal, and therefore transform the original interval data by taking its logarithm before applying standard statistical techniques.

Additionally, we can then generalize the above approach for binomials to interval measures as follows:

For benchmark measures, such as mean time to FOC for flow-through orders within 20 minutes, use the coefficient of variation of the measure (standard deviation divided by the mean, this is σ , the scale parameter of the lognormal) to determine where the median of the lognormal distribution is: $\text{median} = \text{mean} / \exp(\sigma^2/2)$. (The log of the median is the location parameter of the lognormal distribution). If σ is 100%, then performance at exactly the level of the benchmark corresponds to a median FOC time of 12:07 minutes. This means that 50% of FOC times would be longer than 12:07 minutes. Performance "twice as good" would only have 25% of FOC times **longer than** 12:07 minutes. This corresponds to a lognormal with median 6:11 mins. Setting $\alpha = .05$ and $\beta(6:11) = .05$ results in a sample size requirement of 24, and a critical value corresponding to a median of 8:40 minutes, which corresponds to a mean of 14:18 mins.

For parity measures, such as average interval to order completion for a particular order-type, suppose the retail average is 11.25 days with a standard deviation of 7.5 days, ie. c.v. = 67%. Then this corresponds to a median of 8.9 days. Performance "twice as bad" would have only 25% of completion times **shorter than** 8.9 days. This corresponds to a lognormal whose median is 14.2 days. Setting $\alpha = .05$ and $\beta(8.9) = .05$ results in a sample size requirement of 24, and a critical value corresponding to a median of 11.25 days, which corresponds to a mean of 14.1 days.

While this generalized approach does not directly address the issue of multimodality, it could be argued that it is less sensitive to multimodality than the standard approach without the log transformation, because it is performing the standard analysis on a mixture of normal distributions rather than a mixture of skewed lognormals, whose multimodality, especially in the right tail of the distribution, will be particularly devastating to the validity of the analysis.

In this section AT&T believes that the sample discussions of calculations concerning the lognormal are not explained very clearly for the benchmark interval measures. For example, there is no indication of where a parameter n (sample size) is incorporated in the calculations so that it can be solved for when α and β are equated.

AT&T believes that the similar discussion for the parity test is less clear and does not seem to consider both U S WEST and CLEC sample sizes.

AT&T recommends that CGT should be more explicit about their exact proposed statistical methods and how the tests are to be conducted. AT&T believes that any transformation of data should be done to result in transformed data that is symmetrically distributed. AT&T agrees that the log transformation is one way of transforming data. AT&T submits that if a log transformation does not work well that a root transformation or some other transformation would. As long as the transformed data is symmetrically distributed, then a standard *t*-statistic or *z*-statistic could be applied to the transformed data. It is unclear if CGT's intent is to use a method to transform data and then use a *t*- or *z*-statistic to conduct the test.

If this is not CGT's intent, then AT&T requests that CGT compare the efficiency of their methods to the efficiency of the standard tests on the transformed data. It is the best interest of all parties to use the statistically most efficient methods possible.

AT&T agrees with CGT's conclusion that multi-modality will probably not be a big issue as long as the transformation makes the general shape of the distribution symmetric with a well shaped central tendency. Based on Monte Carlo studies that AT&T participated in of the modified *z*-test using the skewed and multi-modal Nevada data set, reliable parity testing was achieved as long as the smaller sample was in the neighborhood of 300. Logging the data could potentially improve considerably on that.

1.6. Required Sample Size

In the case of binomials, the required sample size increases as the performance level gets closer to 0% or 100%, according to the following table:

Parity/ Compliance Performance Level	Parity with Retail		Benchmark	
	Alternative	Sample Size	Alternative	Sample Size
50.00%	25.00%	38	75.00%	38
66.67%	33.34%	22	83.34%	70
75.00%	50.00%	38	87.50%	102
80.00%	60.00%	54	90.00%	133
85.00%	70.00%	80	92.50%	186
90.00%	80.00%	133	95.00%	291
95.00%	90.00%	291	97.50%	606
96.00%	92.00%	370	98.00%	764
97.00%	94.00%	501	98.50%	1027
98.00%	96.00%	764	99.00%	1553
99.00%	98.00%	1553	99.50%	3130
99.25%	98.50%	2078	99.63%	4181
99.50%	99.00%	3130	99.75%	6284
99.90%	99.80%	15747	99.95%	31519

Actually, in the case of parity, this calculation also ought to depend on the sample size for the retail analog. The numbers above are reasonably accurate as long as the retail volumes are at least an order of magnitude larger. Where the retail volumes are smaller, the required sample size for wholesale will need to be even substantially larger than indicated above. Due to the rapidly increasing sample sizes required as we get closer to 0% or 100% performance levels,- it may be appropriate in the interests of practicality, if we modified our concept of materially different alternate performance to “twice as bad but at least 3% worse” in the case of parity, and “twice as good, but at least 3% better” in the case of benchmarks. (This will prevent us from evaluating compliance with benchmarks higher than 97%). On the other hand, at levels of performance between 20% and 80% for parity and 40% and 60% for benchmarks, we may want to modify “twice as bad / good ” to “1.5 times as bad / good”, since we are more likely to have sufficient sample sizes there. These modifications might be further refined to be applied with a smooth transition function.

A simple model is often adequate for coming up with sample sizes, but when the data are actually collected a different method of testing may be needed, eg, the permutation test or an exact proportion test. Generally, the test would be done to control α at its specified value. This leaves β only approximately controlled because the sample size was found by a different model. This is one other reason why AT&T believes it is important that the null hypothesis always assumes “disparity.” Since the burden of proof is on U S WEST, having “disparity” as the null means that a little variability in the actual β is less troublesome.

AT&T checked, using exact methods, a few sample size values in the benchmark column of the table in section 1.6 and found them to be quite good with α and β generally differing by no more than 0.015.

In the paragraph after the table, CGT seems to admit that the retail analog sample size was not accounted for in the earlier example. CGT asserts that if its volume for the retail orders was an order of magnitude larger then the second sample size it can be ignored. AT&T is not sure if this assertion is true. It may be possible for sample size calculation. However, the assertion may not be correct when doing the actual test if system variance is large.

1.7. Cells

The above required sample sizes would be required on a per cell basis. This would mean that to whatever level of disaggregation each measure is to be evaluated at, we would require 133 orders of say, each product-type, dispatched, in an MSA.

A way that the required number of orders may be somewhat further reduced is by using more liberal Type I and Type II error probabilities, say 15%, on tests performed in individual cells, when several cells will be aggregated, and requiring the above stringent sample sizes on the aggregate tests only. This will require that we not make final pass-fail decisions on the basis of per-cell tests, but use them only to evaluate if there is a pattern to the failures which indicates problems with a particular product type, order type, region type, etc.

AT&T requests that CGT provide more explanation on the aggregate test.

1.8. Military-Style Testing

If military-style testing is being performed, we do not need to be overly concerned with the problems of multiple significance, as long as there is no “double-jeopardy” involved. That is, once a measure has passed, even if another measure fails partially or fully, retesting for

the failed measure (potentially after such improvements to the system which are not considered to adversely affect future results of the passed measure) will yield data both on the previously failed and previously passed measure. However, we will not fail the previously passed measure once we have previously passed it.

AT&T has some concerns about the military testing. Since a failure to establish parity means additional testing until parity is established, then that means effectively a much smaller value for β than originally planned. The exception to this is if system changes to correct problems can be documented after the first test. Then the second test is unrelated to the first test. Without such documentation then it would seem the effective α and β are not equal. This could argue for $\alpha < \beta$ in the initial design in anticipation of additional testing should proof of parity fail to appear.

1.9 Caveat

The above proposal is based on CGT's best current knowledge of Arizona's data. The statistical design, significance level, power, specification of material difference from parity / benchmark, the retail parity levels of performance, the interval measure distributions and their standard deviations, the appropriate levels of disaggregation for analyzing each measure, and the target order mix are all intertwined and together have major practical implications on the number of LSR's, ILEC and CLEC facilities, and friendly lines required, and upon the sample order mix. Consequently, more detailed knowledge of many of these aspects will be required before a final list of scenarios and iterations is constructed. Therefore CGT reserves the right to use its best statistical judgment in balancing these various requirements with what is feasible in the context of OSS testing in Arizona, both in finalizing the design and performing the subsequent analyses of generated data.

Given that Sections 2 – 7 are in outline form only, AT&T was unable to provide any meaningful comments. AT&T reserves the right to provide additional comments should CGT provide additional information for these sections.

2. **Design Strategy**
 - 2.1. **Looser Strategy for Per-Cell Proofs**
 - 2.2. **Sizing of Samples on a Per-Cell Basis with Distributional Assumptions, $\alpha=.05$, $\beta=.2$ for Benchmarks, $\alpha=.15$, $\beta=.05$ for Parity**
 - 2.3. **Distribution-Free Per-Cell Tests for Diagnostic Purposes**
 - 2.4. **Strictest Assumptions (Distribution-Free) for Aggregate Proofs, $\alpha=.05$, $\beta<.2$ for Benchmarks, $\alpha=.05$, $\beta=.05$ for Parity**
 - 2.5. **Random Mix of Per-Cell Failures: OK**
 - 2.6. **Pattern to Per-Cell Failures: Retesting Required**
 - 2.7. **Extreme Failure in Particular Cell: Retesting Required**
 - 2.8. **Multiple Significance**
3. **Benchmarks**
 - 3.1. **One-Sample Test**
 - 3.2. **Structured to Prove Compliance Beyond Reasonable Doubt**
 - 3.3. **Conclusion if Reject Null Hypothesis: Compliance Proven**
4. **Parity with Retail**
 - 4.1. **Two-Sample Test**
 - 4.2. **Structured to Prove Non-Parity Beyond Reasonable Doubt**
 - 4.3. **Conclusion if fail to reject Null Hypothesis: Not far from Parity (Increase Power?)**
 - 4.4. **Effectively switch directionality by switching sizes of α & β**
 - 4.5. **Where Retail Volume Permits, use only those Retail orders in same hour as Test LSRs**
5. **Problems with Binomial near $p=0$ or $p=1$.**
 - 5.1. **Need over 120 tests (per cell) to prove compliance with a benchmark of 97.5%**
 - 5.2. **Similar issue with parity when retail $p > .95$ or $<.05$.**
 - 5.3. **Try to define measures in terms of an aspect of the service which is more testable**
 - 5.4. **Are there sufficient retail volumes per cell ?**
6. **Data Requirements**

- 6.1. Actual service time for each LSR – both in test and retail**
- 7. Matrix of Minimum Numbers of Test Cases for each Product...**
 - 7.1. Need to know retail performance for each parity measure before being able to determine required sample sizes**
 - 7.2. Need to know mean and standard deviation of the log of each service time measure (whether parity or benchmark) before being able to construct test critical values.**
 - 7.3. Need to know facility constraints of CLECs and ILEC.**
 - 7.4. Need to know exactly which order types are counted for each measure and other associated business rules in order to construct an appropriate order mix which will efficiently achieve sufficient power for all practically testable sub-measures using minimum number of tests overall.**

AT&T generally agrees with CGT's analysis of the statistical approaches used in CA, NY, TX, PA and FL. AT&T did not have an opportunity to review in detail the statistical approaches in the referenced states or talk with AT&T representatives who participated in the other state proceedings. However, assuming that CGT has characterized the statistical features of those orders correctly, then AT&T believes that the CGT criticisms are right on target.

AT&T will offer some improvement suggestions to the CGT statistical approach in Arizona and believes that there are still many details to be worked out for Arizona statistical testing. However, AT&T's general impression is that CGT's statistical approach in Arizona is far superior to those of other third party testers in the eastern states.

AT&T will provide other minor comments in the body of this document.

Contrast of Different Statistical Approaches to 271 Parity / Compliance

States Examined: CA, NY, AZ, TX, PA, FLA

1. Introduction

CGT has proposed an approach to statistical design of OSS Testing which is perhaps novel in this field, yet is based on standard statistical methodology as practiced in all areas of application. A draft form of this approach, titled "ACC Design Concept," has been submitted to the ACC for its consideration in regards to CGT's implementation of it in its role as Test Administrator for Arizona's OSS test. The purpose of this document is to compare and contrast CGT's proposed approach with other approaches taken in 271 testing.

2. Purpose of OSS Test

While performing OSS Testing has many purposes, such as having an objective third party "live the CLEC experience", evaluating if the performance measures are being properly calculated, etc., from a legal and statistical perspective a primary purpose is "that the applicant BOC must demonstrate that it provides non-discriminatory access to its OSS in local service". In August 1997, the FCC's Ameritech Opinion analyzed the non-discriminatory access requirements of Sec 251(c) to a BOC's Sec 271 application and clarified that for those OSS subfunctions with retail analogs, a BOC "must provide access to competing carriers that is equal to the level of access that the BOC provides to itself, its customers, or its affiliates, in terms of quality, accuracy, and timeliness". For those OSS subfunctions with no retail analogs, a BOC must offer access sufficient to allow an efficient competitor "a meaningful opportunity to compete".

Both legally and statistically, then, a prime purpose of the test is demonstrate parity with retail analogs and compliance with agreed upon benchmarks.

3. Purposes of Some Other 271-Related Investigations

In some cases, investigations have been made of historical or current commercial data to determine whether parity / compliance existed in the past or exists currently. In these cases, no experimental design is constructed to determine what or how many orders to test; rather the analysis is based on whatever data exists. For this reason, in many of these cases statistical power was not controlled, because sample sizes could not be controlled, and were generally sufficient for many of the aggregate conclusions made. Yet, even in these cases, when it was desired to perform analyses in the most disaggregated form possible, and then aggregate these results upward, the recommendation was made by a joint team representing both the ILEC (Drs. S. Hinkins, E. Mulrow, and F. Scheuren of Ernst & Young LLP (consultants for BellSouth

Telecommunications)) and the CLECs (Dr. C. Mallows of AT&T Research), that: “The testing methodology should balance Type I and Type II Errors”¹.

This joint recommendation was made even in a situation where the purpose of the analysis was to determine significance of reported differences with a view towards exacting “incentive payments” for demonstrated sub-parity or sub-compliant service. In that kind of scenario, one could argue that the whole focus is only the statistical significance of the results, and that the power, or Type II error probability does not play as major a role, yet nonetheless, proper statistical practice does mandate its consideration. AT&T believes that CGT’s comments on statistical testing for ongoing performance monitoring are outside the scope of this OSS test. Therefore, AT&T believes that any comments on ongoing performance monitoring in this document are inappropriate.

4. OSS Test: A Designed Experiment

While looking at commercial data, particularly historically, provides two advantages, namely that blindness is well controlled, and that there are substantial volumes of retail and wholesale data to compare, nonetheless, running an OSS Test as a Controlled Scientific Experiment can provide several other potentially more valuable benefits – amongst these are:

- Greater ability to compare like-with –like.
- greater confidence that similarities or differences which turn up really provide significantly positive evidence of benchmark compliance / (closeness to) parity with retail, or of non-compliant / sub-parity performance
- precise control over the risks of both Type I and Type II Errors – thereby the ability to ensure that the test is fair to both sides, and that the evidence will be decisive.

5. The CGT Approach

CGT’s statistical approach is to let the purpose of the test, the requirements of significant decisiveness, and fairness to both sides define the statistical strategy chosen and implemented. If and when this runs into feasibility constraints, CGT typically explains the risks to the client and works with them to reach an approach with which as much of the stated goals as possible can be achieved within the constraints of feasibility.

CGT’s proposal directs hypotheses in accordance with what is to be demonstrated by the test as per its stated purpose, uses .05 chance of error as its standard criterion of significance, and equalizes the probability of both kinds of error, mistakenly rejecting noncompliance / parity when it is true, and mistakenly concluding noncompliance / parity when the true CLEC performance is materially better than compliant or worse than parity.

All of this is achievable because we can perform the OSS Test as a controlled scientific experiment. Since this has not been the focus of many previous 271 investigations, the approach appears somewhat novel in this field, but it is standard statistical practice.

6. Differences between the Arizona Approach vs. California Approach

California: The approach proposed in California by CGT is essentially the same as that proposed for Arizona with one exceptions: in California, the Parity Test is at the disaggregated cell level, CGT proposed

¹ Statistical Techniques For The Analysis and Comparison of Performance Measurement Data. Submitted to Louisiana Public Service Commission (LPSC), Docket U-22252 Subdocket C.

$\alpha=.15$ and $\beta=.05$ with the understanding that non-extreme disparities at the disaggregated level would not be viewed as definitive Pacific Bell failures, but rather we would expect several of them and consider greater numbers of them than expected or distinctive patterns among them as suggestive of areas on which to perform limited root-cause analysis as to the particular region, product, order-type, etc., which may need to be addressed regarding the specific measure. The greater emphasis on restricting the Type II Error in CA is due to the fact that the burden of proof is on the ILEC to establish (at least something approaching) parity, and CGT does not consider non-evidence of dis-parity to necessarily be sufficient evidence of parity.

Arizona: The approach CGT is currently proposing in Arizona equalizes the risks of Type I and Type II Error at the disaggregated cell test levels (both perhaps as high as .1 depending on sample size feasibility). This modification came about as a result of discussions with the ILEC's Performance Measures and Statistical representatives in which the ILEC insisted that its risk of not getting a pass on any test when parity is being provided be bounded by .05 and we were insisting that the risk of not getting a fail when performance is materially worse than parity (our standard for materially worse within OSS testing only and within the range of 80-95% good performance or 5-20% bad performance for the retail analog, is "twice as bad" as parity) be bounded by .05. For those cases where disaggregation would limit the feasible sample sizes, both sides felt they could compromise to the point of equal risks which may be higher than .05 on the disaggregated cell-level tests without giving up the .05 significance and power requirements on the aggregated tests.

7. The New York Approach

Please refer to <http://www.dps.state.ny.us/tel271.htm>

On that page, go to the bottom section and click on Appendix C for the "Statistical Approach". More detail is found in the results section, specifically Section H (POP8) of the PreOrder, Ordering, and Provisioning, Section IV, Part2 (Pages 151-314)

Comments: The approach used in NY differs from that proposed here in the following material ways:

- **Benchmarks:** No statistical hypothesis testing performed (ACC. To Appendix B of FCC report on NY State Approach – FCC does not make a statement on the appropriateness of this but leaves door open to test these statistically in future 271 applications).
- No control of power or Type II Error or proposal as to what constitutes a materially disparate / subcompliant performance level for each measure. *(Without relating at all to the appropriateness or not of the seeming lack of consideration of power and Type II Error in the NY decision, the FCC goes into a substantial discussion of these issues to little apparent context to NY Third Party OSS Testing in their Appendix B. One possible interpretation is that they want to see such issues addressed more thoroughly in future 271 applications.*

The following quote from the FCC NY decision (with FCC footnotes below) illustrates that they are thinking in the usual, general case, where sample size is not typically capable of being controlled, such as when looking at past commercial data to determine incentive payments:

“When we look at the differences in metric values, we will assume that parity exists unless the competitive LEC scores are worse than those for the BOC, and the difference is statistically significant at the 95 percent confidence level for a one-tailed test.² We use the 95 percent confidence level because it is a commonly used standard, and because it gives us a reasonable likelihood of detecting variations in performance not due to random chance, with few false conclusions that variations are not due to random chance.³ At the 95 percent confidence level,

² A difference in metric values that is statistically significant, however, does not necessarily mean that the BOC's service is discriminatory. We will examine the totality of the evidence before making a determination whether the BOC is providing parity.

³ Khazanie, *supra* n.4 at 506; Neter, Wasserman, and Whitmore, *supra* n.3 at 298. We note that Bell Atlantic argues that the 95 percent confidence level is appropriate. Bell Atlantic Dowell/Canny Decl. Attach. B, App. K; Bell Atlantic Duncan Reply at para. 36-38.

even under parity an average of 5 percent of the tests should fail (this is the probability of a Type I error).⁴ At higher confidence levels this probability would be lower, but then the probability of not detecting unexplained variations in performance if they do exist (the probability of a Type II error) would increase. The 95 percent confidence level appears to be a fair compromise. We do not comment here on AT&T's proposal to choose a confidence level of 85 percent, which it says will balance the probability of Type I and Type II errors.⁵ We find that AT&T has not put sufficient evidence on the record for us to determine that setting the confidence level at 85 percent⁶ will in fact balance the probability of Type I and Type II errors.⁷ “

In the specific situation of OSS Third Party Testing CGT believes a different framework is possible, and the need to control the risk of declaring parity when in truth we are materially far from it can be balanced with the need to control the risk of declaring disparity when in truth parity exists, without sacrificing the .05 level for either, at least in aggregate tests. If either risk ought to be emphasized more than the other in OSS Third Party Testing, then a very strong case can be made that it ought to be the risk that involves deciding to make a change to the status quo ante, the default assumption we have before conducting the test, that parity / compliance is not present, and declare that instead we have found such a preponderance of the evidence that it forces us to change our mind and declare that service to the CLECs is actually in parity with retail / compliant with benchmark. Such decision is the one which requires the greater burden of proof, and which is a prime goal which OSS Third Party Testing has been devised to demonstrate and establish.

- The only mention of the word “design” (as in statistically designed experiment) in the entire report is where some 33 calls are made from each of four locations to a help desk to test help center responsiveness. Generally, no evidence of reasons why different sample sizes are used to test different metrics is provided, and the sample sizes used are in most cases as small as 1 or 2. Being that that is not powerful enough to find evidence of dis-parity, the tables presented indicate “In Parity”.
- In about half the cases where close to sufficient sample sizes are taken, the result is “Out of Parity”, with a very high degree of significance $<.01$. It is surprising that this was not noted in the Executive Summary and that the NY PSC’s Chairman’s comments seem to not take note of it.
- In a few of these cases, substantially larger sample sizes were taken than needed to establish whether the metric is in or out of parity.
- Table IV-8.18 on P P8 IV-184 gives evidence that there were at least 1795 orders which went through the ordering system. Tables IV-8.19 thru IV-8.26 give evidence that at least 377 Resale and 2320 UNE orders were provisioned to at least the point of determining whether there was an appointment missed (Metrics 58-62). Appendix E seems to indicate that there were a total of 4769 LSRs and 3400 PreOrdering queries sent in the course of normal testing – (excluding volume testing).
- Based on what I see in Table IV-8.19-26, reasonably powerful parity testing could only be performed at a very limited level of disaggregation. Eg. Non-dispatch orders could be evaluated but not dispatch orders. *CGT’s approach in CA and as proposed in AZ would attempt to produce sufficient orders to evaluate parity / compliance with perhaps looser but still reasonable power and significance in each disaggregated cell, and then aggregate the results back up, weighting each disaggregated test result by its cell’s expected frequency in the target future order mix. The design would require the aggregated*

⁴ Type I and Type II errors are described above. See *supra* para.

⁵ AT&T argues that choosing a critical value to balance the probabilities of Type I and Type II errors is desirable, because it balances the interests of BOC and competitive LECs by setting equal the chances of falsely finding discrimination and of falsely missing discrimination. While acknowledging that the critical value to achieve this balancing (“balancing critical value”) will depend on the number of BOC and competitive LEC observations, they argue that using a fixed critical value based on an 85 percent confidence level is a reasonable approximation of the balancing critical value, given typical competitive LEC sample sizes. AT&T Pfau/Kalb Aff. at paras. 88-93 and n.97 and Attach. 2 at 27-30.

⁶ This would mean using a critical value for the z-test of 1.04.

⁷ AT&T’s proposal to balance the Type I and Type II error probabilities does appear to have the attractive feature that the interests of the incumbent LEC and the competitive LECs are given equal weight, so that the probabilities of falsely concluding the incumbent LEC may be discriminating and of missing existing discrimination are balanced (so $\alpha=\beta$). Such an approach could be used in future section 271 applications. We would be more likely to accept use of such an approach if the state commission and parties have agreed on its use, particularly since there are details that need to be worked out before it is used. For example, the relevant alternative hypothesis must be agreed upon. We note that the New York Commission has not accepted AT&T’s proposal. Bell Atlantic argues that AT&T’s proposal is not standard and is difficult to implement. Bell Atlantic Duncan Reply at paras. 36-38.

test results to then have the tighter .05 levels for both Type I and Type II Error (at a feasible materially different sub-parity alternative CLEC performance level).

8. The Texas Approach

Refer to the following links at <http://www.puc.state.tx.us/telecomm/projects/20000/download.cfm>

ossreport.pdf,
attach-c.pdf,
attach-k.pdf

Telecordia seems to have performed at least 648 orders, and at least 2733 pre-orders. For the 121 measures, they ended up having what they considered sufficient sample size to perform statistical evaluation on a total of 17, seemingly only at the state-wide level, of which 12 passed and 5 failed (Attachment K-01B).

Initially they proposed 569 test scenarios, expanded that to 612 to incorporate a shift to multi-line. Their number represents what they believed would be sufficient to generate sample sizes of 30 for several measures. It seems they had 201 / 220 unique test scenarios, some of which they iterated from 2 to 18 times to generate the 569 / 612 numbers. They elsewhere make the comment that an LSR typically generates 3 orders to SORD. This may explain why they seemed satisfied with 17-18 scenarios per product when they were trying to get samples of size 30, with perhaps not all orders qualifying for each measure. They further argued that the TAG’s 1009 number of scenarios was over-inflated because several of the TAG-team’s 425 proposed scenarios did not add functional value, so they suggested eliminating 224 of them. (Attachment C).

They later had to perform retesting of several measures in the billing area, primarily because the initial sample size was insufficient, which they ascribed to the situation in reality not reflecting the information they were provided with up front.

Telecordia statistically evaluated benchmark performance, requiring as with parity, only that deviations in the direction of poor performance be significant at the .05 level with the LCUG’s modified z-test.

Generally, other than attempting to use a sample of size 30 for performing statistical evaluations of parity / compliance for certain measures, they did nothing to ensure reasonable power that material deviations from parity / compliance would be found.

CGT Statistical Comment:

A sample of size 30 is generally regarded as sufficient to allow the use of a standard normal distribution instead of the t-distribution. It also is considered to be sufficient when testing the means of arbitrary distributions (provided they are not overly skewed – our interval measures are highly skewed, which we can remedy by performing our tests on the natural logarithms of the time intervals instead of upon the intervals themselves). But while 30 may be enough for the test to have validity as a statistical test, the question remains, of what ?

I will here indicate the “material difference” from parity that such a test is capable of detecting with 95% power (ie. $\beta=.05$):

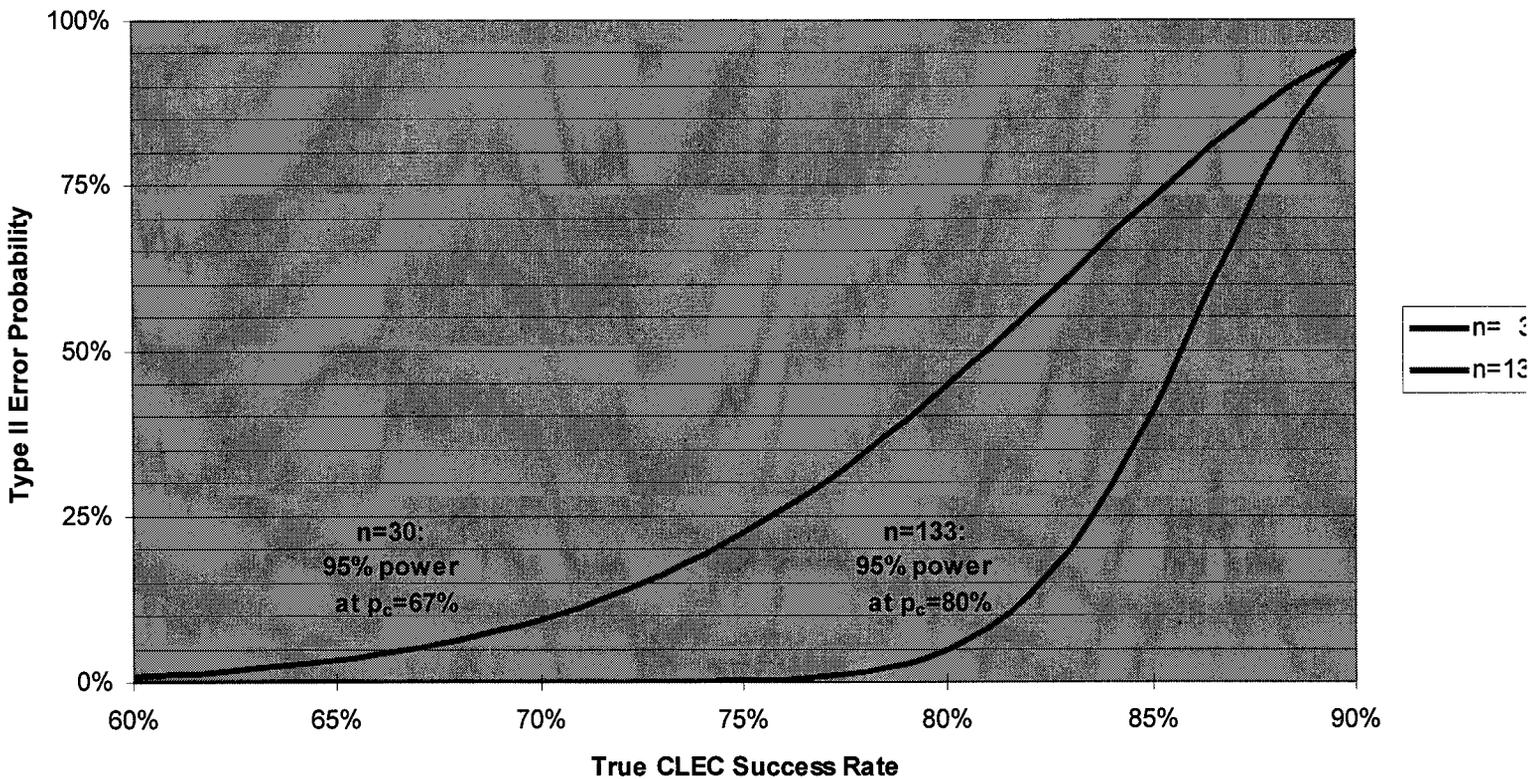
Parity/ Compliance Performance Level	Parity with Retail	
	Materially Different Alternative	Sample Size

50.00%	22.00%	29
66.67%	37.50%	30
75.00%	47.00%	30
80.00%	52.50%	29
85.00%	59.00%	29
90.00%	66.00%	29
95.00%	75.00%	29
96.00%	77.00%	29
97.00%	79.50%	30
98.00%	82.00%	30
99.00%	85.00%	29
99.25%	86.00%	29
99.50%	87.00%	29
99.90%	89.70%	30

The question may then be framed as follows: “If ILEC performance (or the benchmark) is at 90%, and the OSS is configured such that long-term CLEC performance is at 70%, then our test will have a substantial chance ($>.05$) of concluding there is parity / compliance. Is this acceptable ?” Following is a graph of the probability of accepting parity / compliance depending on the true underlying OSS configured CLEC order success rate:

The preceding table was not understandable to AT&T as to the purpose of the table or the point the table was attempting to make. AT&T requests a more complete explanation of the table.

Probability of Accepting Parity / Compliance with Benchmark/Retail value of 90% with sample sizes of 30 (TX) and 133 (AZ proposed)



9. The Pennsylvania Approach

Refer to Appendix C in the WinZip file downloadable from:

<http://puc.paonline.com/Telephone/OSSTesting/Master%20test%20plan/Master%20test%20plan%20page.htm>

This is the Statistical Approach section of KPMG's Master Test Plan for PA dated March 29, 1999.

Same as NY with following exceptions:

- **Benchmarks:** Statistical "Hypothesis testing" proposed – example provides a 2-sided test-criterion when 1-sided is appropriate – so direction of hypotheses proposed is neither that of CA or the other way around.
- **Parity:** Example provides 2-sided LCUG-modified z-score of 3 equivalent to a 99.7% confidence interval. This indicates that performance either significantly better or significantly worse than retail will be cause to reject the "null" hypothesis of parity. As with benchmarks the two-sided alternate hypothesis implicitly proposed here is not compatible with either the NY approach or the CA approach, or with any 271 approach anywhere. Everything I've read indicates that people recognize that the only deviations from parity that are of concern are those which indicate sub-parity, ie. worse service. Further, such a small probability of Type I Error as .003 in this environment is an attempt to put blinders on to make sure that in case there are any differences we don't see them unless they are very extreme.
- As in NY nothing regarding OSS Testing as a statistically designed experiment, no control of risk to the CLECs whatsoever via requiring tests of adequate power at a measure-specific agreed-upon materially different alternate dis-parate value. This results in no required sample sizes, and consequently even less ability to find any disparities that may exist.

The above PA Statistical Approach is marked Draft Copy, except for the final page which is entirely blank and marked Final Copy. All pages are marked CONFIDENTIAL: For PAPUC, Bell Atl., and KPMG Internal use only.

I don't believe this will be their final approach. I don't think this section was written by a statistician, a statistical professional, or anyone who had already been even remotely connected with the analysis KPMG performed in NY.

10. The approach contemplated in Florida

Currently Florida is investigating Third Party Testing. I could find no information on their web site other than that they plan to use the approaches of New York and Pennsylvania, perhaps because they may already have chosen KPMG as their Test Administrator.

11. Benefits of CGT Approach as compared to other States

- Efficient allocation of resources (test orders) to test important measures at both aggregate and disaggregated levels, with control of the probabilities of both risks of false positives and false negatives.
- Achieves balanced error probabilities and .05 level of significance – best of both AT&T and RBOC approaches.

See note 50 in Appendix B of FCC decision on NY 271 application:

“AT&T’s proposal to balance the Type I and Type II error probabilities does appear to have the attractive feature that the interests of the incumbent LEC and the competitive LECs are given equal weight, so that the probabilities of falsely concluding the incumbent LEC may be discriminating and of missing existing discrimination are balanced (so $\alpha=\beta$). **Such an approach could be used in future section 271 applications. We would be more likely to accept use of such an approach if the state commission and parties have agreed on its use, particularly since there are details that need to be worked out before it is used. For example, the relevant alternative hypothesis must be agreed upon.**”

- ❑ After the FCC’s thorough presentation and discussion in its NY decision of the general theory of hypothesis testing, especially including the concept of Type II Error at a materially different alternative, proper use of the concept should no longer be viewed as “not standard.” While there may be some difficulty of implementation in the general case where there is no control over sample size and a straightforward mechanized approach to determining incentive payments might be desired, however, Third Party OSS Testing is different. In Third Party OSS Testing, where sample size can be controlled and set so that the different risks involved are acceptable to the parties involved, implementation of

this approach is not overly difficult – it is just standard statistical methodology of experimental design for hypothesis testing.

- ❑ Testing for most measures and disaggregation levels should be capable of being performed with substantially less orders than the 4000-5000 in NY.

CERTIFICATE OF SERVICE

I hereby certify that the original and 10 copies of the Comments of AT&T and TCG Phoenix on Cap Gemini's Statistical Plan Docket No. T-00000A-97-0238, were sent via overnight delivery this 17th day of January, 2000, to:

Arizona Corporation Commission
Docket Control – Utilities Division
1200 West Washington Street
Phoenix, AZ 85007

and that a copy of the foregoing was sent via overnight delivery this 17th day of January, 2000 to the following:

Carl J. Kunasek, Chairman
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

Jerry Porter
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

Jim Irvin, Commissioner
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

Patrick Black
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

William A. Mundell, Commissioner
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

Hercules Alexander Dellas
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

Christopher Kempley
Arizona Corporation Commission
Legal Division
1200 West Washington Street
Phoenix, AZ 85007

Deborah Scott
Director - Utilities Division
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

David Motycka
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

Mark A. DiNunzio
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

Maureen Scott
Legal Division
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

Jerry Rudibaugh
Hearing Officer
Arizona Corporation Commission
1200 West Washington Street
Phoenix, AZ 85007

and that a copy of the foregoing was sent via United States Mail, postage prepaid, this 17th day of January, 2000 to the following:

Timothy Berg
Fennemore Craig, P.C.
3003 North Central Ave., #2600
Phoenix, AZ 85012

Joan S. Burke
Osborn Maledon
2929 N. Central Avenue, 21st Floor
Phoenix, AZ 85067-6379

Thomas M. Dethlefs, Esq.
U S WEST Communications, Inc.
1801 California Street, #5100
Denver, CO 80202

Thomas H. Campbell
40 N. Central Avenue
Phoenix, AZ 85004

Thomas F. Dixon
MCI WorldCom, Inc.
707 – 17th Street, #3900
Denver, CO 80202

Michael M. Grant, Esq.
Gallagher and Kennedy
2600 North Central Ave.
Phoenix, AZ 85004-3020

Scott Wakefield
Stephen Gibelli
Residential Utility Consumer Office
2828 North Central Ave., #1200
Phoenix, AZ 85004

Michael W. Patten
Brown & Bain, P.A.
P. O. Box 400
2901 North Central Ave.
Phoenix, AZ 85001-0400

Daniel Waggoner
Davis Wright Tremaine
2600 Century Square
1502 Fourth Avenue
Seattle, WA 98101-1688

Darren Weingard
Stephen H. Kukta
Sprint Communications Company L.P.
1850 Gateway Drive, 7th Fl.
San Mateo, CA 94404-2467

Doug Hsiao
Rhythms NetConnections
7337 So. Revere Parkway, #100
Englewood, CO 80112

Carrington Phillip
Fox Communications, Inc.
1400 Lake Hearn Drive, N.E.
Atlanta, GA 30319

Karen Johnson
Electric Lightwave, Inc.
4400 NE 77th Ave
Vancouver, WA 98662

Bill Haas
Richard Lipman
McLeod USA Telecommunications Services, Inc.
6400 C Street SW
Cedar Rapids, IA 54206-3177

Charles Kallenbach
American Communications Services, Inc.
131 National Business Parkway
Annapolis Junction, MD 20701

Richard M. Rindler
Morton J. Posner
Swidler & Berlin Shereff Friedman, LLP
3000 K Street, N.W. – Suite 300
Washington, D.C. 20007-5116

Mark Dioguardi, Esq.
Tiffany and Bosco, P.A.
500 Dial Tower
1850 North Central Ave.
Phoenix, AZ 85004

Richard Smith
Director of Regulatory Affairs
Cox Communications
2200 Powell Street, Suite 795
Emeryville, CA 94608

Joyce Hundley
United States Dept. of Justice
Antitrust Division
1401 H Street NW, Suite 8000
Washington, DC 20530

Jim Scheltema
Blumenfeld & Cohen
1615 MA Ave., Suite 300
Washington, DC 20036

Alaine Miller
NEXTLINK Communications, Inc.
500 108th Avenue NE, Suite 2200
Bellevue, WA 98004

Thomas L. Mumaw, Esq.
Snell & Wilmer, LLP
One Arizona Center
Phoenix, AZ 85004-0001

Raymond S. Heyman, Esq.
Randall H. Warner, Esq.
Roshka Heyman & DeWulf
Two Arizona Center
400 N. Fifth Street, Suite 1000
Phoenix, AZ 85004

Jeffrey W. Crockett
Snell & Wilmer, LLP
One Arizona Center
Phoenix, AZ 85004-0001

