ALJ / JAR / k47

Mailed 1/22/2001

Decision 01-01-037 January 18, 2001

BEFORE THE PUBLIC UTILITIES COMMISSION OF THE STATE OF CALIFORNIA

Order Instituting Rulemaking on the Commission's Own Motion into Monitoring Performance of Operations Support Systems.	Rulemaking 97-10-016 (Filed October 9, 1997)
Order Instituting Investigation on the Commission's Own Motion into Monitoring Performance of Operations Support Systems.	Investigation 97-10-017 (Filed October 9, 1997)

INTERIM OPINION ON PERFORMANCE INCENTIVES

TABLE OF CONTENTS

TABLE OF CONTENTS

TITLE **PAGE**

TABLE OF CONTENTS

TITLE **PAGE**

Summary

The Telecommunications Act of 1996 (TA96 or the Act) was a major step in the process of opening previously monopolistic local telephone service markets to competition. To foster competition, the act requires the incumbent local exchange carriers (ILECs) to provide competing carriers access to any necessary ILEC infrastructure, including the incumbents' operations support systems (OSS). OSS includes pre-ordering, ordering, provisioning, maintenance, billing, and other functions necessary to providing various telephony services. For competition to occur, the competitive local exchange carriers (CLECs) must be able to access these services in the same manner as the ILEC.

For example, for pre-ordering, a CLEC must be able to access customer information relevant to the service being ordered, so that the CLEC can tell its customers what options they have. For ordering, a CLEC needs to be sure that the ordering process for its customers takes no more time than for ILEC customers. Similarly, for provisioning, a CLEC needs to be sure that the time the ILEC takes to actually install or provide a new telephone service for CLEC customers is no longer than for ILEC customers. Delays or inaccuracies in these and the other OSS functions could discourage potential customers from doing business with the competitors.

Under its authority to implement the Act, the Federal Communications Commission (FCC) has strongly encouraged that regulatory remedies be established to ensure ILEC OSS performance does not present barriers to competition. While not an outright prerequisite for FCC approval of Regional Bell Operating Companies' (RBOC) applications to provide in-region interLATA service under § 271, the FCC has indicated that such applications must be in the public interest. In its evaluation of the public interest, the FCC states that, "the

fact that a BOC will be subject to performance monitoring and enforcement mechanisms would constitute probative evidence that the BOC will continue to meet its section 271 obligations and that its entry would be consistent with the public interest."[1] As a consequence, we will establish a performance remedies plan to identify and prevent or remove any barriers. The three critical steps for any performance remedies plan are performance measurement, performance assessment, and the corrective actions necessary if performance is deemed harmful to competition.

The California Public Utilities Commission (Commission or CPUC) has established performance measures in a parallel proceeding in this docket. Our decision today establishes an interim performance assessment plan. We have created a set of procedures for assessing the performance measurement results to identify competitive barriers. In effect, we have set forth a self-executing decision model that applies barrier-identifying criteria to the performance measurement results. A self-executing plan is one that requires no further review and no new proceedings. Explicit, objective, data-based standards are established that automatically calculate and determine the existence of "competitive barrier" performance. Statistical tests identify barriers when ILEC performance to its own customers can be compared to ILEC performance to CLEC customers. Explicit performance levels, called benchmarks, identify barriers when there is no comparable ILEC performance.

This decision model now enables us to proceed to the final step of the remedies plan, establishing the incentives that will be tied to any deficient

[1] *Bell Atlantic New York Order* ("FCC BANY Order"), 15 FCC Rcd at 3971, ¶ 429.

performance identified by the model. The overall goal of the plan will be to ensure compliance with the FCC's directive that OSS performance shall provide competitors a true opportunity to compete.

Background

On October 9, 1997, the Commission instituted this formal rulemaking proceeding and investigation to achieve several goals regarding Pacific Bell's (Pacific) and Verizon California, Inc.'s (Verizon CA)[2] OSS infrastructure. One objective of this docket (the OSS OII/OIR) is to assess the best and fastest method of ensuring compliance if the respective OSS of the ILECs do not show improvement in implementation or meet determined standards of performance. Another related objective is to provide appropriate compliance incentives under Section 271 of TA96, which applies solely to Pacific[3], for the prompt achievement of OSS improvements.

To further these specific objectives, the ILECs and a number of interested CLECs participated in a series of meetings jointly conducted through the OSS OII/OIR proceeding and the 271 collaborative process[4]. In October 1998, a group

---

[2] Verizon CA was previously named GTE California Incorporated. Hereafter, Pacific and Verizon CA will be referred to collectively, as the ILECs.

[3] As a Bell Operating Company (BOC), Section 271 specifically applies to Pacific.

[4] From July through mid-August 1998, Pacific, AT&T Communications of California Inc. (AT&T), MCI WorldCom (MCI W), Sprint Communications, Electric Lightwave, Inc., ICG Telecom Group, Inc., Covad Communications (Covad), MediaOne Telecommunications of California, Inc., Cox California Telecom, LLC, Northpoint Communications, California Cable Television Association, and staff entered into a collaborative process and jointly worked on developing solutions to the flaws in Pacific's 1998 draft 271 application. Verizon CA observed one collaborative meeting on penalties, but otherwise did not participate. (Verizon CA Response to Motion to Accept

Footnote continued on next page

of the interested parties filed joint comments setting forth their various positions on the issues discussed during the meetings. Following a pre-workshop conference in January 1999, the assigned Administrative Law Judge (ALJ) and the Telecommunications Division staff (staff) convened a 7-day technical workshop[5] on the respective performance incentive plans of Pacific and the participating CLECs. Pacific and the CLECs filed concurrent opening briefs on March 22, 1999, and concurrent reply briefs on April 5, 1999.

Pursuant to ALJ Ruling, Verizon CA filed its proposal on incentives for compliance with performance measures on May 3, 1999. The CLECs responded to the proposal on May 11, 1999. On July 12-14, 1999, the ALJ and staff convened a technical workshop on Verizon CA's performance incentive plan in relation to the CLECs' plan[6]. The parties filed concurrent opening briefs on July 28, 1999, and concurrent reply briefs on August 4, 1999. On August 12, 1999, Verizon CA petitioned to have submission set aside and supplemental comments accepted. The CLECs responded to the petition on August 27, 1999.

On November 22, 1999, the assigned Commissioner noted in a ruling (the ACR) that staff and its technical consultants had advised him that the performance incentive plans that the parties had submitted were significantly flawed. The ACR set forth the framework of a performance remedies plan that it encouraged Pacific, Verizon CA and the CLECs to analyze and comment upon

---

Joint Comments regarding Report on Performance Incentives, footnote 2 at 2 (October 20, 1998)).

[5] February 5, 8-11, and 23-24, 1999.

[6] The CLECs submitted their plan in both the Pacific and Verizon CA portions of the proceeding.

with the overall goal of developing a common and acceptable approach to implementing the performance plan. The parties filed opening comments on the ACR on January 7, 2000. Pacific and the Office of Ratepayer Advocates[7] (ORA) included new performance incentive plan proposals with their initial comments. The parties filed reply comments on January 28, 2000.

On March 27, March 28 and March 30, 2000, the ALJ, assisted by staff, convened a facilitated workshop that focused exclusively on the performance assessment part of three performance remedies proposals: (1) the ACR-proposed plan; (2) the new Pacific plan, and (3) the ORA plan. The parties submitted opening and closing briefs on April 28 and May 5, 2000, respectively.

Performance Remedies Plan Fundamentals

The TA96[8] and the FCC's implementing rules require Pacific and Verizon CA to provide CLECs with nondiscriminatory access to unbundled network elements (UNEs), including OSS. The FCC commented generally that ILECs must provide the CLECs with access to the pre-ordering, ordering, provisioning, billing, repair, and maintenance OSS sub-functions pursuant to the Act such that the CLECs are able to perform such OSS sub-functions in "substantially the same time and manner"[9] as the ILECs can for themselves.

---

[7] ORA had monitored this phase of the OSS OII prior to its January 7th submission.

[8] Section 251(c)(3).

[9] *Implementation of the Local Competition Provisions of the Telecommunications Act of 1996, CC Docket No. 96-98*, First Report and Order, 11 FCC Rcd, at 15763-64 (1996) (Local Competition First Report and Order).

The Act does not expressly mandate the establishment of either performance measures or incentives, though the FCC has stated that the most probative evidence that the CLECs are provided with nondiscriminatory access to OSS will be evidence of actual commercial usage evaluated under a set of Commission-approved performance measures. Similarly, TA96 and the implementing rules have no stated requirement for an additional customer economic effect test. The FCC has stated that an ILEC may demonstrate statistically that the differences in measured performance are the result of random variation in the data, as opposed to underlying differences in behavior. The phrase "underlying differences in behavior" means differences in the statistical distributions of the ILEC and the CLEC that are generating the performance outcomes.[10] Thus, equality of distributions (when the ILECs' and the CLECs' distributions are the same) is a sufficient condition for parity according to the FCC.[11]

The cornerstone of any performance incentive structure is how parity is defined, since it is on those occasions when the ILECs are out of parity that

---

[10] Roughly speaking, distributions are different when average performance and range of performance (variability, distribution) are different. For example, CLEC customer phone service provisioning could take 7 days on the average, whereas ILEC customer service provisioning could take 6 days. In this example, average performance for the ILEC is better than for the CLEC by one day. For variability, even with equal ILEC and CLEC averages of 7 days, CLEC provisioning times could range between 1 and 13 days, whereas ILEC averages could range between 6 and 8 days. In this example, performance for the ILEC is less variable, and thus more predictable. ILEC customers could be told that their new service would be installed in 8 days or less, in contrast to CLEC customers who could only be told that their service would be installed in 13 days or less.

[11] Id.

incentive payments will be made. This Commission's definition of parity generally incorporates the above-stated objectives of the TA96 and the FCC. Thus, parity means that the ILEC is providing services in substantially the same period of time and manner (including quality) to the CLECs as it is providing to itself. Further, it will be helpful to rely on statistical testing and benchmarks to infer whether or not parity has been achieved. Consequently, we endeavor to ensure that the CLECs have OSS access that is at least equal to the ILECs' own access.

Initial Proposed Plans

This section provides an overview of the history of this proceeding, and focuses on the parties' various positions and plans. Brief explanations of statistical concepts are presented with the limited purpose of identifying parties' positions. A more detailed explanation of statistical concepts accompanies our deliberation in the section titled "Selection of the Statistical Model."[12]

Plan Principles

Pacific initially developed a statistical approach to determining compliance with TA96's nondiscriminatory access standard structured on three central principles. First, the remedy plan must not impose payments on Pacific when nondiscriminatory or parity treatment is provided.[13] However, Pacific conceded that, given the nature of the statistical models applied, it was difficult to drive the parity payment amount closer to zero without lowering the out-of-parity

[12] Readers wishing elementary or more detailed statistical explanations before reading this section may wish to first read the section titled: "Selection of the Statistical Model."

[13] "The expected cost for parity treatment should be zero."

payments substantially. (Pacific's 1999 Opening Brief on Performance Remedies at 2-3.)

Second, if Pacific does not provide parity treatment, then payment amounts to the CLEC should have some reasonable relationship to the level of performance provided.[14] Pacific argued that remedy amounts should not be enormous when the level of performance deviates from parity by only small amounts or in isolated incidents. Thus, the levels of remedies should start relatively low and increase commensurately with the level of nonperformance. Id. at 3.

Third, remedy payments should motivate Pacific to provide nondiscriminatory service, but should not motivate the CLECs to favor receiving large remedy payments.[15] Therefore, the remedy amounts must not be so high that a CLEC would be more desirous of receiving poor service and collecting large payments than receiving nondiscriminatory service. Id.

The CLECs also based their initial incentive proposal on three principals. They declared that the incentives must be in an amount sufficient to cause Pacific to meet its parity obligations. Second, the incentives must be self-executing without broad opportunity for circumvention or lengthy delay in the payment of the consequences. Finally, the CLECs asserted that the structure of the plan must be fairly simple to implement and monitor.

---

[14] "Payments should bear a reasonable relationship to level of performance."

[15] "CLECs should not be motivated to receive large remedy payments."

Parity and Statistical Model Elements

In its initial performance incentive proposal, Pacific defines parity to mean delivering services to CLEC customers from the same processes as delivered to ILEC customers. When organizationally it is not possible to have the same processes, Pacific then defines parity to mean that the ILEC must deliver services with the same properties to the CLEC as delivered to the ILEC. The definition for parity, and the test for parity, appears to be the same, i.e., 1.645 standard deviations from the mean.[16] (Pacific 1999 Opening Brief at 5-6 and 13-15.)

Verizon CA contends that parity only requires that CLEC ordering processes be performed in "substantially the same time and manner" as the ILEC's like processes. It claims that ILECs have unavoidable variations in their own processes, and as long as the ILEC and CLEC distributions are substantially the same, parity is present. Verizon CA also considers the appropriate test for parity to be average performance within 1.645 standard deviations of the mean. (Verizon CA 1999 Opening Brief at 5.)

The CLECs define parity as equal service for the ILEC and the CLEC. The CLECs want zero (0) standard deviations from the mean for the definition of

[16] A standard deviation is a standardized statistic measuring how dispersed scores are. A low standard deviation indicates scores are grouped closer to the mean than scores with a higher standard deviation. When applied to a normal or "bell-shaped" curve, the standard deviation provides helpful information about the dispersion of scores: 68.3 percent of all scores lie within one standard deviation of the mean (plus or minus one standard deviation, 95.4 percent lie within 2 standard deviations, 99.7 lie within 3 standard deviations, and so forth. In the present application, 1.645 standard deviations above the mean encompass 95 percent of the scores. So under conditions of random selection, a score greater than 1.645 standard deviation would be selected 5 percent or less of the time.

parity, but have offered that a test for determining parity could be one (1) standard deviation from the mean. (CLECs' 1999 Opening Brief at 4-15.)

In its May 3, 1999 preliminary statement, Verizon CA embraced each of the core principles Pacific and the CLECs set forth, and asserted that the concepts need not be mutually exclusive. Moreover, it added the following seven principles of its own to the "ideal" incentive plan. First, a design objective of the plan should be that no incentive payments should be made when parity exists. Consequences should be economically significant, not just statistically significant. Further, the incentive structure should provide that the incentive payment equals the resource cost of meeting the standard. Regular review periods are necessary. The incentive mechanism should not result in large administrative costs. There must be some "off-ramps" in a self-executing incentive system to deal with certain circumstances. Finally, with an eye to the future, the plan should be symmetrical across all parties. (Verizon CA Brief on OSS Performance Incentives at 2-5.)

Test for Determining Compliance with Parity

Pacific originally proposed using a standard Z-test[17] for purposes of determining compliance with parity. The CLECs objected to the standard Z-test,

---

[17] **Standard Z-test** : Z = Difference / Standard deviation of the difference
Where: Difference = Pacific Average – CLEC Average.

Standard deviation of the difference = Square root of ((Variance of Pacific x 1 / Pacific sample size) + (Variance of CLEC x 1 / CLEC sample size)).

Or, assuming the variances for Pacific and the CLEC are equal, the variances are pooled together: Standard deviation of the difference = Square root of ((Pooled variance of Pacific and CLEC samples) x (1 / Pacific sample size + 1 / CLEC sample size)).

which utilizes the individual variances of the Pacific and CLEC samples, arguing that Pacific could manipulate the variance of the CLEC sample. Pacific responded that the standard Z- test was adequate because any alleged manipulation of the CLEC sample variance would be readily apparent.

The CLECs speculated that Pacific could increase the variance of the CLEC sample, which would reduce the probability that Pacific would be found out-of-parity.[18] In response, they proposed the "Modified Z-test,"[19] which modifies the standard Z-test by using only Pacific's sample variance. In the "spirit of collaboration," Pacific offered to use the CLECs' proposed Modified Z-test on a trial basis, and then test it in order to evaluate whether the Modified Z-test yielded "fair and accurate results." Verizon CA agreed to use the Modified Z-test to assess parity subject to review and modification following a six-month interim implementation period.

---

[18] An increased CLEC variance theoretically could increase the size of the Z-test denominator without affecting the numerator, thus reducing the resulting Z-test statistic and reducing the chances of identifying out-of-parity situations.

[19] **Modified Z-test:** Z = Same as Z-test.

Where:

Difference = Same as Z-test.

Standard deviation of the difference =

Square root of (Variance of Pacific x (1 / Pacific sample size + 1 / CLEC sample size)).

Minimum Sample Size

Pacific initially desired a minimum sample size of thirty occurrences.[20] In the "spirit of cooperation," Pacific was willing to lower the sample size to twenty, with the caveat that the impact of smaller sample sizes be evaluated during a review period in the not too distant future. Pacific also accepted benchmark measures for a specific list of rare submeasures.[21] That is, parity measures with rarely occurring activity were essentially to be converted to benchmark measures.

The CLECs acknowledge that many of their number will have fewer than thirty observations (e.g., orders) in a month for some measures. They want to ensure that a requirement of a larger sample size does not passively provide an acceptable level of performance to the ILEC. Therefore, the CLECs preferred sample sizes as small as one, but suggested a minimum sample size of five for parity submeasures. The CLECs also accepted the benchmark measures for the specific list of rare submeasures.

Verizon CA supported the use of "table lookup"[22] for sample sizes exceeding 50 CLEC transactions. Noting that there is a lack of experience using

---

[20] A sample size of thirty is a standard textbook "rule-of-thumb" sample size cutoff for parametric statistical testing such that distributional assumptions can be anticipated to be met for most situations.

[21] A "measure" defines how performance will be measured for a specific OSS function, such as ordering, across several service types, such as residential telephone service, business telephone service, DSL service, etc. A "submeasure" applies the specified "measure" methods to individual service types, for example, either residential telephone service, or business telephone service, or DSL service, etc

[22] The statistical test produces a test value. The test value can then be "looked up" in a table to determine statistical significance. In most cases a normal approximation or a "t"

Footnote continued on next page

the Modified "t" statistic[23] for non-normal samples, Verizon CA advocated using permutation tests for sample sizes between 20 and 50. (Verizon CA 1999 Opening Brief at 33-34.) For sample sizes less than 20, Verizon CA originally proposed that the CLECs and it should explore, during the interim development period, use of: (1) permutation tests; (2) aggregation of results across sub-measures; (3) aggregation of results across CLECs; and (4) possible exclusion of a given measure from performance incentive assessment. During the interim period, Verizon CA stated that it would also rely, to the extent practicable, on "exact methods"[24] to determine achieved significant levels for small sample tests on proportions. (Id. at 34.)

Alpha Level/Critical Value

Pacific and Verizon CA proposed a Z statistic of greater than 1.645 standard deviations (critical value) to determine "out-of-parity." A 1.645 standard deviation corresponds to a five percent (one-tailed) Type I error, or "alpha." A Type I error is rejecting the null hypothesis (i.e., parity service)[25] when it should not be rejected. A Type II error is accepting the null hypothesis when it should not be accepted. "Alpha" is the probability of a Type I error and "beta" is the probability of a Type II error. Values of 1, 5, and 10 percent alpha levels are the most common "textbook" values.

---

distribution table is used to determine the Z or t statistic that must be exceeded for a performance failure finding.

[23] The "Modified t-test" is a variant of the Modified Z-test used for sampling distributions of small sample mean, as discussed later in this Decision.

[24] The term "exact methods" is defined as performing all possible permutations.

[25] A "null hypothesis" proposes that there are no differences between the true means.

The null hypothesis in this application poses that ILEC and CLEC performance are in parity. A Type I error is identifying the ILEC as not providing parity service (i.e., the ILEC is providing worse service to CLECs than to itself) when in fact the ILEC is providing parity service. A Type II error is identifying the ILEC as providing parity service when in fact it is not providing parity service. Pacific wanted to be limited to a five-percent probability of being identified as not providing parity service when in fact it is providing parity service.

The CLECs recommended an equal error methodology be employed for setting the errors. This essentially calculates and equates the Type I and Type II errors for each submeasure each month. The CLECs ultimately suggested that a Z statistic of greater than 1.04 standard deviations (critical value) should identify "out-of-parity" conditions. A 1.04 standard deviation corresponds to a fifteen percent (one-tailed) Type I alpha level. The CLECs were concerned with Type II errors, not just Type I errors. By making the critical alpha level larger, the CLECs worried less about the beta error.[26] Thus, the CLECs wanted at least a fifteen-percent probability limit for identifying Pacific and Verizon CA as not providing parity service when in fact they are providing parity service, because they believed that this would correspond more closely to an equal probability of identifying non-parity service as parity service.

Assigned Commissioner's Ruling and Proposed Plan

By ruling issued November 22, 1999, the assigned Commissioner assessed the submitted proposed plans and set forth his concerns about them (the ACR).

[26] As the critical alpha level is increased (e.g., from 0.05 to 0.15), beta decreases.

The ACR noted that the existent ILEC models and the CLECs' model appeared distinct and incompatible. In addition, the parties revealed considerable misunderstanding and confusion about the two sets of respective model assumptions and calculations. It was difficult to sort out the relative impacts of each of the respective components of the two differing model approaches. Moreover, the end result outcomes of the two models were highly uncertain because both the modeling approaches were trying simultaneously to design and implement the total model (both the performance assessment model elements and the incentive plan elements) without the benefit of an implementation and data calibration structure.

While the plans' proponents had articulated numerous core concepts, no distilled set of principles supported both plans. There also appeared to be little rationale for the incentive levels implicit in either plan. It is unlikely that either plan could be implemented as designed. Moreover, both models might impose costs when evidence suggests parity service, and both models might not impose costs when evidence suggests non-parity service. During the February 1999 technical workshop, each proposed plan produced dramatically different payments due to different input assumptions. Both plans were also very sensitive to minor changes in assumptions. These problems were not due to an attempt to keep the plans simple; both the ILECs' and CLECs' plans were very complex. Accordingly, we affirm the ACR's evaluation of the initial ILECs' and CLECs' plans.

The ACR expressed the need to have one common interim model framework of analyses for review and discussion, and for use by all concerned parties in order to implement the performance remedies plan. One interim performance remedies plan model and set of explicit assumptions, would allow

common quantitative analyses to be performed and estimates to be developed. All key model assumptions would be explicit, and the policy ramifications of these assumptions would be clear.

The ACR proposed that a common and feasible approach to implement the necessary performance remedies plan[27] be developed with the assistance of the ILECs and the CLECs. It noted that to achieve the single common model framework, there needed to be an unwinding of the performance assessment model elements and the incentive plan elements that the parties merged together from the outset. To that end, the ACR proposed an initial conceptual performance measurement statistical model, and asked the parties to respond to specific questions about the model. Further, it proposed that the Commission implement a fully functioning, self-executing performance remedies plan during a six-month pilot test period.

We concur with the ACR assessment that a single model approach would allow the Commission to make informed policy decisions about the performance remedies plan. A single model approach focuses on the goal of parity service by the ILECs, economic incentives paid by the ILECs, and/or a change in ILECs' operations support to the CLECs. The end goal is certainly not just to have complex statistical measurement theory applications. There may be a variety of statistical measurement approaches that can all achieve the same basic economic and operations incentives by using different incentive plan structures and amounts, in combination with different measurement approaches.

---

[27] To avoid confusion with the work going on in the Performance Measurement segment of this proceeding, what is essentially the "performance measurement, assessment and incentive" plan will be referred to as the "performance remedies" plan.

A single common interim model and a single set of explicit assumptions should allow calibration of end result economic outcomes both before and after a six-month pilot test period using actual empirical data. The interim pilot test period can assist the Commission in determining the appropriate levels of long-term economic incentives. Long-term incentive impacts can be calibrated in relation to one model, one common set of assumptions, and actual test period empirical data. Penalty amounts and structures can still be set and paid during the pilot test period, and they can be applicable only during this interim period, unless otherwise determined.

ACR Plan Statistical Model Elements

Noting the ILECs' and CLECs' distinct views on standard and Modified Z-tests, the ACR questioned whether there would be a way to determine if the Modified Z-test yields "fair and accurate results." Of interest are differences in the results if the standard Z-test was used rather than the Modified Z-test. Such differences would be due to disparities between the variances of Pacific and the CLECs. Regarding the CLEC position that the variance of the CLEC sample could be potentially manipulated, the ACR stated that concern about the possibility of manipulation should not direct the test procedure.

The ACR suggested that the optimal course might be for the Commission to proceed with the standard Z-test on a trial basis to be evaluated after a six-month test period. The proposed Modified Z-test[28] applies an experimental

[28] It also holds the possibility of manipulation.

argument[29] to an observational situation. There are no other academic precedents for our application of this particular modified calculation. The ACR stated that it was doubtful at this point whether any further complicating modifications to the statistical methodology for determining compliance with parity would be worth the benefits without first trying the standard Z-test.

The standard Z-test is the most common method to compare two population means, under the following key assumptions:

1. Underlying distributions are not too skewed (i.e., they are not too different from a normal bell shaped curve).
2. Sample sizes are reasonably large.
3. Observations are independent measurements from the same processes (e.g., phone service installation operations).

If the variances are known to be equal, then a pooled, or common, variance estimate is used. If the variances are known to be unequal, then both separate variances are used. If it is unknown, *a priori*, whether the population variances are equal or not, then an initial test compares the variances. Based on this first test, either the separate or pooled variance estimate is used.

The genesis of the Modified Z-test assumes the contention that Pacific could manipulate the variance of the CLEC sample. While such manipulation might be possible, it seems equally likely that Pacific could simultaneously manipulate the mean of the CLEC sample, and the variance and mean of the corresponding Pacific sample. The ACR proposed to first test for variance equality between Pacific and CLEC results. If the variances prove to be unequal,

[29] Brownie, Cavell, Boos, D., and Hughes-Oliver, J. *Modifying the t and ANOVA F Test When Treatment Is Expected to Increase Variability Relative 2 Controls*, 46 Biometrics at 259-266 (1990).

the ACR suggested that it might be necessary to use the standard Z-test with both variances. In either case, parity will be assumed to exist when the differences in the measured results for both the ILECs and the CLECs in a single month, for the same measurements, are less than the critical value[30] of the Z-test.

Early on, the CLECs implied that the difference between the standard Z-test and the Modified Z-test could measure Pacific's ability to manipulate the data. Since both Pacific and the CLECs have agreed to use the Modified Z-test during a pilot test period, the ACR raised the possibility that both the standard and Modified Z-tests might be calculated and evaluated over the six-month pilot test period. However, the ACR further proposed that if both tests were run, actual calculations during the trial test period would be based on the standard Z-test. The results of the evaluation might suggest that the decision as to which form of Z-test to use might be moot, since all choices might identify the same situations as being out-of-parity.

The ACR also suggested that during the six-month pilot test period, sample distributions could be reviewed to explore whether the distributions meet the above-stated underlying assumptions of the Z-test. At the end of this six-month pilot test period, there could be a reconsideration of whether any variety of Z-test should be used, or whether nonparametric tests[31] might be more appropriate. All of the Z-tests described by Pacific and the CLECs are parametric

[30] The critical value of the Z-statistic corresponds to a critical alpha value. The rejection region encompasses the critical Z-statistic and larger Z-statistic values, which correspond to critical alpha and smaller alpha values.

[31] Distribution-free tests based on medians or ranks; that is, tests not dependent on assumptions about distributions, such as normality.

tests. They assume observations are independent and are generated from the same process with a relatively well-behaved distribution.[32] However, the ACR questioned the independence of the observations and the shapes of the distributions, especially the CLEC distributions. The ACR suggested that if these characterizations were accurate, over the long-term it might be better to use nonparametric tests.

Finally, the ACR noted that there appeared to be some confusion regarding the concept of samples versus entire populations. If, as the ACR surmised, it would be appropriate to assume we had the entire population of measurements during a time period, as with production output, then it might make sense to ultimately utilize concepts of statistical process control to monitor and modify the procedures when they appear to have gone, or likely will be going, out of control. For example, a production monitoring and control methodology[33] could utilize the mean and variance of the ILEC (essentially as a benchmark against which CLEC measurements are compared). This could be performed using a Z-test-based chart set only on the mean of CLEC measurements against the historic mean and variance[34] or other statistics of the ILEC. Or similarly, a permutation test could be used.

---

[32] "Well-behaved" refers to distributions where a resulting distribution of sample means is not deviant enough from a normal distribution to cause inaccuracies – discussed later in this decision.

[33] For example, a Shewart Quality Control chart. R. Mason, R. Gunst, and J. Hess, Statistical Design And Analysis Of Experiments With Applications To Engineering And Science at 65 (1946).

[34] Or cumulative values.

The ACR suggested that the real problem here might be that many performance measures ostensibly constructed from "samples" really are constructed from the complete set of actual observations. The ACR reasoned that frequently, a one-month observation is really a "sample" of the entire length of the production process, but is not a random sample, unless selected from among all of the months of production using some random procedure. In many instances, the proper statistical application may be statistical quality control viewing data as a time series. At the end of the six-month pilot test period, the confusion surrounding the sample versus population issue should be resolved. The ACR indicated that it would be very important to analyze the key underlying assumptions during the six-month pilot test period in order to establish the reasonableness of these assumptions and to understand the potential impact of any divergences from them.

Initially, the ACR plan did not contemplate a Z-test, or any other statistical test, for benchmark measures. It proposed to regard any measure that exceeds the benchmark value as a performance failure. Consequently, it envisioned that any performance worse than a benchmark would not be tolerated, and if exceeded, at least some penalty would be assessed. The ACR recommended monitoring the number of observations (e.g., orders) and improving benchmark measures over time taking into account the actual number of observations realistically expected to occur. For the immediate future, the ACR suggested treating benchmarks as absolutes, but moderating the impact of exceeding the benchmarks by means of smaller penalties for each occurrence. It also suggested that penalties should be greater for larger deviations from the benchmark.

Treating benchmarks as absolutes assumes that the parties established the benchmark values with some knowledge of the anticipated ability to meet them

and / or the relative frequency of time they reasonably could be met. The frequency and value of the ILECs' inability to provide service meeting the benchmarks could be monitored and re-evaluated during the initial six-month pilot test period. Any dramatic differences between assessing performance with parity versus benchmark measures could eventually be resolved either by readjusting the alpha values, or benchmarks, or the incentives.

Minimum Sample Size

The ACR concurred with the concept of converting parity submeasures with rare activity to benchmarks. It suggested that additional rare activity submeasures should be converted to benchmarks. The ACR stated a preference for benchmark measures over parity measures for performance remedies, because benchmark measures do not require any complicating summary statistics. Early estimates indicated approximately forty percent of all measures were benchmarks, and that sixty percent were parity measures. Approximately fifteen percent of all measures had both parity and benchmark submeasures. The ACR expressed the hope that over time, the parties would agree to convert even more parity measures to benchmark measures.

The ACR surmised that sample size proposals were justified more by pragmatic concerns than by statistical principles. Proposed sample size specifications reflect negotiated values more than statistical criteria. For example, selecting a minimum sample size of five suggested one of two things: (1) either the cost to collect each observation is extremely expensive, or (2) there

is an insufficient population from which to sample.[35] The issue of minimum sample size is relevant only for the first situation.

If all five observations occur during a particular time period, this is the entire population of measurements instead of a sample. The only sampling analog is to assume that the five observations are a sample of the potential observations that could have occurred during that same time period. Usually measurements are made with sufficient frequency to allow for corrective action if the process is beginning to "go out of control," or because management prefers to review data on a set periodic basis (i.e., hourly, daily, weekly, monthly, etc.). Such "periodicity" of measurement is usually established independent of sample size concerns. The ACR suggested that if too few observations occur in an established time interval, either the time interval can be lengthened, or the test can be performed using an aggregated measure incorporating more than one measurement. Or, the consistency of measurements could be tracked over time (e.g., number of "misses" for percent success measures) using statistical quality control charts.

The current assumption is that the time period for measurement is monthly. The ACR proposed lengthening the time period when the number of observations (e.g., sample size) is very small. However, the ACR recommended that this time period should not be so long as to enable the ILEC to manipulate results, and/or escape detection for providing non-parity service to the CLECs.

The ACR proposed to proceed with a minimum sample size of thirty, which could be aggregated in up to three-month time periods. Thus, a minimum

[35] For instance, such as might occur in the case where there are only five observations within a specified time period.

sample size of at least thirty would be obtained through an accumulation of up to three months, if necessary. If any sample size, aggregated or not, were to reach thirty in one, two, or three months, then the test would be performed when the number of observations first reached thirty. If, at the end of three months, the sample size had still not reached a minimum of thirty, the test would be performed using whatever sample size was achieved, regardless of the sample number. Ultimately, the measurement probably would be included in the rare occurrence benchmark list if fewer than thirty measurements happened during three months.

The ACR also advised that the appropriate length of time period for aggregation would be evaluated during the six-month pilot test period to better understand the frequency of measurements. Such an evaluation would aid in answering the question: "How many of each type of measurements can reasonably be expected to be made during any one month?" Any additional rare submeasures that could become benchmarks would also be evaluated during this pilot test period.[36] The ACR proposed to analyze any relatively large CLEC or ILEC values that skew the general tendency of the other values. (ACR at 24-25.)

Alpha Level/Critical Value

The ACR observed that it appears not to matter which critical value is actually employed, since the amount of the penalty can be adjusted to provide equivalent expected outcomes for the different possible critical values. The ACR proposed to track the actual alpha level outcomes, and ultimately calibrate the

[36] As stated, if there is no sample of observations, but instead, the population of CLEC values and/or ILEC values, the issues of errors and distributions are not really relevant.

size of payments as a function of the actual values. The greater the Z-statistic value (corresponds to a smaller Type I alpha error), the larger the penalty. The ACR proposed that in this proceeding, there should be no single critical cutoff value but a range of values. However, the ACR proposed that if one discrete cutoff value must be selected, it be a ten-percent Type I alpha level for parity tests. Preliminarily, ten-percent was a split between the suggested five and fifteen-percent values, and it is a commonly used critical value. This alpha level corresponds to 1.282 standard deviations.

The ACR described the CLECs' critical value proposal to be more of an "equal error" proposal than the "equal risk" proposal as the CLECs introduced it. Equal error refers to decisions with the same Type I and Type II error probabilities. Equal risk refers to decisions where the consequences of the decisions are equal, such as equal dollar losses. Their ultimate proposal does not equate the two expected dollar losses. In addition, the significance level that equates Type I and II errors varies by sample sizes and underlying distributions. The ACR also noted that the CLECs indicated concern with the Type II error, not just the Type I error. While fifteen-percent alpha levels are not commonly used for hypothesis testing, they are sometimes used for monitoring.

In their initial brief, the CLECs suggested that a performance payment be made for any occurrence beyond the acceptable level in a benchmark. (CLECs' 1999 Opening Brief at 3.) The ACR offered a similar recommendation, and pointed out that the CLECs also proposed that a specific table[37] be used to detail the small sample size benchmark standard comparable to the table agreed upon

[37] CLECs' 1999 Opening Brief at 33.

for large sample sizes (i.e., thirty or more observations). The ACR noted that the proposed table was negotiated, and did not systematically adopt the “closest” percentage possible compared to what would be expected from a large sample. It was unclear whether Pacific accepted this particular CLEC proposal.

The ACR remarked that while the concept of payments for all missed benchmark measures is easy to implement, it assumes accurate measurements. The ACR proposed discarding the benchmark table entirely at this juncture, and going with some level of graduated penalty for any measurement over the benchmark. For example, very small benchmark penalties could be assessed for very small frequencies of occurrences, and much larger penalties could be set for larger frequencies of occurrences.

For small sample sizes, the CLECs suggested permutation-testing procedures to compute the exact alpha and beta calculations.[38] (CLECs' 1999 Opening Brief at 30.) Pacific accepted this suggestion, specifying that the sample size should not be less than ten, if and when the Commission orders permutation testing. The company commented that permutation testing "is not an intuitive process for most people." Pacific recommended studying the validity and feasibility of utilizing permutation testing and that the approach be revisited

[38] Permutation testing involves direct estimation of probabilities from the actual data distribution, rather than inferences drawn from normal distribution “look-up tables.”

after a trial test period. (Pacific 1999 Opening Brief at 2.) The ACR suggested that permutation-testing procedures might be a reasonable application.

Desiring larger numbers of observations so that there would be little need for permutation testing procedures as a result of sample size, the ACR outlined its concern. Proposed statistical procedures use one-tailed tests to indicate when penalties should be assessed against the ILEC for poorer service to the CLECs, but do not yield any incentives to the ILEC for providing exceptional service. Still, the ACR acknowledged that permutation-testing procedures could have some role in assessing more exact measures of error. The ACR recommended that during the pilot test period, there be an evaluation of this application of permutation testing.

The ACR asked the parties to respond to four specific questions[39] and to submit comments on the overall statistical model approach presented in the ruling. The parties[40] filed opening and reply comments on January 7, and January 27, 2000, respectively.

Responses to the ACR Questions and Comments on Its Overall Statistical Model Approach

Use of Standard Z-test or Modified Z-test

In response to the ACR's initial question of why the standard Z-test should not be used in the model, Pacific advocated retaining the Modified Z-test for three reasons. First, the standard Z-test yields inaccurate Type I error rates under the conditions apparent in performance remedies plans, i.e., in the absence

[39] The ACR questions are reproduced in the attached Appendix A.

[40] Pacific, the CLECs, Verizon CA and ORA.

of normal distribution and with relatively few large samples. Second, the Modified Z-test is easier to compute. Third, the Modified Z-test is sensitive to differences in the CLECs' variances. (Pacific Opening Comments on the ACR at 2 and 5.)

The CLECs urged using the Modified Z-test, yet agreed that the standard Z-test could be used. (CLECs' Opening Comments on the ACR at 3.) Verizon CA endorsed use of the standard Z-test, with modifications. It maintained that parties should be able to calculate and evaluate both the standard and Modified Z-tests during the evaluation or pilot test period. (Verizon CA Opening Comments on the ACR at 2.) ORA argued that since the underlying series or performance measures are not normally distributed, the true probabilities are unknown and the Z-test is of little value. It opposed using formal statistical tests for performance incentives. (ORA Opening Comments on the ACR at 5.)

Use of Benchmarks without Statistical Tests

To the ACR proposal to use benchmarks without statistical tests, Pacific asserted that benchmarks without statistical tests require an ILEC to meet an unreasonably higher standard of performance for small sample sizes than for large sample sizes. Pacific stated that statistical tests for benchmark measures make it possible to achieve a uniform Type I error rate for all measures under conditions of parity and compliance. Pacific segued from this question into an introduction of its white paper concept of converting all benchmarks to " standards" against which all the CLECs' results could be statistically tested. (Pacific ACR Opening Comments at 6 and 9.)

The CLECs remarked that the ACR's desire "to see more parity measures turned into benchmarks [ACR at 27] " was "troubling and difficult to understand." (CLECs' ACR Opening Comments at 8.) The CLECs continued to

support a limited benchmark approach with no associated statistical component (except for the use of a table for small sample sizes). However, they maintained that unlike the parity standard, which requires the use of statistics to compare distributions, a benchmark standard requires no comparison other than the benchmark itself. The CLECs urged the enforcement of the benchmark standards adopted in D.99-08-020. (CLECs' ACR Reply Comments at 5.)

Verizon CA supported using benchmark measures without any statistical tests during the pilot period for all previously designated benchmark measures. Verizon CA agreed that the ACR's simple approach could be used during the pilot. Notwithstanding, Verizon CA proposed examining other alternatives such as tables for small sample sizes and the use of statistical tests with benchmarks. (Verizon CA ACR Opening Comments at 12.) ORA argued that benchmarks should be based on historical and not future data, and should be limited to those measures in which there is historical data available on at least 20 observations. (ORA ACR Opening Comments at 6.) ORA asserted that benchmarks should be defined as the historical mean of the series plus one standard deviation.

Use of Special Tables for Benchmark Measures

Pacific urged, and the CLECS agreed to, the use of special tables for percentage-based benchmarks with small samples. The CLECs favored the use of a table for benchmarks with small sizes. While allowing that the ACR's simple benchmark approach could be used, Verizon CA advocated alternatively examining the use of tables for small sample sizes. Verizon CA endorsed Pacific's adjusted table of percentages for benchmarks. As noted, ORA opposed the use of any formal statistical tests for performance measures.

Use of Minimum Sample Size of Thirty

In response to question 3, Pacific agreed that samples of thirty are adequate for average-based parity submeasures. It did not agree that a sample size of thirty is appropriate for benchmark measures that are interpreted as absolute standards and for percentage-based measures for which the benchmark is near zero (0) or 100 percent. (Pacific ACR Opening Comments at 12.) Pacific initially desired a minimum sample size of 30 occurrences, which is the standard "rule of thumb" for parametric statistical testing. As a compromise, Pacific was willing to lower the sample size to 20, with the caveat that the impact of the small sample sizes be evaluated at the end of the six-month trial test period. It also accepted benchmark measures for a specific list of rare submeasures, i.e., rare parity measures essentially become benchmark measures.

Pacific did not agree to use the sample size at whatever number of cases is available after three months if a CLEC does not have thirty cases. Stating that neither the CLECs nor the ILECs have proposed that sample sizes less than five (5) be considered for assessing remedies, Pacific did not want to set the minimum sample size at one (1) case. Pacific argued that aggregating over months introduces additional complexity and accounting expenses into the measurement reporting process and that a simpler rule for sample size examines results over one month. Pacific concluded that "while it may be possible to program these aggregation rules, they will make it difficult for the CLECs to monitor Pacific's performance and difficult for Pacific to manage its business." (Id. at 13.)

The CLECs disagreed with using a minimum sample size as large as thirty (30). They argued that many CLECs would have fewer than 30 observations in a month for some measures. They also noted that Pacific reported that in the period of July through November 1999, approximately 100 CLECs had reportable

data on 18,555 instances of parity submeasures. Of these reported submeasures, 62 percent of the CLECs had sample sizes of less than thirty cases. The CLECs further argued that the majority of all submeasures would have sample sizes less than thirty (30). (CLECs' ACR Opening Comments at 11.) Consequently, a majority of submeasures would not be subject to incentive payments. The CLECs have suggested a minimum sample size of 5 for parity submeasures. (CLECs' ACR Reply Comments at 8.)

The CLECs advocated using permutation testing for small sizes. (CLECs' ACR Opening Comments at 10.) They also disagreed with aggregating sample sizes over three months, or any time, because the ILECs could perform poorly for more than a month without correction. The CLECs insist that the only reason to favor a minimum sample size of thirty (30) for measured variables is that this might make a normal distribution an acceptable approximation to the distribution of the Z-test. Regarding minimum sample sizes for benchmark measures, the CLECs continued to advocate use of the table as the cleanest, easiest means of maintaining consistency with the adopted benchmarks. (Id. At 11.)

Verizon CA stated that aggregating small sample sizes over three months raises some potentially difficult and complex implementation issues. It advocated the standard Z-test with unequal variances employing exact distributions. For parity measures, Verizon CA favored using exact distributions for small sample sizes less than fifty (50). It also supported the Pacific-CLECs tables for benchmark measures with small sizes. Verizon CA disagreed that 30 observations for parity measures are appropriate with the Modified Z-test. It maintained that neither the standard nor Modified Z-test should be used with less than fifty observations. (Verizon CA ACR Opening Comments at 15.) ORA

commented that the minimum sample size is not a "trivial issue" that should be arbitrarily set at thirty. It recommended a minimum sample size of 20 based on a formula (N (sample size) = 1/a where a = .05). (ORA ACR Opening Comments at 8.)

Use of Ten Percent Alpha Level versus Fifteen or Five Percent

Pacific argued against the use of the 10-percent alpha limit and instead proposed a 5-percent Type I maximum error rate. The company asserted that a 10-percent alpha limit is unreasonably large and will yield an unfair proportion of Type I errors. It maintained that 5 percent represents a just compromise between unfairly detecting discrimination where none exists (Type I error) and failing to detect discrimination where it exists (Type II error). (Pacific ACR Opening Comments at 14-15.) Pacific focused on their desire to mitigate the effects of random variation. It commented that forgiveness rules help with the mitigation of random variation, but are complex and expensive to administer.

The CLECs continued to recommend an alpha value of 15 percent. They contended that it is a reasonable approximation of an alpha value that will balance Type I and Type II errors. The CLECs assert that they cannot ignore the impacts of a large Type II error. They also stated that any risk adjustment, such as a forgiveness plan, must reflect the alpha chosen by the Commission. The CLECs argued that an alpha value that more easily detects discriminatory behavior combined with a valid mitigation plan can achieve the goals of a high-powered test while minimizing payments under parity conditions. (CLECs' ACR Opening Comments at 13-14 and Reply at 10.) They agreed that there is no statistical reason why a 10-percent alpha cannot be used. In addition, they recommended that the Z-test for all parity submeasures be calculated throughout the six-month pilot test period at the five, ten, and fifteen percent levels in order

to determine how many submeasures pass or fail depending on the critical value chosen. (CLECs' ACR Opening Comments at 13.)

Verizon CA commented that a 5 percent alpha remains a more balanced and reasonable choice. It asserted that a 10-percent critical value leads to a greater number of instances where a finding of "no parity" will follow from application of the test, when in fact, parity service is present. However, Verizon CA concurred with the CLECs that the result should be examined at all three proposed levels: five, ten, and fifteen percent. (Verizon CA ACR Opening Comments at 17 and Reply Comments at 4.)

ORA stated that an alpha level of 10 percent is simply too large. It argued that a more standard alpha level of 5 percent should be used. ORA stated that the use of a larger than normal alpha level means an increase in the probability of incorrectly declaring that the ILEC is out-of-parity. ORA urged the Commission to reject multiple alpha values as an attempt at data mining. (ORA ACR Opening Comments at 13.)

ORA also noted that the proposed remedies plan has no provision to prevent service deterioration, thus posing an unacceptable risk to ratepayers. It asserted that service levels can only be maintained if standards are based on prior historical data and not on future data. Performance measures used in the test period should be limited to those measures in which there is historical data available on at least twenty (20) observations. One of the two major goals that ORA identified for the Performance Remedies Plan is to maintain service levels at least at historical levels for all ratepayers. Its other goal is to ensure that customers of both the CLECs and the ILECs receive “statistically equal” service. Finally, ORA insisted that a benchmark should also be based on historical, and not present or future data. (ORA ACR Opening Comments at 6.)

March 2000 Workshop

In their reply, the CLECs recommended that the Commission hold a workshop on the new Pacific "white paper" proposal. ORA recommended that the Commission convene workshops to review all the various proposals. In all, the comments raised several issues requiring further discussion. To respond to the recommendations and address the unresolved issues, the assigned ALJ and staff facilitated a three-day workshop on March 27, March 28 and March 30, 2000. The workshop was divided into three daylong segments devoted to exploring the respective new Pacific and ORA plans, and further refining the components of a hybrid model.

The three segments of the workshop focused exclusively on the performance assessment part of the overall performance remedies plan (i.e., performance measurement, performance assessment, and incentive payment parts). The sessions did not include any substantive discussion of the performance measurement and incentive payment components of the remedies plan.[41]

For the purposes of the workshop sessions, the parties were to assume as given all prior work on performance measurements and benchmarks (on the separate parallel track pursuant to Commission Decision 99-08-020), including any current constraints. Parties were also to assume that any emergent performance measurement plan would use the performance measurements and

[41] By ruling, the assigned ALJ advised the parties that the Commission would address the incentive components (including incentive structure, incentive amounts, and who receives incentive payments) after it determined the performance measurement and assessment plan components.

benchmarks resulting from the concurrent performance measurement phase of the proceeding. Finally, the parties were asked to delay incentive payment data modeling until the Commission selected the performance assessment model, or models.

The goal of the workshop was to develop more fully the three distinct performance measurement plans. These three plans were (1) the Pacific plan, (2) the ORA plan, and (3) a hybrid plan. All workshop participants were to assume on each specific plan's day that they were advocates for that particular plan and that all participants would be jointly developing the "best" possible model for that specific plan type (i.e., hybrid, ORA, or Pacific). Where there were problems with various aspects of any plan, participants were asked to cooperatively recommend potential solutions for those deficiencies.

Participants also were asked to jointly determine if any of the plans were "fatally flawed" in any area, and if so, why. They were asked to follow the plan principles presented in the November 22, 1999 ACR, and to assume that the task before them would be to refine each particular plan type so as to be practical, capable of implementation and as simple as possible. Workshop participants were given an opportunity to advocate on behalf of their own plan on that specific plan's day, and to critique a competing plan on that plan's day. However, the intent of the sessions was to help refine each plan so that any one or all could be applied during the six-month pilot test period.

For each of the three plans, the assigned ALJ and staff focused on the respective model, element by element. There was a "straw man" or hypothetical proposal for each model element and either (1) a group decision was reached on that element or (2) a group modification was made to the hypothetical proposal. Discussion remained on each model subcomponent until a group "best" decision

was reached, or it was evident that no decision could be reached and that the participants could only "agree to disagree." At the end of each plan subcomponent, a court reporter transcribed the group's findings on that plan element for the record.

Workshop Recommendations and Positions

Hybrid Performance Measurement Plan

At the workshop, Pacific, Verizon CA, the CLECs and ORA all agreed to use the Modified Z-test to develop a hybrid performance measurement plan.[42] Most of the parties also agreed that since they had selected the Modified Z-test, the use of a two-step standard Z-test procedure and other modifications[43] were no longer applicable in terms of the "Hybrid model." Verizon CA, however, supported using permutations, deltas and exact distributions in conjunction with the Modified Z-test.

The CLECs agreed to the initial hypothetical recommendation to treat benchmarks as limits without relying on statistical tests. Pacific and Verizon CA concurred with this as long as special tables based on statistical charts are used for all benchmarks. Pacific and the CLECs further agreed to produce two sets of consensus tables of acceptable misses for sample sizes scaled from 1 to 100 at a 10-percent alpha level. One set of the tables would represent percentage-based benchmarks, and the other would represent average-based

[42] Parties' consents to develop a hybrid plan did not imply their agreement with any resulting hybrid, as each party qualified its consent. For example, in response to the draft decision, ORA stressed that it did not support the hybrid plan or the Modified Z-test. ORA Comments at 6 (December 18, 2000).

[43] Such as the unequal variance Z-test, exact distributions, permutations, and deltas.

benchmarks. ORA opposed the proposition of treating benchmarks as absolutes without reliance on any statistical testing. (Reporters' Transcript (RT) at 1107, lines 10-24.)

All the parties assented to the second hypothetical Hybrid model recommendation that the Commission re-evaluate the benchmarks after a six-month pilot test period. However, Pacific's concurrence was subject to some preliminary data calibration occurring prior to the pilot. Moreover, the CLECs stressed that real penalties and incentives should be in effect during the six months.

In the discussion on sample sizes, Pacific, Verizon CA, the CLECs and ORA all supported the hypothetical recommendation that the time period for measurement of the sample be kept on a monthly basis. The second recommendation was that each party should precisely specify what minimum sample size it selects between five (5) and twenty (20). Pacific stated that it would go to a sample size of 5, with the proviso that there be mitigation measures to offset such a small sample size. Pacific further maintained that although it would apply the Modified Z-test for parity measures down to a minimum sample size of 5, it would not agree to use data or apply a permutation test below 5. Pacific argued that permutation testing was costly. In substantiation, Pacific agreed to submit operational cost calculations for permutation testing[44].

The CLECs selected a sample size of 5 and declared that if the minimum sample size were to be below 5, they would prefer permutation testing

[44] 2000 Pacific Workpaper #9

to be used. ORA would support a minimum sample size of 5; however, below 5 it would not support using the data. Verizon CA would support a minimum sample size of 20 with permutation testing. Below, Verizon CA would prefer to discard the data. Between 5 and 20 Verizon CA would prefer to use permutation testing, but without being subject to incentive payments. (Verizon CA May 5, 2000 Reply Br. at 5) Verizon CA strongly advocated permutation testing, and agreed to jointly submit with the CLECs after the workshop a description of a permutation testing protocol[45].

Following the 1999 performance incentive workshops, the parties identified six sub-measures[46] as "rare sub-measures." The parties purported to have agreed that there would not be an application of the minimal sample size to those measures or sub-measures identified as "rare." However, it was unclear from the briefs submitted after the 1999 workshops whether the parties still agreed as to what constituted the list of rare sub-measures. Thus, the third hypothetical sample size recommendation was to identify the measures or sub-measures upon which there was agreement that there would not be an application of the minimal sample size.

The parties agreed that rare measures or sub-measures would be those that rarely saw activity, yet were important to the CLECs. Pacific and the CLECs agreed to reanalyze the issue and submit as a workshop document any suggestions, additions or deletions to the group of six rare measures and

---

[45] 2000 GTE Workpaper #8.

[46] Sub-measure Nos. 26, 27, 30, 40, 41 and 43.

submeasures.[47] The rare measure list identifies those measures (or submeasures) where the measure would still be used at a sample size of one.

The parties also discussed how to make the Hybrid model operational for parity measures with no permutation testing and with sample sizes between five and twenty. To further the analysis, Pacific acceded to provide in two parts the "data on sample size for CLECs by submeasures." Pacific specified that one part of the analysis would show the percentage of the total data elements that would be used (not discarded). The second part of the analysis would show the percentages for the resulting sample sizes that would be used, relative to the entire set of samples. The company also offered to provide the absolute numbers, not just the percentages, for the previous two months of data.[48] (RT at 1135, lines 12-28.)

Pacific suggested that staff consider different remedy amounts when analyzing this data in the context of the "small sample world" versus the "large sample world." It questioned the reliability of the data if used with certain of the recommendations in the small sample realm. The CLECs proposed two recommendations to make the Hybrid model operational. First, small CLECs could be pooled into a sufficient aggregation to meet the minimum sample size. Second, a "mean plus standard deviation" similar to the ORA proposal could be used. (RT at 1136, lines 7-12.) Verizon CA supported the small CLECs pooling proposal, stating that it merited further exploration. (RT at 1136, lines 13-15.)

---

[47] 2000 Pacific / GTE Workpaper #10

[48] 2000 Pacific Workpaper #12.

Staff set forth two hypothetical recommendations on the Commission model's alpha level. Staff proposed that a 10-percent alpha level be used for the Modified Z-test. All the parties agreed to compromise at the 10-percent alpha level for the sake of developing the Hybrid plan. To the second proposal, that parties not calculate multiple alpha levels going forward, Pacific alone agreed to refrain.

In the January opening comments on the Hybrid model proposal, Pacific asserted that certain performance measures are based on failure rates for which no standard deviation has been defined. Thus, while a test similar to a Modified Z-test might be crafted for most of these measures, a Z-test could not be calculated for at least one of them as currently defined. (Pacific Opening ACR Comments at 5, footnote 5.) During the discussions on the Hybrid model the parties identified Measures 15, 16 and 19 as measures that might require special treatment or alternative application rules. At the conclusion of the Hybrid model discussion, Pacific, the CLECs and Verizon CA agreed to recommend a common solution to staff for these three measures.

Pacific, the CLECs and Verizon CA each detailed their respective lists of necessary enhancements to the Hybrid model. Pacific identified three necessary elements. The need to: (1) mitigate for random variations; (2) develop a procedure that deals with such excludable events, such as force majeure; and (3) establish an absolute cap for maximum exposure. Pacific noted that it was willing to pay up to $120 million in payments without evidentiary hearings in its latest incentive proposal. (Pacific ACR Reply Comments at Appendix 1.)

The CLECs maintained that their essential enhancement to the Hybrid model would be to convert all benchmarks associated with averages into percentage-based benchmarks. As a result, the benchmarks would be simplified

and unified into one category.[49] Verizon CA specified five necessary enhancements to the Hybrid model. It would like the Hybrid model to either consider or perform correlation measures during the six-month trial period. Verizon CA would like the Hybrid to treat small sample sizes as they are being treated under the Bell South model.[50] It would also like the Hybrid model to consider real customer materiality[51] in contrast to statistical measurement differences. Verizon CA emphasized that all of the different measurement components are tied together, and some of its parts may have an aggregate effect that the Hybrid needs to consider. Finally, Verizon CA asked the staff to consider Pacific's white paper proposal as a tool to resolve many of the sample size issues or to satisfy the concerns about mitigation.

ORA Performance Measurement Plan

Foremost, ORA's plan attempted to adhere to the ACR's core guiding principal that any model under the Performance Remedies Plan be simple to implement and monitor. Thus, the first ORA proposal stressed simplicity as one advantage of its model. During the facilitated workgroup

---

[49] The CLECs stated that they would also be proposing this within the Performance Measurement Phase of this proceeding.

[50] "Statisticians for Bell South and AT&T have recently proposed a 'correction' to the Modified Z test that accounts for the skewness in the underlying distributions. They believe that this correction makes the 'modified-modified *t*' essentially equivalent to permutation testing at small sample sizes." Verizon CA Opening Brief at 23 (April 28, 2000).

[51] "Customer materiality" refers to whether the customer could actually perceive a difference between the performance to ILEC customers versus to CLEC customers, regardless of any statistical difference.

discussions, Pacific noted that while striving for simplicity was one of its concerns, there are more pressing substantive issues. The CLECs urged completeness and effectiveness in the remedies plan over mere simplicity. Verizon CA stated that the emphasis should be on fairness and accuracy, and simplicity should be one of several core principles. However, it asserted that if there were two plans equally effective and fair, it would prefer the simpler plan. Ultimately, Verizon CA suggested, ORA's plan may not be operationally simple.

ORA observed that the ILECs and the CLECs have proposed a mixed system with benchmark measures without any statistical tests to determine performance failure for some measures. ORA opposes using a mixed system. It argued that the same system should be applied to all performance measures, and that statistical tests are either relevant or they are not. (ORA ACR Opening Comments at 4.)

In its white paper proposal submitted in January, Pacific embraces the concept of a "same" system for both parity measures and benchmarks. However, Pacific asserts that benchmarks without statistical tests demand of the ILEC an unreasonably higher standard of performance (to avoid missing the benchmark) in the context of small sample sizes as opposed to large sample sizes. In contrast to ORA, Pacific asserts that statistical testing is relevant.

Both ORA and Pacific propose moving to a uniform system, but in different directions. The Pacific white paper plan advocates converting every performance measurement to a statistical test. The ORA plan advocates converting every performance measurement to a simple means and variance analysis, without any more formal statistical tests. The CLECs disagree that there is a need for a "same system." They contend that parity measures and benchmark measures need to be treated differently. Finally, Verizon CA notes

that while the second ORA recommendation of consistency in terms of a "same system" concept is laudable, it is unnecessary.

The ORA plan argues that there are no provisions to prevent service deterioration in the Performance Remedies Plan. It states that current service levels can only be maintained if standards are based on historical, rather than future data. The current plans may have built-in reversed incentives such that if the ILECs were to increase the variability of their own processes, they could reduce incentive payments even though the CLECs receive worse performance. That is, the poorer the ILEC performs, the poorer the parity performance for the CLECs, but the larger variability would effectively prevent discrimination detection. To militate against this possibility, one of the straw man recommendations under the Hybrid plan was to monitor ILEC means and variances and compare them to historical values[52].

Responding to ORA's recommendation to base standards on historical data, Pacific questioned how the historical period would be defined and how the historical data concept could be operationalized. Pacific stated that it saw the ORA white paper as a conceptual approach that had not yet been specified to an operational level. It also requested a more detailed description of what monitoring ILEC means and variances would entail.

The CLECs advised that when one uses historical data in the context envisioned, there is a need for a lot of data. Overall, the CLECs were content with real-time data over historical data. However, they support monitoring the

[52] Pacific and Verizon CA agreed to provide staff with the incumbent local exchange carriers' historical means, variances and sample sizes for their retail parity measures and submeasures from September 1999 going forward through June 2000.

means and variances in order to mark improvement in Pacific's performance and to record where the CLECs stand in terms of Pacific's historical performance. Verizon CA noted that the data fluctuates substantially from month to month. Verizon CA maintained that there are inherent limitations in the depth and breadth of historical data necessitating adjustments. In addition, Verizon CA supported continuing to monitor the ILEC means and variances.

In its white paper proposal, ORA argued that neither Z-test, nor any other parametric test, should be used during the performance remedies plan six-month pilot period because many of the underlying performance measurement series are not distributed normally.[53] ORA argued that such abnormal distributions violate a fundamental assumption of the Z-test. Pacific supported using the Z-test during the pilot. It indicated a willingness to look again at the Z-test after the pilot, but wanted more specifics on what this would encompass.

Verizon CA commented that ORA's proposal to reject all statistical tests during the pilot is too extreme. Yet, it acknowledged that ORA's concern about normality was justified. Verizon CA suggested that ORA's approach should be considered after the six-month pilot is completed. At the workshop, Verizon CA cautioned that two factors should be taken into consideration. First, how to calculate the test statistics; and, second, how to use the calculation. Verizon CA noted that given assumptions of normality are met, one could consult "look-up" tables. Outside the range of normality, one could use permutation testing and exact distributions.

---

[53] As Pacific characterizes it: "no normal distributions and relatively few large samples." In fact, "samples" in question may not really be "samples," but rather time-series population observations.

The CLECs alone directly addressed the ACR's questioning of the use of any Z-tests. The CLECs recommend the use of existing parametric tests. However, they maintained that if actual experience does not justify confidence in the results, the test simply should be based on the number of observations that fall above some specified level. Essentially, this would convert measurement cases into counting cases. At that point, the CLECs propose to use the upper ten percent quantile[54] of the observed ILEC sample. CLEC statistical expert Dr. Colin Mallows of AT&T performed simulations and found that for some alternatives this non-parametric test is much more powerful than the Z-test. (CLECs Reply ACR Comments at 4-5.) In the workshop, the CLECs supported using "some flavor" of the Z-test during the pilot.

The ACR urged moving toward more aggregation of the measures over time in order to simplify the performance remedies plan. The aggregation effort should take all double counting out of the measures to the extent that there is correlation and interdependence between a number of the measures. In response, ORA stated that there are a total of 44 performance measures with over 1000 submeasures. It expressed concern about possible correlation between these measures. ORA argued that the ILECs' OSSs could be adequately measured with fewer performance measures, since many of the measures may be cross-correlated with each other and may not be needed. ORA's plan recommends that correlation tests be run for all the performance measures. It

[54] A quantile is a portion of a distribution. An upper ten-percent quantile designates the highest ten-percent of results in a distribution, i.e., those results above the 90th percentile.

also submits that no performance measure should be included if it has a correlation greater than 0.80 with any other performance measure.

During the workshop, Pacific supported the hypothetical ORA plan recommendation for correlation testing. Pacific agreed that eliminating measures would help. To date, there has been no correlational statistical analysis or scientific modeling of the measures. However, given the contentiousness surrounding the issue, Pacific is willing to address the matter at a later time. Pacific admitted that the issues of correlation and interdependence had not yet been raised in the Performance Measurement Phase.

The CLECs pointed out that there was not a lot of data until recently to determine correlation. They do not want to get sidetracked with correlation issues at this point. While not adverse to a goal of reducing or adding measures if there is a legitimate rationale, the CLECs are opposed to a casual reduction of measures. They maintained that, at this point, Pacific and the CLECs see no further correlation between any of the submeasures. Verizon CA concurred with the plan recommendation as well as the ACR's desire to reduce the number of performance measures, if supported by the data. It asserted that the data is not currently available, and will not be available until after the six-month pilot. Verizon CA stated that the performance incentive phase would be the proper forum to address the issues of correlation and interdependence.

ORA's plan recommends a minimum sample size of twenty (20). It argues that a performance measure should only be used in the pilot if two requirements are met. First, that it satisfies a minimum sample size of twenty; and, that the measure is not highly correlated (greater than 0.80) with any other measure. At the workshop, Pacific, Verizon CA and the CLECs opposed ORA's recommended minimum sample size.

ORA's plan also recommends that parity be defined as a situation in which the average measured results for the CLECs served by a particular ILEC are within one standard deviation of the average measured results that the ILEC provides to its internal company units. ORA proposes that the ILEC average be based on historical and not future data.[55]

In terms of the workshop discussion, ORA's recommendation was to use the most recent historical fiscal or calendar year for the ILEC. None of the other parties supported ORA in its selection of one standard deviation. Assuming that ORA refers to one standard deviation of the mean, a one-tail test, and assuming normality, one standard deviation corresponds to approximately 84 percent the normal distribution, or a 16-percent alpha. However, this interpretation is somewhat inconsistent with the Office's prior recommendation of a 5-percent alpha, at least for large samples. For a one-tail test a 5-percent alpha corresponds to approximately 1.645 standard deviations. Assuming ORA was referring to the standard deviation of the mean, to facilitate the workshop discussions staff proposed using a 10-percent alpha or approximately 1.282 standard deviations for the sake of developing the ORA model.

However, a close read of ORA's proposal shows that ORA refers to one standard deviation of the observations, not one standard deviation of the mean.[56] In this case it is not possible to determine one critical alpha level

---

[55] In its comments on the draft decision, ORA states that historical data should be used for "the longest period for which data is available." ORA Comments at 7 (December 18, 2000).

[56] Statistical notation consistently used by ORA indicates its plan is based on one standard deviation of the observations: σ. *See* ORA Comments at 15 (December 18,

Footnote continued on next page

equivalent even with normal distributions, as one standard deviation of the mean is a function of the standard deviation of the observations and the sample size.[57]

ORA proposed that the benchmark be defined as the historical mean of the series plus one standard deviation. Consequently, any performance worse than the benchmark would trigger a penalty. ORA argued that the best demonstration of parity would be actual, not estimated, performance, even when experts using reasonable information make the estimates in good faith. The Office contended that proxies could be used in place of benchmarks in many cases. Since they are based on actual data, proxies are clearly superior to arbitrary benchmarks. ORA recommended that the Commission investigate their use before adopting arbitrary benchmark measures, and urged that benchmarks be used only in cases where there are no retail analogues and no proxies for those retail analogues.

Pacific, Verizon CA and the CLECs rejected ORA's benchmark proposal. They maintained that the reason they initially established benchmarks was because there were no retail analogues. Technically, there is no historical time-series data to calculate the mean and standard deviation for benchmarks under ORA's definition. Ideally, normalizing the benchmarks through proxies (assuming fairness and simplicity) is preferable to the current negotiated values. However, re-creating benchmarks distinct from those established in D.99-08-020

2000) and Reply Comments at App. A, at 4 (December 22, 2000). One standard deviation of the mean (standard error of the mean) would be designated: σ_m.

[57] *See* W. Hays, Statistics at 214-215 (5th ed. 1994).

would be impractical, contentious and time-consuming at this juncture. The parties accepted staff's recommendation to treat benchmarks as limits, as agreed to in D.99-08-020, in the context of the ORA plan.

Finally, staff asked the parties to help identify any other requirement conditions that need to be specified to make the measurement component of ORA's plan operational. In response, WorldCom introduced the "SiMPL Plan"[58] during the workshop. The SiMPL Plan would calculate the ILEC's historic performance percentiles and compare the relative CLEC performance results in those intervals. For example, non-parity would be identified if more than 10 percent of the CLEC's results were above the ILEC's 90th percentile. Other percentile comparisons would be made as well. WorldCom explained that this feature could assist in shaping CLECs' service expectations. It also contended that the plan is easy to administer since ILEC compliance is based upon whether the count of ILEC and CLEC events within each of three performance zones is proportional. (2000 MCIW Workpaper # 3 at 4.[59]) WorldCom characterized the SiMPL Plan as an alternative to the Modified Z-test in furtherance of the workshop assignment to collaboratively refine each model into the best that it could be. (Post-Workshop Opening Brief of AT&T, Covad, MGC Communications and WorldCom at 4-5.)

Pacific objected to WorldCom not presenting the SiMPL Plan in writing in advance of the workshop, and asserted that it saw "only minimal

---

[58] The Simplified Measurement of Performance and Liability Plan. 2000 MCIW Workpaper No. 3.

[59] Dr. George Ford's paper on the SiMPL Plan.

connections, at best" between the SiMPL[60] and ORA plans. (Pacific Opening Comments on Performance Remedies Workshop at 6.) Pacific described the SiMPL Plan as "fatally flawed"; simple only in that it does not require statistical testing to make the final determination of which measures were missed; and "inherently unfair to the ILEC." (Id. at 6-7.) Pacific concluded that the net result of the SiMPL Plan would be either to guarantee superior service to the CLECs or to plunge the ILEC into a spiraling series of costly service improvements that ultimately would not shield it from remedy payments. (Id. at 7.)

Pacific's White Paper Proposal

Pacific's revised Performance Remedies Plan, issued in January 2000, incorporates a number of new principles. First, Pacific maintains that there should be minimal payment of remedies when the ILEC provides parity service that is compliant with the standards of acceptable performance. This revised principal is similar to the ACR principal that "the plan should impose smaller penalties on Pacific for discriminatory performance that could be merely the result of random variation, and impose larger penalties for seriously deficient performance." (ACR at 12.) The ACR recognized this principal as a relative one, offset by benefits that the ILEC receives when it is not actually providing parity service but also is not measured as out of compliance.

Underscoring its first new principal, Pacific states as a supporting principal that the plan should not provide incentives for the ILEC to engage in behaviors to escape remedy payments other than performance improvements. It also insists that the plan should provide payment to the CLEC only for poor

[60] Pacific refers to it as "the Ford Model."

performance by the ILEC and not as a normal course of business. Further, Pacific restates the CLEC principal that the risks of Type I and Type II errors should be shared equally between the CLECs and the ILEC. Finally, Pacific asserts in its revised plan that samples of various sizes should be used provided the data they supply support valid decision rules.

Pacific's revised plan distinguishes between two definitions of parity service delivery. The company selected the definition that it contends recognizes and incorporates the variability of service delivery processes, i.e., the impossibility of delivering service exactly the same way every time. Thus, Pacific prefers the assertion that "parity of service delivery is achieved whenever the results for the CLEC and the ILEC are not *significantly* different." It notes that the key is to find a way to operationalize the meaning of "significant" when applied to ILEC and CLEC results. Pacific states that this is a statistical question that may be answered using models of the processes that produce the data to be evaluated. It is possible to calculate the probability of observing any particular difference between the results of the ILEC and CLEC given the assumption that parity service is being delivered. The probability of the observed difference in results is the mechanism for deciding the significance of the difference between ILEC and CLEC.

Pacific's white paper proposal advocates a definition of compliance that it maintains diminishes the disadvantages of measuring compliant service where there are no retail analogues. Instead of comparing CLEC results in absolute terms against a benchmark, CLEC results are compared in relative terms against a standard. CLEC results are compared to a standard using a statistical test to evaluate the compatibility of those results with the standard.

Consequently, "the results for the CLEC are compliant if they are not *significantly* different from the standard."

Pacific's revised plan contends that a key aspect of the use of standards and statistical tests is that the same criterion for the probability of failure (under the assumption of compliance) can be used as is used for parity measures. (Pacific's Opening Comments to the ACR, Attachment I at 4) Moreover, this probability can be made nearly constant for all sample sizes. Pacific disputed the CLEC's assertion that introducing standards at this late stage of the development of the remedy plan threatens to jeopardize all the difficult negotiations that went into the setting of benchmarks. The company insists that all standards may be derived from already agreed upon benchmarks by using a straightforward, objective formula.[61] The agreed upon benchmarks would remain intact and both sides would reap the benefits of using standards. (Id.)

In the revised plan, Pacific continues to propose a 5-percent alpha for parity measures. The white paper is not clear on what alpha level equivalent Pacific recommends for benchmarks with statistical tests. Pacific also contends that it is willing to go to a minimum sample size of 5 for parity measures, provided its white paper proposal for benchmarks is used. It recommends using the same minimum sample size of 5 for benchmarks.

Finally, Pacific recommends setting aside the forgiveness rules of its original plan, and sets forth an alternative mechanism for mitigating random variation. With this mechanism, Pacific proposes to focus on the CLEC as the unit of analysis and determine whether the total relationship between the ILEC

[61] Id. at 20, Appendix III.

and the CLEC shows evidence of discrimination or whether any failures observed can be ascribed to random variation. (Id. at 14.) Thus, Pacific would use a table to evaluate all the sub-measures for a single CLEC in lieu of forgiveness rules.

At the workshop, the CLECs disagreed with a performance assessment that uses statistical significance testing on benchmarks. They maintain that such a focus increases the complexity of the FCC's "a meaningful opportunity to compete" standard. The CLECs also contend that benchmarks are a surrogate for parity. Thus, benchmarks should not be treated the same way as parity measures. The CLECs support the existing treatment of benchmarks as tolerance limits not targets, as Pacific's plan would suggest. (RT at 1170-72.) Further, the CLECs continue to assert that there is a need for a mitigation plan for both Type I and Type II errors, and that all submeasures should be treated the same over time regarding both these categories of errors. (RT at 1170, lines 16-20.)

Verizon CA argued at the workshop that overall it supported Pacific's white paper model; however, it would like to see how certain specific elements of the model would be implemented. Verizon CA prefers permutation testing below a sample size of 50, and thinks the Modified Z-test down to a sample size of 5 presents problems. Within the context of the Pacific model, Verizon CA favors a 5-percent alpha and supports the concept of benchmarks with statistical testing. (RT at 1174, lines 7-24.)

ORA, noting concerns about the assumptions inherent in any parametric testing, reiterated that if the Commission adopts either the Hybrid or Pacific model we base them on historical data. ORA also suggested that we reassess the choice of alpha level, specific level of benchmarks, and the values of

the small sample tables when more historical data becomes available. Moreover, ORA did not accept Pacific's argument that false negatives (Type II error) are unimportant because they do not harm the CLECs. It stated that performance incentives are fundamentally aimed at encouraging ILECs to provide parity of service and to dissuade attempts to discriminate, with the goal being to allow competition to proceed uninhibited. The fact that the attempted discrimination was unsuccessful does not mean that the performance incentive plan should not consider the attempt. (ORA Opening Brief at 3.)

Selection of the Decision Model

Our task now is to select a decision model consistent with several levels of policy goals. At the highest level, our model must effectively assist in converting a historical natural-monopoly market to a competitive market. This requires us to ensure that incumbents allow nondiscriminatory access to their infrastructures so competitors can provide local telephone services. That is, the CLEC's customers must not receive significantly worse performance from the ILEC than the ILEC's customers receive. Our decision today is at an even finer level of detail. We must specify a model that will accurately assess and identify discrimination. We must specify accurate calculations, accurate analyses, and accurate discrimination-identification decisions.[62]

We have reviewed the proposed models and the parties' comments regarding each of these models. While we had hoped that the parties would agree on a model and all the necessary implementation specifications, this did not occur. To the contrary, the parties disagreed on the models and on most of

[62] We assume accurate data. Data accuracy is a topic in parallel proceedings.

their elements. While the workshop hybrid model[63] seemed to come closest to a successful compromise, the parties did not fully endorse it. At best, each party accepted the proposed hybrid model only insofar as we would modify it to address their particular interests.

Thus, we must review and approve or reject proposed models and/or elements, especially to resolve issues where there was no agreement. Unfortunately, virtually all model specifications by each party generated disagreement from at least one other party. The following is a list of the issues we must resolve now to specify the decision model for the next phase of this proceeding.

- Shall we select the workshop hybrid model, or any party's decision model, in its entirety, or should we select the best elements of different models to create a new hybrid?
- What statistical test[s], if any, shall be used to assess parity measures, including average, percentage, and rate measures?
- Where statistical tests are used, what decision criteria shall be used to identify results as parity or non-parity, or in other words, what criteria shall be used to identify test passes and failures?
- Shall a determination of material differences be a factor in non-parity identification?
- What sample size rules should be used?
- Shall data be transformed to closer approximate statistical test assumptions?

[63] When we refer to the "workshop hybrid model" we are referring to the outline model first described in the ACR, then subsequently revised in the workshops. Beginning with modifications in the workshops, this model was referred to as the "hybrid model" since it incorporated components from the different models.

- Shall benchmarks be used as limits or as targets, and shall statistical tests, or tables based on statistical analyses, if any, be used for: (1) Some benchmark measures, (2) All benchmark measures?
- Shall correlational analyses be employed to assess and reduce redundancy between performance measures?
- Shall historical data be used as a decision criterion, or be monitored separate from the identification of passes and failures?
- Shall existing benchmarks be modified to address new developments in this assessment phase of the proceeding?
- Should we specify different models for the different ILECs?
- Should we plan to adjust payments retroactively after the six-month trial period?
- What other specifications should we order to enhance the use and understanding of our decision model?

We will base this decision on the following criteria:

- Accuracy: Identify discrimination when it exists, and do not identify discrimination when it does not exist.
- Correctability: When more important criteria do not provide conclusive guides to our decisions, we will select the elements that offer the most opportunity for correction in later phases of this proceeding.
- Academic soundness: Our rationale shall be based on recognized applicable statistical assumptions and principles, and confirmed by data when possible.
- Policy goals: Our rationale will be consistent with competition-enhancing policy and law providing substantially equal access for all potential local phone service providers, whether small or large.
- Simplicity: Without sacrificing higher-order goals such as accuracy, we will prefer the more simple models and elements.
- Fairness: We will strive to be as even-handed as possible to optimize competitive market potential and benefits.
- Openness: We will document and explain the criteria we use in selecting the model and its elements so that all parties can

knowledgeably comment and knowledgeably argue for modifications to the model.

- Consensus: We will prefer models or elements where a consensus exists, unless there are differences on more important criteria.
- Experimentation: Rather than consider the initial model to be a final product, we will consider this initial implementation to be an experiment that will inform future model development.
- Costs: Unless a more costly model or element is likely to better satisfy important criteria, we will prefer less costly approaches.
- Understandability: When differences on more important criteria are minimal, we will prefer more easily understandable models and elements. We will also take care to explain models, elements, and analyses in sufficient detail and at a level to help the reader understand the model we specify and the reasons we have selected the model and its elements.

From the parties' proposals and comments, relevant statistical sources, and staff's analyses, using the above criteria we have selected a decision model.[64] The model is presented in Appendix C. The following is a discussion of the model and our rationale for selection of the various model elements.

Decision accuracy

While the above criteria lists may seem self-explanatory, we believe it important to discuss at length the first and most important criterion, decision accuracy. We begin with a brief overview.

Once performance measures are established and results are obtained, accurately assessing the existence of competitive conditions then becomes a

[64] Accordingly, we take official notice of several academic sources. They are referred to throughout the following discussion and are listed in Appendix B. Additionally, we take official notice of several analyses performed by staff which are included as appendixes to this Decision.

decision-making task. Since these decisions must be self-executing, the Commission must construct a decision model that can automatically identify performance result levels that reveal competition barriers and that will trigger incentive payments. There are two fundamental categories of performance measures that must be assessed. These categories' definitions are based on the characteristics of the service an ILEC provides a CLEC and the CLEC's customers. Where there is an ILEC retail analogue to the service given the CLECs and their customers, the FCC has stated that parity of services is evidence of open competition.[65] Where there is no ILEC retail analogue to service given the CLECs, then open competition is gauged by performance levels that provide a "meaningful opportunity to compete."[66] These performance levels that have no retail analogue are designated "benchmarks." Thus, the two categories of measures have been termed "parity" and "benchmark" measures.

Decisions regarding parity measures

In identifying parity or non-parity, accurate remedies-plan decision-making is not simply a matter of accurately calculating average ILEC and CLEC performance and identifying non-parity if ILEC service to CLEC customers is worse than ILEC service to ILEC customers. Given that there is variability in ILEC performance in its own retail services to its own customers, a measurement

---

[65] Parity of services refers to "access to competing carriers in 'substantially the same time and manner' as it provides to itself" and "access that is equal to (*i.e.*, substantially the same as) the level of access that the BOC provides itself, its customers, or its affiliates, in terms of quality, accuracy, and timeliness." *Bell Atlantic New York Order* ("FCC BANY Order"), 15 FCC Rcd at 3971, ¶ 44.

[66] *Id.* at 3971-72, ¶ 44-45.

result of inferior service to CLEC customers could be due either to this variability, or actual discrimination, or both. In other words, if we sample the ILEC's service results to its own customers, we will get different results, some better, some worse than the average. Service to a CLEC may be viewed as a "sample" of the ILEC's services.[67] Theoretically speaking, if the performance measured from the CLEC "sample" is typical of the performance for similar ILEC customer "samples," then there is no evidence of discriminatory service, even if it is somewhat worse than the ILEC average. However, if the CLEC "sample" performance is worse than most ILEC customer "samples," then there appears to be evidence of discrimination.

In statistical terminology, the non-discriminatory variability between multiple ILEC samples is termed "sampling error" or "unsystematic variability," referring to the fact that the variability is simply due to random sampling outcomes. Discriminatory variability is the case where the performance in a CLEC sample is worse than what would be reasonably expected from sampling error. Discriminatory variability is variability that goes beyond sampling error and is termed "systematic variability," meaning that something is systematically causing the differences between the samples. Since these two types of variability cannot be directly observed, discrimination or non-discrimination must be indirectly inferred.

---

[67] By using the word "sample" we do not mean to imply that the correct model is a sampling model in the traditional parametric statistical use of the term. The record does not help us resolve, nor do we resolve, the underlying assumptions of whether the combined ILEC and CLEC performance results for each month should be viewed as a "sample" of an underlying process distribution, or if each month's results should be viewed as the entire population of events.

A decision outcome matrix illustrates this problem. Figure 1 presents the four possible decision outcomes about parity. The four outcomes represent conclusions of either parity or non-parity of service under conditions of either actual parity or non-parity. The decision outcome matrix simply recognizes that when we make a dichotomous decision, there are four possible outcomes, two correct and two incorrect. In the context of this proceeding, the decision outcome matrix illustrates decision goals: (1) to detect differences when they exist, and (2) to not detect differences when they don't exist.

Figure 1: Decision Matrix

	Parity Identified (Decision: No Discrimination)	**Non-Parity Identified** (Decision: Discrimination)
Reality: Parity (No Discrimination)	**Correct Decision** (True Negative)	**Incorrect Decision** (False Positive)
Reality: Non-Parity (Discrimination)	**Incorrect Decision** (False Negative)	**Correct Decision** (True Positive)

Figure 2 expands this illustration. Given that decisions regarding parity are based on measurements that are comprised of both "true" values and "error," these outcomes can represent both correct and incorrect decisions, depending on the relative amount of error in the measurement. Figure 2 portrays sampling error effects.

Figure 2: Decision Matrix Showing Sampling Error Effects

	Decision: Parity	**Decision: Non-parity**
Reality: Parity (No discrimination)	**Correct Decision** Relatively low sampling error	**Incorrect Decision** Sampling error creates spurious difference
Reality: Non-parity (Discrimination)	**Incorrect Decision** Sampling error masks real difference	**Correct Decision** Relatively low sampling error

Figure 3 illustrates the contribution of statistical testing. The potential for errors is the same as in the first two matrices where no statistical testing is applied. The only contribution of statistical testing is that it allows us to estimate decision accuracy, or in other words, to calculate the decision error probabilities. These probabilities can then assist decision-making by quantifying the different error probabilities and comparing them to standards of confidence that we wish to apply. These standards of confidence are expressed as: (1) the power of the test, and (2) the confidence level.

Figure 3: Decision Matrix with Statistical Tests

	Decision: Parity	**Decision: Non-parity**
Reality: Parity (No discrimination)	**Confidence level** Probability = ***1 – alpha***	**Level of significance** Probability = ***alpha*** Type I error
Reality: Non-parity (Discrimination)	**Test insensitivity** Probability = ***beta*** Type II error	**Test power** or sensitivity Probability = ***1 – beta***

Test power refers to the ability of the test to actually find true differences, that is, the confidence that you found what you were looking for, when it existed. "Confidence level"[68] refers to the ability of the test to reject spurious differences, that is, the confidence that when you identified something, it actually existed. Together, these probabilities represent the amount of confidence one can have in decision quality. The higher the test power, the greater the confidence one can have that true differences were uncovered. The higher the "confidence level" the greater confidence one can have that discovered differences are real differences. Other things being equal, as one level of confidence is increased, the other decreases. In other words, the more powerful the test, the more likely there will

[68] While by convention 1–alpha has been termed the "confidence level," in reality both 1–alpha and 1–beta are confidence levels. They are distinguished by the type of confidence they estimate.

also be differences found solely due to random variation, and the higher the confidence level, the more likely true differences will be missed. Neither confidence standard is inherently more important than the other. Each application of a statistical test implies different trade-offs between these two confidence standards, and their corresponding error probabilities, depending on the consequences of the two different errors.[69]

In the present case of restructuring a historical natural-monopoly market to create a competitive market, the primary function of performance measurements and the decisions about performance measurements is to detect and prevent barriers to competition. To maximize goal attainment these decisions must be as accurate as possible, to find and prevent actual barriers, and to avoid identifying barriers when they do not exist. However, there is no legislative or regulatory guidance specifying the relative importance of the two decision errors.

On one hand, if we do not detect barriers when they occur, competition may fail, and the fundamental purpose of the legislation will have been thwarted. On the other hand, if we identify barriers when they do not exist, then we are likely to take unfair punitive action. Therefore we will use statistical testing to assess the balance between these two competing outcomes, thus enabling greater decision quality and attainment of legislative goals. Figure 4 summarizes the statistical decision matrix and identifies the probabilities that correspond to the four possible decision outcomes.

---

[69] *See* W. Hays, Statistics at 267-303 (5th ed. 1994), and B.J. Winer, Statistical principles in experimental design at 10-14 (1971). We discuss these issues in more detail in a following section.

Figure 4: Decision Matrix Statistical Testing Summary

	Decision: Parity	**Decision: Non-parity**
Reality: Parity	No barriers exist. No barriers identified. (***1 – alpha***)	No barriers exist. Barriers identified. (***alpha***) Type I error
Reality: Non-parity	Barriers exist. No barriers identified. (***beta***) Type II error	Barriers exist. Barriers identified. (***1 – beta***)

Using measures of performance averages and variability, statistical analysis provides estimates of: (1) the probability that a result of a certain magnitude would be detected when it exists (test power and corresponding error *beta*) and (2) the probability that the result is due to random variation when in fact there are no differences (confidence level and corresponding error *alpha*). The methodology for using these estimates to establish dichotomous decision criteria is called null hypothesis significance testing. The analyst specifies a null hypothesis to pose that there are no differences between two performance outcomes, selects a confidence level that strikes the appropriate balance between the two types of error, calculates the probabilities, and compares them to the selected significance level. If the probability is less than the selected significance level, then the analyst rejects the null hypothesis and accepts the alternative hypothesis that there are real differences.

In the two approved Section 271 applications to date, Bell Atlantic New York and Southwestern Bell in Texas use a "Z-test" statistic to calculate these

probabilities. Conceptually, the Z-test statistic compares the ILEC's average (mean) performance to the CLEC's mean performance, and then compares the difference between the means to the difference that would be expected from random variation at a selected confidence level. The expected difference is calculated from the variability in the samples of performance. The greater the variability, the greater the expected difference, and the less likely a true difference will be detected. In the Z-test, the difference between means is compared to (actually divided by) an expected difference term that is calculated from the sample size (n) and the variability in those samples (variance).

Thus the sample size, the variability in the samples, the power of the test, the confidence level, and the size of the true differences between means affect decision quality.[70] These elements are interdependent such that changing one will have an unavoidable effect on at least one of the others. A convention has existed for several decades to pre-select a fixed confidence level (or alpha) and adjust the other elements if desired. For example, if a test with the common 95% confidence level (0.05 alpha) lacked adequate power to detect true differences, the sample size could be increased. Methods have been developed to calculate the minimum sample size required to attain adequate test power.[71]

Additionally, since much of science depends on replication, test power is relegated less attention because of the expectation that experiment replication will address this issue. However, this convention which evolved in the 1920's, called null hypothesis significance testing, has been questioned over the last

[70] W. Hays, *supra* at 289-293 (1994).

[71] For example, *see* W. Hays, *supra* at 333-334 (1994).

three or four decades. At least one professional standards board was recently established to consider abandoning such testing in favor of new methods that strike a more even balance between test power and confidence levels.[72] Illustrating this concern about ignorance of test power, the following comments reveal some of the intense dissatisfaction with current research relying on 0.05 critical alpha levels:

> Whereas most researchers falsely believe that the significance test has an error rate of 5%, empirical studies show the average error rate across psychology is 60%--12 times higher than researchers think it to be. The error rate for inference using the significance test is greater than the error rate using a coin toss to replace the empirical study. . . . If 60% of studies falsely interpret their primary results, then reviewers who base their reviews on the interpreted study "findings" will have a 100% error rate in concluding that there is conflict between study results. (p. 3.)[73]

The balance between these interdependent elements that affect decision outcome quality is problematic not only in pure research contexts, but also in applied contexts such as engineering and operations management.[74] As parties have greater vested interests in different outcomes, the greater the argument

---

[72] R. Hubbard; R. Parsa; M. Luthy, The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917-1994, 7 Theory & Psychology at 545-554 (1997).

[73] J. Hunter, *Needed: A ban on the significance test*, 8 Psychological Science at 3-7 (1997).

[74] For example, *see* C. Das, *Decision making by classical test procedures using an optimal level of significance*, 73 European Journal of Operational Research at 76-84 (1994); R. Verma & J. Goodale, *Statistical power in operations management research*, 13 Journal of Operations Management at 139-152 (1995); and K. Brubaker & R. McCuen, *Level of significance selection in engineering analysis*, 116 Journal of Professional Issues in Engineering at 375-387 (1990).

there is over the appropriate balance. This is certainly the case in the present proceeding. Parties disagree on what is appropriate for all elements: the appropriate tests, confidence level, test power, sample size, test statistic, and other elements and nuances of a statistically based decision structure.

Determinations regarding benchmarks

Unlike performance measures where there is a retail analogue, benchmarks cannot compare ILEC service to CLEC service since there is no ILEC service analog. Instead, benchmarks are judgments about the levels of ILEC performance for CLEC competitive service that are necessary to "allow a meaningful opportunity to compete." Benchmarks have been constructed as tolerance limits. For example, one measure specifies that *99 percent of billing invoices shall be available within 10 days of the close of the billing cycle*.[75] The issues for statistical analysis accuracy are not the same as for parity measures. However, small sample benchmark applications raise similar decision matrix issues that we discuss after we address the more complex issues of the statistical models for parity performance measurement results.

Statistical models

As discussed, several models for parity assessment have been presented during the course of this proceeding. Some were intended to be complete, such as Pacific's most recent model. Other models were intended to present conceptual frameworks that would resolve various problems and which could be implemented with further negotiation and development. Examples of these

[75] Performance measurement No. 30, Wholesale Billing Timeliness, D.99-08-020, *mimeo.* at 43.

include ORA's model, MCI's SiMPL model, and the ACR's proposal.[76] We find that none of the presented models are acceptable in their entirety. Our rationale for this finding is best explained by discussing our evaluation and selection of the model elements that we will specify in what will be a new "hybrid" of elements from each of the different models presented in this proceeding.

Statistical tests

Three types of parity measurements have been developed for monitoring ILEC performance: averages, percentages, and rates. Each measurement type requires a different statistical test or a variant of the same test.

Average-based measures

The choice of a statistical test for average-based parity measures came as close as any model element to being accepted by all parties. Pacific and the CLECs have agreed that the Modified Z-test should be applied to average-based measures. Verizon CA also agreed to use the Modified Z-test, albeit with modifications. Only ORA disagreed, although they consented to its use in the development of a "hybrid" model. (RT at 1103.) All parties have agreed that a one-tailed test should be used. A one-tailed test is appropriate for situations where we are only interested in outcomes in one direction, in this case where the CLEC performance results are worse than the ILEC results. This is consistent

[76] In comments to the draft decision ORA asserts that its proposal specifies an implementable model. We appreciate ORA's sincere efforts to present a simplified model which is intended to avoid recognized problems with other models, such as data distribution non-normality. However, ORA's proposal leaves unclear critical components, such as calculation of the "standard deviation" as discussed *supra*. If ORA wishes to explore its proposal further, we urge them to present explicit formulas and data examples to the other parties, and ultimately, to us during the next phase.

with academic texts[77] and with the FCC's view of the appropriate statistical application regarding the requirements of the Act.[78]

Standard Z-test

The standard Z-test compares the difference between means to what is essentially an expected difference between means that could be explained by random variation. The expected difference is calculated from the variation (variance) in both the ILEC and CLEC results. The ACR proposed that the ILEC and CLEC variances be screened for statistically significant differences as a first step, then either the pooled or equal variance standard Z-test statistic would be calculated as a second step depending upon the results of the first step. Verizon CA described several concerns with the ACR's proposed two-step standard Z-test method and suggested several corrections.[79] However, in response to the CLECs' concerns that ILEC discrimination could increase the CLEC variance, and thus make it more difficult to detect any discrimination, all parties agreed to use a Modified Z-test instead of the standard Z-test.

Modified Z-test

This test was first adopted by the NYPSC for the BANY 271-application performance remedy plan.[80] Similar to our situation, since the CLECs were concerned that by providing highly variable service to the CLECs, the ILEC

---

[77] Hays *supra* at 293-294 (1994); and Winer *supra* at 20 (1971).

[78] *Bell Atlantic New York Order*, 15 FCC Rcd at 4191, App. B, ¶ 18.

[79] Verizon CA ACR Opening Comments at Apps. A and B (January 7, 2000).

[80] *Bell Atlantic New York Order*, 15 FCC Rcd at 4182-4188, App. B., ¶¶ 1-13.

theoretically could increase the expected difference and thus mask real differences, the parties in the BANY application proceedings agreed that the CLEC variance would not be part of the expected difference calculation. This alteration has been given the name "Modified Z-test." The FCC considers this test reasonable,[81] and it has been favorably presented in statistical academic literature.[82] The FCC subsequently approved Southwest Bell's performance remedy plan for Texas, which also uses the Modified Z-test.[83]

Only ORA objects to use of the Modified Z-test, although for the purposes of developing a hybrid model, ORA is willing to proceed using the test. (RT at 1103.) ORA's primary concern is based in their opinion that use of any Z-test requires that the data be normally distributed. According to the statistical literature, this may be only partially correct; Central Limit Theorem states that for sufficiently large samples, non-normality in the data does not affect the test.[84] With large samples, the distribution of sample means will be normal, whether or not the raw data distribution is normal. The means of

---

[81] *Id.* at 4188, App. B ¶ 13 and n. 37.

[82] C. Brownie, D. Boos & J. Hughes-Oliver, *Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls,* 46 Biometrics at 259-266 (1990).

[83] *See* SWBT interconnection agreement, *Texas T2A Agreement*, Attachment 17: Performance Remedies Plan, ¶ 2.0 at 1.

[84] "If a population has a finite variance σ^2 and a finite mean μ, the distribution of sample means from samples of N independent observations approaches the form of a normal distribution with variance $\sigma^2/[\text{sqrt}(N)]$ and mean μ as the sample size increases. When N is very large, the sampling distribution is approximately normal. Hays (1994) at 251. See also, R. Khazanie, Statistics in a world of applications at 344-345 (4th ed. 1997).

sample sizes of 30 or more are typically considered sufficiently normally distributed to have minimal effect on a Z-test.[85] The BANY performance remedy plan addresses this issue by using the Modified Z-test down to a sample size of 30, and is temporarily using the t-test for smaller samples until permutation testing is established.[86] In comments to the draft decision, ORA asserts that only its proposal is consistent with Central Limit Theorem. ORA Comments at 9. We are not persuaded and remain concerned that no proposal has adequately addressed what a "sufficiently large" sample is. For example, ORA states that over time, distributions will approach normality because the number of observations will increase. However, there is no evidence that the distribution of the *observations* will be normal for very large samples.[87] Our understanding is that only the distribution of sample means will approach normality as sample sizes increase. Yet even ORA's model appears to depend on results limited to a month interval. ORA Opening Comments on the ACR at 9. Additionally, the adverse affects of non-normal data may be quite limited. For example, a statistical text cited by ORA to support its views on Central Limit Theorem also states,

---

[85] *Id.* at 349-351.

[86] *Bell Atlantic New York Order*, 15 FCC Rcd at 4187, App. B., ¶ 11. We assume that the t-test used by BANY is the Modified Z-test with the resulting Z-statistics compared to critical values in a t-distribution table rather than a normal curve table. *See also* Khazanie, *supra*, at 410-411 (1997), and Brownie, et al., *supra*, at 260-261 (1990).

[87] See the graphs and data tables presented in conjunction with the discussion herein of data transformations. The presented data is actual commercial performance data. It is extremely non-normal even at sample sizes of as large as 179,000 cases.

> Regardless of the shape of the population from which we draw our samples, the sampling distribution of means will be normal *if the sample size is sufficiently large*. What is a "sufficiently large" sample? There is no easy answer, because the required sample size depends on the shape of the population distribution. You will find some statistics texts specifying an *N* of 30 and others an *N* of *50*; certainly an *N* of 100 would remove all doubt about the resultant shape of the sampling distribution. In any event, the central limit theorem enables us to solve problems without worrying whether or not the population from which we are sampling is normal. (p. 151, italics in original text, underlining added.)[88]

We appreciate ORA's persistence in raising this concern, and agree insofar as we acknowledge that non-normality is a problem of an unknown extent. We will not act on this until we receive more evidence on the extent of the problem before prescribing for the final decision model any statistical tests that may be adversely and meaningfully affected by non-normality.

Verizon CA agrees to use the Modified Z-test, although its agreement is conditional. Most importantly, Verizon CA agrees to use the Modified Z-test for average-based measures if a permutation test is used for small samples. As discussed below, we agree with the concept, but have concerns with the implementation.

Permutation tests

To remedy the problem of small samples, which may not meet the "normality" assumptions of the Modified Z-test, Verizon CA proposed that a permutation test be used for average-based and other performance measures. The permutation test is a statistical test that, independent of any underlying

[88] A. Bartz, Basic Statistical Concepts at 150-151 (1988).

distribution, assesses the probability of an outcome. As such it is termed a "distribution free" or non-parametric test in contrast to the parametric Z-test which is based on distribution assumptions.[89] The reasoning behind its use is that when the Z-test normality assumption is violated, a permutation test is more appropriate and accurate since it compares the actual CLEC data directly to the ILEC data without making distribution inferences. Theoretically, the test is only necessary for smaller samples where Central Limit Theorem does not predict normality, because the two tests should produce similar results for larger samples. Differences in distributions do not affect permutation test results, and "look-up" distribution tables, such as "Z" or "t" tables are not necessary.[90] In theory, the benefit of permutation testing is that it can increase the accuracy of the error estimates, thus enabling more accurate decisions.

Only Pacific objects to the use of permutation tests.[91] Pacific originally objected to the assumed costs of such a procedure, but continues to object even though those costs have turned out to be much smaller than originally assumed.[92] Pacific now objects to the procedure as being inadequately

---

[89] *See* generally, P. Good, Permutation tests: A practical guide to resampling methods for testing hypotheses (2nd Ed. 2000).

[90] *See* Mallows Aff., FCC CC Docket No. 98-56, ¶¶ 25-29 at 15-17 (May 29, 1998).

[91] Pacific Reply Brief at 14-15 (May 5, 2000).

[92] Pacific originally estimated the implementation cost of permutation at .75 to 1.2 million dollars (Pacific Bell response to staff questions, February 11, 1999 workshop). Recently Pacific updated their estimate, showing a $300,000 initial implementation cost, with $24,000 to $36,000 yearly maintenance and operational costs (Pacific Bell, deliverable no. 8, April 13, 2000), although we are not aware of any competitive bids that might serve to reduce this estimate further.

tested and too complex,[93] although earlier had acknowledged its feasibility at least for Pacific samples less than 5,000 or 10,000.[94] Regarding the feasibility of its use for such large samples, Verizon CA has presented procedures for implementing permutation on samples of any size.[95]

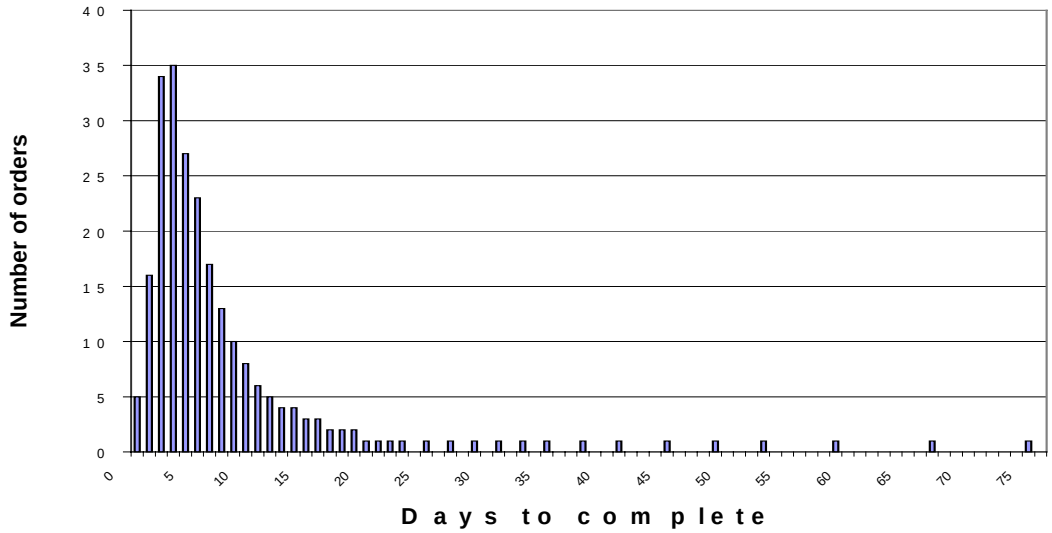
The selection of the appropriate test for small samples should be based on the relative accuracy of the different tests. The permutation test has the potential for being a more accurate test that can handle small samples. Contrarily, the Z-test relies on the resulting sampling distributions being normal. Evidence in this proceeding is compelling that normality cannot be assumed for small samples since measures of time-delay are commonly skewed – the distribution is "bunched up" for shorter delays, and tapers off slowly for longer delays. (*See* Figure 5 for a hypothetical example of a provisioning frequency distribution.)

[93] Pacific Reply Brief at 14-15 (May 5, 2000).

[94] 2000 Pacific Workpaper No. 9 (April 13, 2000).

[95] I.e., resampling techniques. Verizon CA Opening Brief, Attachment 1 at 1 (April 28, 2000). *See also Bell Atlantic New York Order*, 15 FCC Rcd at 4189, App. B, n. 38, and P. Good, *supra* (2000).

Figure 5: A skewed distribution



Given the Z-test's problems with non-normal data, and the fact that the permutation test is unaffected by different distributions, it is possible that the permutation test will be more accurate, and thus would be the preferred test. Theoretically, one should expect that the permutation test would calculate alphas that diverge from Z-test-produced alphas increasingly as sample sizes decrease – the smaller the sample, the larger the discrepancy. On the other hand, as sample sizes increase, the alphas from the two methods should converge toward equality for large samples. Unfortunately, the few data examples we have available to us do not show this expected relationship.[96] The examples show the expected divergence for small samples, but not the expected convergence for larger samples, contrary to the theoretical expectation that the

[96] John D. Jackson, *Using permutation tests to evaluate the significance of CLEC vs. ILEC service quality differentials,* Verizon CA Opening Brief, Attachment 1 at Appendix 2 (April 28, 2000).

results should be the same for large sample sizes.[97] These results raise doubts that the record before us is sufficiently developed to allow us to confidently select the permutation test as a superior test. Either the permutation test is treating data differently than we would expect, or a sample size of 30, or even 131, is still too small to expect sample mean distribution normality for these performance measures. We note that the permutation test is relatively insensitive to outliers[98] compared to the Z-test. This insensitivity occurs because in the final step, the permutation test treats the data as ranked data where an extreme score's value does not influence the outcome.[99] In contrast, extreme scores influence the Modified Z-test. [100]

This result raises the question whether extreme scores would have insufficient influence in a permutation parity test, insofar as these extreme scores might be some of the most publicly noticeable indicators of

---

[97] *See* Jackson, *supra*, at 2-9.

[98] In this application a statistical outlier refers to rare extreme scores, for example, a large but rare performance failure such as an unusually long provisioning time.

[99] R. Khazanie, *supra*, at 720 (1997).

[100] This insensitivity can be illustrated by examining the data example originally presented by Dr. Mallows, but elaborated by Dr. Jackson. (*See* Verizon CA Opening Brief, Attachment 1 at Appendix 2 (April 28, 2000).) In this example, if one were to change the value of the highest CLEC result, 5, to 10, the permutation statistic would not change and remains at an alpha of about 0.15 – a "pass" at a critical alpha level of 0.10. In contrast, the Z-statistic would increase considerably, as the CLEC mean would increase from 4.0 to 6.5. The Z-statistic would increase from 1.2 (0.12 alpha) to 3.0 (0.001 alpha), changing this result from a "pass" to a "failure." Generally, non-parametric tests are considered less powerful insofar as they rely on ranked rather than interval data. R. Khazanie, *supra* at 720 (1997).

discrimination. For example, an unusually long delay in obtaining a needed phone service can be especially troubling. Other issues regarding the selection of the Z-test or the permutation test are more fundamental. If it is more appropriate to view the ILEC and CLEC performance results as samples of a theoretically larger process, then the Z-test may be the more appropriate test. If it were more appropriate to view the ILEC and CLEC performance results as the whole population of production output, then the permutation test would be more appropriate. This underlying issue was raised in the ACR, but has not been resolved by the parties or the record in this proceeding. Until we can determine which test is the more appropriate treatment of the data, including underlying issues such as "production output" versus "larger process population sampling" and more specific issues regarding outlier treatment, we are not in a position to either order or approve use of the permutation test. The most important question of decision accuracy is not resolved. Additionally, we need to better understand what the appropriate sample sizes are for using the permutation test versus the Modified Z-test.

Consequently, we will order the Z-test used during the trial period for all average-based performance results. Most importantly, we will not order Pacific to implement a permutation test data analysis system since even the new lower cost estimates warrant a greater confidence than we currently have in the test's benefits relative to its costs. However, we recognize the permutation test's potential for being the more accurate test, especially if it is appropriate to view a CLEC result as a sample of a fixed production output result. As we believe it would be a mistake to leave unresolved the questions surrounding this test's potential, we direct the parties to conduct or fund a research inquiry to answer these questions. We prefer a collaborative research approach where all

interested parties would collectively influence the research proposal, and thus would be more inclined to accept the results. But in the interim, the Z-test is the most developed and accepted alternative to permutation testing. We shall order that the Modified Z-test be used for average-based parity performance measures. We discuss further the problem of small samples in a following section.

Percentage-based measures

Modified Z-tests

While the parties have proposed Modified Z-test variants for percentage-based measures, and those variants are being used in New York and Texas, these measures present new difficulties for Modified Z-test application. For example, the test requires an ILEC variance. When there is perfect ILEC performance, the Modified Z-test statistic is not calculable.[101] Pacific proposed a modification to the Modified Z-test for percentages based on the CLEC variance. The CLECs and Verizon CA proposed use of permutation tests, or more specifically, exact tests, which do not require calculation of ILEC variance.

Exact tests

Exact tests are called "exact" because if used consistent with necessary assumptions they calculate the exact probabilities of frequency (counted, rate, proportion) data.[102] They represent a special case of permutation testing. The advantage for our statistical model is two-fold: (1) calculations are

---

[101] Pacific Opening Brief at 9-10 (April 28, 2000).

[102] *See* CLECs' Reply Brief at 12 (May 5, 2000) and 2000 GTE/CLEC Workpaper No. 7: D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures at 221-225 (1997) (March 30, 2000).

made directly from the raw data, and (2) exact tests have the potential to produce more accurate results for small samples. In the case of the percentage-based performance results data, the Fisher's Exact test is appropriate.[103]

The Fisher's Exact test calculates the probability of an obtained or worse result when the data conform to a two-row by two-column table. Such is the case in the analysis of percentage-based measures where, for example, the first row represents CLEC percentages with the number of "missed dates" for orders in the first column and the actual number of "met dates" in the second column. The second row similarly represents the ILEC data, creating a two-row by two-column data table, or a "2 x 2" table. Given such a table, there is a limited number of possible unique combinations, or permutations, of entries in each of the table's four "cells." The Fisher Exact test determines the probability of each individual combination that is as extreme or worse than the obtained combination being tested. The sum of these probabilities is the probability that the obtained result could occur if the results are only due to random variation.

This probability is "alpha," the probability of a Type I error. Unlike for average-based permutation applications, outliers cannot affect the result, as the data consist only of "cell counts." Additionally, unlike for average-based permutation applications, the results from the percentage-based Modified Z-test and the results from the Fisher's Exact test converge towards equality as theoretically expected.[104] Additionally, the FCC has approved an application that

---

[103] *Id.* at 221.

[104] We take official notice of sample Fisher's Exact test and Z-test calculations performed collaboratively by staff and Pacific's consultant that show this convergence. The results of these calculations are presented in Appendix D. During the calibration

Footnote continued on next page

uses the Fisher's Exact test for percentage-based measures.[105] We shall order that the Fisher's Exact test be used for all percentage-based parity tests.[106] The evidence before us indicates that it provides accurate decision error probabilities, is consistent with theoretical assumptions, and solves the Z-test application problems.

Rate-based measures

The problem, and our solution, for rate-based performance result analysis is similar to the case of percentage-based performance measures. In this case, a binomial exact test is applied to rate data because the Fisher's Exact test's assumptions are not met. Specifically, the Fisher's Exact test is not appropriate where the row totals are not fixed, or where an entity being observed can contribute more than one cell entry. In the case of percentage-based measures, the Fisher's Exact test is warranted because the row totals are always 100 percent, equal to the total number of CLEC or ILEC orders, and every order only creates one cell entry. In contrast, row totals for rates vary directly with the performance result. For example, the most common rate measure is service "troubles." The rate is typically taken as the rate of troubles per number of lines. This figure can theoretically vary from zero to a number greater than the number of lines

---

phase, parties will be able to confirm these results for the data that is available to them by their own agreements.

[105] *Bell Atlantic New York Order*, 15 FCC Rcd at 4188-4189, App. B., ¶ 13 and n. 39.

[106] Since larger samples cause computer resource problem, an upper sample size limit will be applied. Since Z-test and Fisher's Exact Test have the same results for large samples, and since calculations over approximately 1000 for CLEC "hits" and "misses" can generate computationally difficult numbers, the Z-test will be used for those samples. (See Appendix C.)

because it is possible to have more than one trouble per line. Consequently the row totals are not fixed. However, in this case, assuming the parameters for a Poisson distribution, a binomial exact test can be applied to calculate the probabilities of rate performance results.[107]

Additionally, like the percentage-based Fisher's Exact test applications, and unlike for average-based permutation applications, the results from the rate-based Modified Z-test and the results from the binomial exact test converge towards equality as theoretically expected.[108] Verizon CA, the CLECs, and ORA agree to the appropriateness of the binomial test[109] and Pacific does not object. We shall order that the binomial exact test be used for all rate-based tests as the evidence before us indicates that it provides accurate decision error probabilities, is consistent with theoretical exceptions, solves the Z-test application problems, and is preferred by most parties.

Confidence levels

Alpha levels

The specific fixed alpha levels that have been recommended in this proceeding are 0.15, 0.10, and 0.05 alphas, which correspond to the 85%, 90%,

---

[107] CLECs' Reply Brief at 11(May 5, 2000).

[108] We take official notice of sample binomial and Z-test calculations performed collaboratively by staff and Pacific's consultant that show this convergence. The results of these calculations are presented in Appendix E. During the calibration phase, parties will be able to confirm these results to the extent that their own agreements allow access to the necessary data.

[109] Verizon CA Opening Brief at 24 (April 28, 2000); CLECs' Reply Brief at 11 (May 5, 2000).

and 95% confidence levels, respectively. The 90% confidence level suggested in the ACR is no party's favored level. The ILECs, Pacific and Verizon CA, prefer a 95% level to minimize the possibility of payments made due to sampling error when there are no real differences. The CLECs and ORA prefer an 85% confidence level to minimize the possibility that the ILECs escape payments when there are real differences, but those differences are masked by sampling error.[110] Each side wishes to protect against the negative effect of random variation. But since there are two possible effects of random variation, and as one is minimized the other is maximized, the two sides differ in the preferred confidence level.

Pacific and Verizon CA assert that the 95% level should be used since it is an accepted convention. We disagree. While we understand that it is a convention is some contexts, it is important to understand those contexts to see if they generalize to the present case. They do not. Academic texts that address the use of the 95% level, and that go beyond simply noting its common use as a convention, are clear in pointing out its arbitrariness in applied decision settings:

> *The widespread convention of choosing levels of 0.05 or 0.01 irrespective of the context of the analysis has neither a scientific nor a logical basis.* The choice of level is a question of personal judgment in the Fisherian approach and one of considering type I and II errors in the Neyman-Pearson approach. Since for a given sample size decreasing one error probability

[110] Our conclusion regarding ORA's position here is based on its preference for one standard deviation being the cut-off for a discrimination finding, and its statement describing its position as similar to the CLEC position. ORA Opening Comments on the ACR at 6.) One standard deviation is approximately equivalent to a 15% alpha, or an 85% confidence level. However, as discussed *supra*, ORA's definition of "standard deviation" is unclear.

> increases the other…, it is possible to argue for a relative balance. In particular, if at $\alpha = 0.05$ the power is very low, one might seriously consider increasing α and so increasing the power.[111]
>
> In our opinion, there is no "right" or "wrong" level here – the decision must be made in full consideration of parameters inherent in the problem itself. It is doubtful that setting *a priori* levels of .05, .01, or what have you settles the matter.[112]
>
> No absolute standard can be set up for determining the appropriate level of significance and power that a test should have. The level of significance used in making statistical tests should be gauged in part by the power of practically important alternative hypotheses at varying levels of significance. If experiments were conducted in the best of all possible worlds, the design of the experiment would provide adequate power for any predetermined level of significance that the experimenter were to set. However, experiments are conducted under the conditions that exist within the world in which one lives. What is needed to attain the demands of the well-designed experiment may not be realized. The experimenter must be satisfied with the best design feasible within the restrictions imposed by the working conditions. *The frequent use of the .05 and .01 levels of significance is a matter of a convention having little scientific or logical basis. When the power of tests is likely to be low under these levels of significance, and when type 1 and type 2 errors are of approximately equal importance, the .30 and .20 levels of significance may be more appropriate than the .05 and .01 levels.* (p. 14, emphasis added.)[113]

---

[111] A.H Welsh, Aspects of statistical inference at 128, (emphasis added) (1996).

[112] J. Skipper, A. Guenther & G. Nass, *The sacredness of .05: A note concerning the uses of statistical levels of significance in social science*, 2 The American Sociologist at 17 (1970).

[113] B.J. Winer, *supra* (1971).

> In principle, if it is very costly to make an error of Type II by overlooking a true departure from [the null hypothesis] but not very costly to make a Type I error by rejecting [the null hypothesis] falsely, one could (and perhaps should) make the test more powerful by setting the value of [alpha] at .10, .20, or more. This ordinarily is not done in social or behavioral science research, however. There are at least two reasons why [alpha] seldom is taken to be greater than .05: In the first place. . . in such research the problem of relative losses incurred by making the two kinds of errors is seldom addressed; hence conventions about the size of [alpha] are adopted and [beta] usually is ignored. The other important reason is that given some fixed [alpha], the power of the test can be increased either by increasing sample size or by reducing the standard error of the test statistic in some other way, such as reducing variability through experimental controls. (P. 290.)[114]

These four quotes point out the dilemma in our applied problem. Unlike in scientific applications where the parameters of an experiment are easily manipulated, we have neither the luxury nor the discretion to change the sample size, the effect size, or the sampling error. Consequently, the Commission must chose an alpha level without regard for conventions developed in qualitatively different contexts.[115]

Additionally, while the authors of the last two quotes appear to differ in their recommendations regarding the relative consequences of Type I versus Type II error, these differences should be viewed in terms of different

[114] W. Hays, *supra* (1994).

[115] Faced with a similar problem in D.97-09-045, we based our decision on the actual probabilities, and not on an arbitrary pre-selected significance level. (D.97-09-045, *mimeo.* at 31-32 (September 3, 1997).)

assumptions regarding the freedom to change sample sizes, error terms, and the strength of experimental treatments, among other parameters. Academic treatises directly addressing these relative consequences have developed formulas that balance the net consequences of any resultant error by establishing loss functions. [116]

For example, while different alpha, and thus beta, levels are appropriate depending on the ratio of the costs of the consequences of both types of errors, when the error consequences are deemed to be equal, losses are minimized when alpha and beta are set to be equal.[117] We have not determined a specific ratio for the relative consequences of failing to identify competition barriers when they exist versus monetary payments made when they should not be made. However, at this point we can only assume from the purpose of the Act and the regulatory policy mandating competition,[118] that the consequences of not identifying barriers is at least equal to any misappropriated payments.[119] As

---

[116] C. Das, *Decision making by classical test procedures using an optimal level of significance*, 73 European Journal of Operational Research at 76-84 (1994).

[117] *Id.* at 78.

[118] For example, the FCC has stated that it based its public interest evaluation and approval of BANY's 271 application on the fact that a primary element of the New York remedies plan was its design to "detect" discrimination. FCC BANY Order at ¶ 429. Test power is the closest index of this fundamental purpose.

[119] In comments on the draft decision, the ILECs dispute that failure to detect discrimination has consequences as harmful as mistakenly detecting discrimination. Pacific Comments at 8-9 (December 18, 2000); Verizon CA Comments at 9 (December 18, 2000. We agree that this issue deserves further discussion, but we are also comfortable moving forward with an interim decision based on our assumption. There will be ample opportunity to further consider these issues before any element of the

Footnote continued on next page

a consequence, our goal will be to choose an alpha level that serves to balance with a beta level.[120] In doing so we are not addressing risk. The question of relative risk is more appropriately addressed in the proceeding's next phase, which will establish the "consequences" for the performance decisions made in the present phase. Balancing alpha and beta to be equal only ensures that the most accurate decision is made, not what the consequences of those decisions will be.

We note that the FCC encourages such a balance.[121] We also note that the NYPSC has adopted as low as an 80% confidence level in certain circumstances, possibly to achieve a better balance. While we have discussed a 90% confidence level as a compromise to facilitate negotiation progress, we are unwilling to permanently select such a fixed level based solely on the midpoint between two negotiating positions.

Pacific argues against the 90% confidence level stating, "There is no forum of which we are aware that supports the use of a 10% error rate." However, we find it notable that the BANY remedies plan uses a 21% error rate (79% confidence level) for conditional failure identifications and what in essence

---

model that depends on this assumption is implemented and before a final remedies plan is implemented.

[120] The parties have argued over balancing for "equal risk" versus "equal error." (*E.g.*, Verizon CA Reply Brief at 9 (May 5, 2000) We note that when the ratio of error consequences is set to "1," the Das (1994) "equal risk" formula simplifies to what essentially is an "equal error" formula.

[121] *Bell Atlantic New York Order*, 15 FCC Rcd at 4190-4194, App. B., n. 50.

is a 10% error rate for final disposition of those identifications.[122] We also note that one of the statistical texts frequently cited in the FCC's BANY 271 approval states, "The value of alpha chosen is usually between 0.01 and 0.1, the most common value being 0.05."[123]

Although Verizon CA presents an academic cite as justification for its preference for a 95% level (.05 alpha), we find that that cite refers only to less formal "rough conventions" and does not refer to any context or consequences of the two different types of error.[124] Additionally, Verizon CA quotes an affidavit in a FCC proceeding citing an AT&T statistician's support for the 95% level. We also do not find that quote necessarily applicable to the problem of balancing the two errors. In that quote, Dr. Mallows states that a 95% level would control Type I error "while making the probability of Type II errors small for violations that are of substantial size."

The Commission cannot base its decision on such a statement when the statement context is not clear. At the time Dr. Mallows made the statement, over two years ago, it may not have been apparent how small the sample sizes

---

[122] *Id*, at 4189, App. B, n. 41.

[123] Khazanie, *supra*, at 506 (1997).

[124] "The hypothesis test of H_0 consists of computing [the achieved significance level], and seeing if it is too small according to conventional thresholds. Formally, we choose a small probability α, like .05 or .01, and *reject* H_0 if [the achieved significance level] is less than α.... Less formally, we observe [the achieved significance level] and rate the evidence against H_0 according to the following rough conventions: [achieved significance level < .10 [is] borderline evidence against H_0." B. Efron & R. Tibshirani, An introduction to the bootstrap at 203-204 (1993) (emphasis added).

were going to be, and thus he may have been referring only to results obtained from fairly large samples. We are concerned that even substantial Type II errors may not be identified with a 0.05 alpha level for small-to-moderate samples. Additionally, Dr. Mallow's statement implied that the statistical test, through its significance level, was used to determine magnitude as well as statistical significance. We cannot know how Dr. Mallows' statement applies to our context without knowing what he meant by the term "substantial." Dr. Mallows more recently has stated that he believes 0.15 is the appropriate level and that the 0.05 level seems too small since it "gives more of a chance of failing to detect a violation than of performing a Type 1 error. . ." (RT at 919, lines 14 to 24) But more importantly, our approach is different. We will address the magnitude issue separately below after the error problem has been addressed.

A deciding factor for us is the potential consequences of the two types of error to our overall performance remedies plan. Given the potential for us to err on one side where we might favor either alpha levels or beta levels to the detriment of the other, the correctability of any such imbalance that might result is an important consideration. On one hand, if we set alpha too large and as a result make Type I errors, we can make up for these errors in the incentive-amount methodology phase of this proceeding. For example, we could adjust the incentive amount to the actual Type I error calculated for each performance result. Specifically, presented for illustration purposes only, we could levy an incentive payment for a result with a Type I error probability of 0.05 at 95% of a pre-determined amount, but levy a payment with a Type I error probability of

0.15 at 85% of the same amount.[125] In contrast, once we have made a Type II error, no correction is possible since parity would have been concluded. In this case the measurement would not make it to the incentive payment phase, and thus would not be correctable.[126]

We note that the NYPSC addressed this issue by selecting three alpha levels: a 0.05 alpha level for immediate non-parity identification, approximately a 0.20 alpha level for conditional parity identifications depending on subsequent months' results, and what in essence is a 0.10 alpha level for final disposition of conditional identifications.[127] The parties have variously proposed

---

[125] The actual alpha probability for each result would be used, not any pre-selected alpha level. For example, if the probability of an obtained result being a Type I error was .03, then 97% of the payment would be assessed, if the error was .12, then 88% of the payment would be assessed, and so forth. Across time, this method may mitigate the problem of Type I error payments. For example, in the long run, there may be no difference between "forgiving" 15% of the incentive payments versus charging only 85% of the levied payments. A probability-adjusted scheme would be even more accurate in the long run. *See* H. Raiffa, Decision analysis (1970). We provide this example for illustrative purposes only and do not suggest that these values would be the specific appropriate ones. Our point is that payments can be scaled to error probability estimates similar to that suggested in the ACR. ACR at 26 (November 22, 1999).

[126] In comments on the draft decision, Pacific disputes our conclusion. Pacific Comments at 7-8 (December 18, 2000). We appreciate its comments and welcome its interest in understanding Type II errors. However, we are not persuaded by its argument. Pacific appears to be discussing a different topic – the likelihood of future discrimination being detected. Our point is that nothing can be done about an erroneous decision to conclude parity, because assumed parity causes no action or adjustments regardless of the degree of the error. On the other hand, when it is concluded that discrimination exists, the degree of Type I error probability is apparent, action is taken, and that action can be "calibrated" to the degree of the error.

[127] *Bell Atlantic New York Order*, 15 FCC Rcd at 4189, n. 41.

the 0.05 or the 0.15 alpha levels, and the ACR recommended a 0.10 level for the purposes of development, inquiry and compromise. However, we are not comfortable selecting alpha levels without discussing and assessing beta and its converse, test power.

Test power

Unfortunately, the record is relatively silent on the actual beta values that various critical alpha levels might produce. The only estimates in the record are that in early tests, AT&T estimated betas to range as high as 0.21 when critical alpha levels were set to 0.05.[128] A beta value of 0.21 corresponds to a test power of 0.79, or 79%. AT&T also estimated that if alpha was set to 0.15, then betas would average a similar level - an average test power of 85% when the average Type I confidence level is 85%. Yet it is unclear if the results from the earlier tests are comparable to the performance results in California. To remedy this lack of critical information, we shall direct the ILECs to calculate both alpha and beta values whenever a statistical test is applied.

Staff has performed some preliminary estimates of beta values using four different alpha levels.[129] The results are discouraging about the ability of our

---

[128] Verizon CA Reply Brief at 8, n.2 (May 5, 2000).

[129] We take official notice of tables prepared by staff summarizing the beta levels that are obtained with different tests and different alpha levels. These tables are presented in Appendix F. These values are based on May 2000 performance data and are preliminary estimates based on the application of the Modified Z-test to average, percentage, and rate-based measures. The alternative hypotheses posed for all estimates were that the CLEC's results were at least 50 or 100 percent worse than the ILEC's results. The formula used is found in Hays, *supra* at 284-289 (1994). Staff presents these values as approximations, and does not represent that these calculations are necessarily the best estimate of beta. We present them here to begin a discussion of

Footnote continued on next page

model to perform its most fundamental task, to detect competition barriers. For example, with a 0.10 critical alpha level, and selecting a 50 percent difference to establish alternate hypotheses, beta values average 0.63 with a median of 0.79.[130] While the selection of a 0.10 critical alpha threshold ensures that 100 percent of the performance results are subject to a 10 percent maximum Type I error, it only provides that 16 percent of the results are subject to a 10 percent maximum Type II error.[131]

Additionally, the parties have not recommended any minimum test power, or its respective error, beta. Since beta is determined by the other elements, the degree of test power ends up being that which results from the other elements. The record is relatively silent on the appropriate test power or beta error level. While unfortunate, this state of affairs is understandable since at the outset alpha can be set, but beta can only be determined upon obtaining the measured performance results. Beta will thus vary for every performance result. For every obtained result, however, it is possible to balance alpha and beta if we can safely make assumptions about two components of the analysis: (1) the relative consequences for each type of error, and (2) the specification of the alternative hypothesis.

As a general policy statement, it is reasonable to assume that a Type II error is at least as important as a Type I error, as discussed earlier. Apparent

---

beta estimation, and believe that the values are sufficiently appropriate for us to base the decisions we make regarding the need for further research and development.

[130] App. F at 2.

[131] *Id.*

discrepancies can be adjusted in the incentive payment phase. However, specification of an alternative hypothesis is more difficult. The alternative hypothesis is the hypothesis that barriers exist - that ILEC service to its own customers is actually worse than to CLEC customers beyond that which could be explained by sampling error. We are aware of three ways to specify the alternative hypotheses. First, the critical value for the alternative hypothesis could be set to equal the critical alpha level value. This would not be much help because the beta error level would always be 50%.

Second, the actual result could be selected as the alternative hypothesis. It would be reasonable to assume that an actual result was the best estimate of the actual underlying process, and as such best represents the alternative hypothesis. A statistical test could then estimate the respective Type I and II errors of this result being a "true" mean, not identified due to sampling error. In this case, the balanced alpha and beta level could easily be determined.[132] It is unclear at this point, though, what the effects of this balancing would be since for very small differences, both beta and alpha might be very large, whereas for big differences, both might be small. If this happens, we would still have to set some alpha/beta thresholds, and/or set some "material" difference thresholds.

Third, the critical alternative hypothesis value could be determined by identifying a performance result or level where ILEC and CLEC service differences become "meaningful." Verizon CA has proposed a few performance

[132] C. Das, *supra*, at 78 (1994).

levels, called "deltas," as a solution to a different problem in this proceeding.[133] However, the record contains no information on what most deltas would be, as no party has submitted any proposal containing a comprehensive set of specific deltas.

In comments on the draft decision, the ILECs assert that establishing alternate hypotheses that represent competition barrier thresholds is a significant problem that may make the exercise fail. Pacific Comments at 9-12 (December 18, 2000); Verizon CA Reply at 4 (December 22, 2000). However, we note that the ILECs are willing to establish nearly identical thresholds to use for "materiality" standards to reduce payments. Pacific Comments at 20 (December 18, 2000); Verizon CA Comments at 11- 12 (December 18, 2000). Adapting such standards for alternate hypotheses, if any adaptation is necessary at all, should be relatively easy.

A fixed alpha is not an adequate long-term solution. As the CLECs have asserted and as staff's data analysis has shown, test power is very low for the small samples that represent the majority of the performance measure results. On the other hand, the ILECs have asserted, and staff's data analysis confirms, that fixed alphas that provide better test power for small samples result in unnecessarily high test power for large samples. This unnecessarily high test-power can easily result in meaningless differences being found statistically significant.[134] We believe that the problems of insufficient test power for small

---

[133] Verizon CA Opening Brief at 10-11 (April 28, 2000).

[134] Verizon CA Opening Brief at 10, n. 6, citing P. Bickel & K. Doksum, Mathematical statistics: Basic ideas and selected topics at 175 (1977)(April 28, 2000).

samples (large beta) and "too much" test power for large samples can be better resolved through even approximate alpha/beta balancing techniques. We direct the parties to develop and implement an alpha/beta balancing procedure for our model. However, to give sufficient time for its development without delaying Pacific's 271 application, we shall adopt a fixed alpha solely for the interim, and shall order that the balancing components to the model be added by the end of the trial period unless the parties reach agreement and move to implement the components sooner.

Fixed alpha

We conclude for the reasons cited above that a fixed alpha critical value should only be used as an interim decision-criterion solution. Setting alpha to remedy one problem only makes another. We select a larger alpha level, 0.10, instead of the 0.05 level to enhance decision accuracy and to avoid uncorrectable decision-making errors while still being able to address correctable errors in the next phase of this proceeding. We select a smaller alpha level than 0.15 because we are concerned about the effect on large-sample results. We have selected the 90% confidence level (0.10 alpha, or 10% significance level) to control the Type I error and to reduce the Type II error to more acceptable levels for the preponderance of the performance results. That is, we choose to be at least 90% confident that any barriers we identify represent real differences, not differences due to sampling error (random variation), while increasing the average

confidence level (power) for detection of actual differences from 30% for the 0.05 alpha to 37% for the 0.10 alpha.[135]

Additionally, because of the low power of these tests, pending further development and consideration we intend to also adopt the 80% confidence level (0.20 alpha) for conditional failure identifications. This threshold is used in the BANY performance remedies plan for conditional identifications where results at 0.20 alpha or less were deemed failures if they occurred in two months of a three-month period.[136] We will not dictate the additional specifications for such conditional identifications, but instead direct parties to set forth those specifications in the next phase. Among other possibilities, our plan could have additional criteria such as (1) successive failures such as in the BANY plan, (2) alpha and beta balance at values less than 0.20, or (3) for CLEC-specific performance assessment, industry aggregate performance out of parity. Noting that if a 80% confidence level (0.20 alpha) was used as the overall fixed threshold instead of the 90% level (0.10 alpha), average

[135] These figures are based on an alternate hypothesis of 50% worse performance for the CLEC and on CLEC samples of only 5 or more. Average power increases from 37% to 49% assuming a 100% worse-performance alternate hypothesis. These estimates were made from existing data and could easily change in the future without any changes in the plan. For example, if the CLECs gain larger shares of the local phone markets and CLEC companies individually place more orders, sample sizes will increase, with a resulting increase in test power, with all other elements held constant.

[136] *Bell Atlantic New York Order*, 15 FCC Rcd at 4189, App. B, n. 41

power would increase from 37% to 48%,[137] we wish to take advantage of this increased power at least on a conditional basis.[138]

Material differences

None of the parties have specified a comprehensive set of minimum differences (effect size) between the ILEC and CLEC performance results that would identify a competition barrier. Two parties have raised the issue. AT&T has somewhat tangentially raised the issue in its discussion of test power[139]. To calculate test power, an alternative hypothesis must be specified as discussed *supra*. AT&T estimated test power across an array of different performance results after subject matter experts made judgments creating competition-affecting performance thresholds.[140] Verizon CA currently proposes utilizing "deltas" which embody virtually the same concept, albeit for different

---

[137] These figures also are based on an alternate hypothesis of 50% worse performance for the CLEC and on CLEC samples of only 5 or more. Average power increases from 48% to 60% assuming a 100% worse performance alternate hypothesis. See Appendix F.

[138] Some commenters raise concerns that the 0.20 alpha level was not addressed in the record. Verizon CA Comments at 12-13 (December 18, 2000). Pacific Comments at 13 (December 18, 2000). However, we note that there was considerable discussion of what the appropriate alpha level would be and at least one party speculated without the benefit of current data that the level would be 0.15. The record has sufficient discussion of appropriate alpha levels for us to order further development on optimal levels, such as the 0.20 level. We advise parties that we cannot guarantee any conditional 0.20 level will be adopted in the final model, especially if we find that no party has specified reasonable conditions for implementing this alpha level. Nevertheless, we note that a closer read of the New York Public Service Commission's use of this alpha level is likely to be informative.

[139] Cited in Verizon CA Reply Brief at 8, n. 2 (May 5, 2000).

[140] Id.

purposes.[141] Whereas AT&T created thresholds to investigate insufficient test power, Verizon CA proposes to create these conceptually identical thresholds to investigate "too much" test power.[142] We find that both efforts to establish "material" thresholds have merit. First, as we have described above, test power is a primary decision-accuracy concern for this remedies plan. The best way to calculate test power is to specify a meaningful alternative hypothesis, and the most meaningful alternative hypothesis is one that embodies the core performance remedies plan goal, barriers to competition. Second, it would be contrary to the same decision accuracy policy goals to impose incentive payments when an ILEC is providing virtually the same service to a CLEC that it is providing to itself with no negative impact on competition. Recent academic discussions have pointed out that in the case of large samples, statistical results right at an alpha level of 0.05, for example, can provide evidence *for* the null hypothesis, rather than against it as designed:

> Results indicate that for point null hypotheses, a statement of [statistical significance at alpha] does not have a straightforward, evidential interpretation. It is demonstrated, that for larger samples particularly, that a report merely that data are [statistically significant at alpha] has no objective

---

[141] Verizon has proposed specific "deltas" for nine of the approximately thirty performance measures that will be included in the performance incentives plan, although Verizon calls these proposals "preliminary values" that "can and should be adjusted as more data is gathered." Verizon CA Comments on Workshop at 9-11 and Attachments 2 and 3 (April 28, 2000). We appreciate this explicit proposal and encourage further development. Notably, no party has proposed material differences for a majority of performance measures or any average-based parity measure.

[142] Verizon CA Opening Comments on ACR at 11 and App. B. at B2 - B3 (January 7, 2000); Verizon CA Reply Brief at 8, n. 2 (May 5, 2000).

> meaning, and under some conditions should be interpreted not as evidence against the null hypothesis, as is usually supposed, but as strong evidence in its favor.[143]

For very large samples, significant differences at or close to the .05 threshold might be so negligible as to be perceptually the same to a CLEC customer as would be the "statistically significantly different" ILEC service, and as a consequence actually be evidence of parity, not discrimination. Statisticians seem to agree that statistical significance is different from substantial significance.[144]

We find that the "material difference" standard has merit and the potential to improve the decision model we specify. However, we are concerned that the task to construct a set of difference thresholds is difficult, and yet to be accomplished in any collaborative forum. We encourage the parties to complete this task as part of the alpha/beta balancing task we order today. However, since other ways to specify an alternative hypothesis may be easier to accomplish, yet sufficient to enhance decision accuracy, we will not order the material differences be defined for every measure. Other methods for balancing alpha and beta errors may resolve the material difference versus statistical difference problem and we choose to allow the parties the discretion to collaboratively determine the best solution before we order our own solution.

---

[143] D. Johnstone & D. Lindley, *Bayesian inference given data "significant at α": Tests of point hypothesis,* 38 Theory & Decision at 51 (1995).

[144] For example, *see* D. Gold, *Statistical tests and substantive significance,* 4 The American Sociologist at 42 – 46 (1969).

Optimal alpha and beta levels

The parties have variously discussed "equal risk," "equal error," and "balancing alpha and beta." "Equal risk" refers to a situation where the expected consequences of the performance remedies plan are the same for an ILEC as for the CLECs. The concept of equal risk is beyond the scope of our decision model as it necessarily requires incentive payment specification which we will not consider until the next phase of this proceeding. "Equal error" and "balancing alpha and beta" refer to a situation where the two possible decision-making error probabilities are the same. We endorse the concept not only because it meets our fairness principle, but also because it maximizes decision accuracy.

Overall decision error is minimized when alpha and beta are balanced.[145] But most importantly, if we are to create a "level playing field," we must be fair in our acceptance of decision error. The data shows that a fixed alpha level of 0.10 can only be suitable for an interim implementation because it favors reducing the error that only the ILECs wish to reduce. There would be no level playing field if we tolerated no more than 10 percent error harmful to the ILECs, yet tolerated 40 to 60 percent error harmful to the CLECs. We only take the 10 percent alpha level as an interim compromise necessary for progress. Additionally, maximizing decision accuracy by equating possible errors is an appropriate first step to optimizing equal risk, and does not necessarily interfere with the consequence-setting function of the next phase of this proceeding.

[145] C. Das, *supra* (1994).

In comments to the draft decision the ILECs raise several arguments against attempting to balance alpha and beta error. Pacific Comments at 9-12 (December 18, 2000); Verizon Comments at 9-11 (December 18, 2000). We take these comments seriously and note that there will be time for further discussion and consideration of these issues before, and if, we decide to implement a executable balancing feature. However, we note that neither ILEC acknowledges the balancing plan's potential benefit of lowering the average Type I error and of reducing small magnitude failure identifications.[146] We are optimistic that parties will find the net result of an error balancing plan mutually beneficial once the details are resolved.

We direct the parties to work collaboratively to develop and implement an alpha/beta balancing decision component for our decision model by the end of the trial period. If the parties are unable to agree on such a model component at that time, we shall direct parties to submit their individual models for our review and decision.

[146] We note that balancing alpha and beta levels can be a "win-win" situation for the parties when compared to a fixed alpha level. Examining CLEC samples, staff has noted that whereas a fixed alpha of 0.10 results in a maximum error rate of 0.10 for all analyses, if alpha and beta are balanced and the maximum error rate is raised to 0.25 for all analyses, the resulting *average* maximum error rate is 0.072, well below the 0.10 fixed alpha level. A summary of staff's analysis is provided in Appendix G, which also shows the error balancing effect to reduce detection of small differences and increase the detection of large differences.

Minimum sample size

Minimum sample size requirements vary depending upon the type of statistical test used. For example, as discussed above, exact tests are not dependent on inferences about the underlying distribution, therefore the accuracy of calculated alphas is relatively unaffected by sample size. Therefore we find it necessary to discuss sample size issues individually for each type of measure.

Average-based measures

Sample size requirements for average-based measures are the most difficult to resolve. On one hand, the CLECs have pointed out the importance of separately assessing performance for even the smallest CLEC with the least activity since these CLECs depend more on each order or service than do the larger CLECs. Harmful ILEC performance in small new or innovative market niches, or harmful ILEC performance to smaller CLECs could be masked by larger market samples or larger CLEC samples when the results for CLECs are combined ("aggregated"). If so, then the smaller markets and the smaller CLECs would not be provided the protection that this performance remedies plan is supposed to provide. Such small CLECs and markets effectively would be unprotected by competitive market reforms, and thus might fail.

Consequently, the CLECs have urged sample sizes small enough to protect these markets. We agree with this principle, and thus, one goal of our plan is to assess each CLEC's performance results for each submeasure. On the other hand, as sample sizes become small, Central Limit Theorem states that the normality desired for Z-tests can no longer be assumed. The accuracy of the error estimates, alpha and beta, becomes suspect with the smaller samples. So we are faced with the potential dilemma of having to choose between achieving

greater decision accuracy or protecting an important sector of the market. The parties predictably were not able to agree on a solution to this dilemma. Proposals ranged from a sample size minimum of 1 to a minimum of 50 or more.

The issue is relatively simple for the ILECs. They are concerned that small samples could produce inaccurate error estimation, which could inappropriately subject them to payments even when their processes are non-discriminatory. However, since the ILECs are more concerned with alpha levels, unlike beta levels, alpha levels can be held constant regardless of the size of the sample. So even though there may be an issue of accurate alpha estimation, there is still some adjustment as sample sizes decrease – alpha error is held constant. Additionally, with alpha error held constant and as sample size decreases, test power decreases, thus reducing the ILEC's potential liability under any performance remedy payment plan. On the other hand, the ILECs may be concerned that smaller samples generate greater incentive payment exposure by the consequent that there are more performance tests. However, this concern is best addressed in the incentive payment phase where it can be accommodated if warranted. The ILECs also prefer aggregation of all results, since in their view, the total result is the best indicator of the parity of the process.[147] As a compromise, the ILECs offered to use sample sizes from 5 to 20, and they have offered to aggregate results in order to achieve these minimum numbers. With a few exceptions, the ILECs wish to exclude, from the performance remedies plan, data that does not meet these sample minimums.[148]

---

[147] Pacific Reply Brief at 11 (May 5, 2000).

[148] The ILECs and CLECs have agreed to have no minimum sample size requirements for "rare submeasures," which are submeasures that rarely see activity, yet are so

Footnote continued on next page

For example, samples that contain four or less observations after aggregation rules have been applied would be discarded unless they are a designated "rare submeasure" that should be analyzed regardless of sample size.

The issues for the CLECs are more complicated. On one hand, since increasing the sample size increases test power as the significance level is held constant, the CLECs would seem to prefer larger samples. Smaller samples often have negligible test power. However, on the other hand, the CLECs prefer no aggregation of results since the actual service each company receives is critical to them. Each company is directly affected by the service it receives from the ILEC independently of the service that other CLECs receive. Consequently, the CLECs have urged inclusion of sample sizes small enough to protect these markets. Second, the CLECs urge that all data be analyzed regardless of sample size. They do not want any data discarded from the performance remedies plan. It is unacceptable to the CLECs to ignore poor performance to a small emerging CLEC, simply because of a minimum sample size rule. However, like the ILECs, the CLECs agreed to a compromise position, accepting some aggregation rules, but firmly rejecting exclusion of any performance results because of insufficient sample size.[149]

Assisted by Pacific's technical expert, staff examined how one possible compromise set of aggregation rules would function.[150] In summary,

---

important as to need close monitoring when any activity occurs. These submeasures are listed in Appendix H, Attachment 1.

[149] CLEC's Reply Brief at 8-9 (May 5, 2000).

[150] We take official notice of staff's summary of this analysis, included here as Appendix H.

the rules were as follows: (1) Samples of 10 or more would be separately analyzed; (2) All samples of less than 10 would be aggregated for a collective analysis if they achieved at least a sample size of 5; (3) Where a minimum of 5 was not achieved, the remaining samples would be aggregated for analysis with all other CLECs for the submeasure; and (4) Where the industry aggregate did not achieve a minimum of 5 the data would be discarded.[151] Using these rules, for the most recent month presented, March 2000, 57 percent of the performance results could be analyzed without aggregation, 39 percent could be aggregated with other small sample results, 1.3 percent had to be aggregated with the rest of the industry, and 2.4 percent of the results had to be discarded.[152] While not having an opportunity to comment on this, the CLECs can be anticipated to object to these rules insofar as they require that 43 percent of the results be aggregated or discarded and that 3.7 percent (127) be either aggregated with the whole industry, possibly rendering their results masked by a much larger sample, or be discarded.[153]

Staff found several unresolved problems with the proposed compromise aggregation rules. First, in some cases, even with very low test power for a reasonable alternative hypothesis,[154] the performance results to a small CLEC were highly statistically significant with an extremely low Type I

---

[151] Pacific's Reply Brief at 11(May 5, 2000).

[152] *Id.* at 12.

[153] CLEC Reply Brief at 2 (May 5, 2000).

[154] The alternative hypothesis was that performance for CLEC customers is at least 50% worse than for ILEC customers.

error, or alpha. However, the aggregation rules caused this result to be combined with and masked by results for large CLECs. Second, in other cases, where several small CLECs experienced better or nearly equal ILEC performance, exceptionally poor performance to one CLEC caused the aggregate performance to be identified as a failure. Such an outcome could trigger payments to each of the CLECs, thus spuriously expanding the ILEC's liability.

Third, the aggregation rules caused some unnecessary aggregation. For some submeasures where only one CLEC did not have the minimum of five or ten results, its results were aggregated across the entire CLEC industry, which often had more than a thousand individual performance results. This would occur even though aggregating with only the smallest CLEC result over five or ten would have provided a sufficient sample size. With the proposed rules the small CLEC result was unnecessarily completely masked by the very large CLEC samples.

Fourth, in cases where there are multiple results for the same CLECs it is not clear which result would be used. For example, when a small CLEC's results are aggregated with larger CLECs' sample sizes that are small, but which are big enough to be analyzed on their own, two different conclusions could be reached. When the larger individual sample results all pass and when the combination of these results do not pass, the individual larger samples will be deemed to have passed individually but not in the aggregate. This result poses a dilemma in that on one hand the aggregate may be the better indicator of the larger process if one assumes a "process model," but on the other hand, assuming a "service model," only the smallest CLEC suffered harm. Each assumption suggests a different remedy.

We believe that it is important to examine performance at the smaller market and smaller CLEC levels. This market arena may be critical for entry and innovation, which in turn are critical to a healthy competitive telecommunications infrastructure. However, given the unresolved issues for sample size and aggregation rules, and the fact that the rules for incentive payments are integrated with the aggregation rules, we are reluctant to permanently order any minimum sample sizes because any such minimums would require some data be discarded. Before finishing this discussion, we examine proposals that might not require sample size minimums.

Permutation testing has been proposed as a solution to the Z-test's small sample normality assumption violations. We prefer use of the permutation test rather than the complicated, and somewhat confusing, data elimination and aggregation rules. However, as we discussed earlier, the record is not sufficiently complete for us to be confident that permutation testing is free of other problems. In New York, while permutation testing is being developed, the New York Public Service Commission has ordered *t*-tests used for small samples as an interim solution for the Z-test small sample problem.[155]

Statistical texts indicate that the *t*-distribution is more appropriate for tests between two sample means, especially for small samples.[156] Use of a *t*-

[155] *Bell Atlantic New York Order*, 15 FCC Rcd at 4187, App. B., ¶ 11.

[156] For small samples the distribution of the means of samples is different from the distribution of the raw scores themselves as expressed in Z-tables. Roughly speaking, the mean sample distribution is narrower and taller in these circumstances than the raw score distribution. Consequently, a *t*-distribution should be used for statistical comparisons of means from smaller samples.

distribution "look-up" table could alleviate some ILEC concerns regarding possible alpha estimation inaccuracy for small samples. For example, with the current fixed critical-Z decision rules, a Modified Z-test statistic of 1.8 would identify a failure at all parties' favored alpha levels since it exceeds the most conservative proposed critical value of 1.645. This result would be the same for all sample sizes including a sample size of one. However, the ILEC's concerns regarding alpha accuracy increase as the sample size decreases. Using the *t*-distribution table would adjust for decreasing sample size. For example, for an ILEC sample size of two ($df = 1$), a critical value of 3.078 must be exceeded for the 0.10 alpha level.

Our example of a Z-statistic of 1.8 would not be significant unless the result sample size was at least four, since the critical *t* for a sample of 3 ($df = 2$) is 1.886 and the critical *t* for a sample of 4 ($df = 3$) is 1.638.[157] Consistent with the academic justification of the Modified Z-test, we shall order the test statistic compared to the *t*-distribution. In this regard, we will refer to the Modified Z-test hereinafter as the Modified *t*-test, also consistent with its academic reference.[158]

Unfortunately however, this adjustment affects only the relatively infrequent small ILEC samples and not the preponderance of small CLEC samples.[159] Additionally, other questions still remain regarding the accuracy of

---

[157] This illustration uses the ILEC sample size for "looking up" the critical *t*-statistic distribution value. The Brownie, et al., *supra*, research indicates the ILEC sample size should be used for the "lookup" step.

[158] Brownie, et al., *supra* (1990).

[159] *Id.*

alpha estimation even with more conservative t-distribution tables. Even though the *t*-distribution is a remedy for small samples, its appropriate use still assumes the population is normally distributed, especially for one-tailed tests.[160]

We find that the controversies over the appropriate minimum sample size involve several unresolved elements of our decision model: alpha estimation accuracy, permutation or Modified Z-test use, aggregation rules, data exclusion rules, and incentive payment rules. For the reasons that there are several possible solutions to the minimum sample size problem, the resolution of any one of these problems may resolve the others, and the ultimate solution may necessarily involve decisions about incentive payment rules, we are reluctant to order a permanent minimum sample size. We are concerned that without further information, research, and calibration information, we would be essentially deciding "in the dark." While we prefer not to delay specifying final model components, in this case the complexity of the problem and the potential for a better solution warrants the delay. A better solution may be achieved during the calibration phase when parties can see how various rules, tests, and distributions work.[161]

However, we also are concerned that the parties may not either create or agree on a better solution to the small sample size problem. If this turns

---

[160] Hays, *supra* (1994) at 327-328.

[161] Even in the unlikely event that parties are unable to resolve the small sample problem in the incentive phase, Pacific will still be able to present a completed performance remedies plan to the FCC, either as the "no minimum" default we order today, or a different sample size plan that we may subsequently order for a completed remedies plan.

out to be the case, then we would in effect be ordering many applications of statistical analyses and decision rules for samples as small as one or two individual performance results. We find that we need to set some minimal rules that, in the case that parties are unable to agree on better solutions, will reduce dependence on such very small samples. We shall order the following rules as an interim solution as a "floor" for sample sizes. These rules are designed to avoid discarding any data, and to increase sample sizes for the very smallest samples with minimal impact on the actual results. These rules are also designed to be easily understood with the results easily reproduced. We find that the previously proposed rules are complicated and fall short of our goal of simplicity.

The following rules shall be used for average-based parity performance measures:[162]

(1) For each submeasure, all samples with one to four cases will be aggregated with each other.

(2) Statistical analyses and decision rules will be applied to determine performance subject to the performance remedies plan for all samples after the aggregation in step (1), regardless of sample size. For example, if samples with as few as one case remain after the aggregation, statistical analysis and decision rules will be applied to determine performance subject to the performance remedies plan to these samples, just as they are for larger samples.

These small sample aggregation rules minimize most of the problems described above for Pacific's proposed plan. (See Appendix I.) We do not presuppose how payments will be triggered or allocated under these

[162] The results of these aggregation rules are illustrated in Appendix I.

aggregation rules. The issues will be addressed in the upcoming incentives phase. For example, the parties can decide whether any CLEC whose results are aggregated into a failing aggregate, yet whose individual results are better than the ILEC parity standard, should receive incentive payments. In this case, an "underlying process" model might suggest that this CLEC receive payment because the process was flawed and the incentive was necessary to motivate process improvement. On the other hand, a "service" model might suggest that this CLEC not receive payment since it suffered no competitive harm. In comments to the draft decision, the ILECs seem particularly concerned about assessing small samples for potential remedy payments. Pacific Comments at 16 (December 18, 2000); Verizon CA Comments at 14 (December 18, 2000). We remain receptive to the proposal that any penalty amounts could be scaled to the transaction volume and to other proposals which would ensure appropriate treatment of small sample results. *See* Verizon CA Comments at 14 (December 18, 2000). Parties will have an opportunity to propose and discuss different treatments of the outcomes from different sample sizes.

Percentage and rate-based measures

The fundamental problem with small sample sizes for average-based parity measures is that they fail to satisfy the normality assumptions for the Modified Z-test. In contrast, percentage and rate-based measures are assessed using exact tests, which do not depend on inferences or assumptions about underlying distributions. Consequently, with these tests there is less concern with the accuracy of the alpha and beta calculations for small samples. We find no other compelling reason to aggregate or discard data, and thus, we

direct that all percentage and rate-based data at the submeasure level for each CLEC be analyzed for parity regardless of sample size.[163]

Data transformations

Pacific proposes a Modified Z-test enhancement to address the data non-normality problem for average-based measures. Pacific asserts that for lognormal data distributions, transforming raw scores to their natural logs can bring the distribution close to normality, and thus satisfy the essential assumption for using a Z-test.[164] The CLECs agree to such transformations.[165] Verizon CA and ORA accept the transformation proposal in concept, but both are reluctant to use it without further research. We agree with Verizon CA and ORA so far as the record is not clear how such transformations might affect decision accuracy. However, academic sources provide guidance. For example, one text states,

> "The logarithmic transformation is particularly effective in normalizing distributions which have positive skewness. Such distributions occur… when the criterion is in terms of a time scale, i.e., number of seconds required to complete a task."[166]

---

[163] I.e., no minimums are necessary. However, per our earlier discussion, maximum sample size limits are necessary for the Fisher's Exact Test because of computational limitations.

[164] Pacific Opening Brief at 8 (April 28, 2000).

[165] CLECs' Reply Brief at 11 (May 5, 2000).

[166] Winer, *supra* at 400 (1971).

This is precisely the type of measure on which the average-based parity performance measurement is based.[167] So from a theoretical perspective, the log transformation is appropriate and reasonable. Additionally, staff has performed analyses on several qualitatively different performance results. From these analyses, staff has concluded that a log transformation (1) brings the distributions much closer to normality, and (2) provides a reasonable interpretation of skewed data. Staff's analyses of several ILEC and CLEC distributions are included as Appendix J. These analyses show the improvement when log transformations are used. In addition, they demonstrate that even in cases where the log transformation dramatically changes results from the non-transformed data, the transformed results are reasonable and appropriate treatments of the performance data.

Transformations also change the effect of outliers. For example, when an outlier exerts influence on the average result in small samples, transformations can change even the direction of the performance result from worse performance to better performance.[168] In another case, we note that the

---

[167] *See* D.099-08-020, performance measure nos. 1, 7, 14, 21, 28, 37, and 44, and staff's analysis of performance measure results frequency distributions in Appendix J.

[168] We take official notice of a lognormal transformation performed by staff on the example simulated dataset in this record. (Verizon CA Opening Brief at 2-9 and 2-13 to 2-17 (April 28, 2000).) The transformation is included in Appendix J, Attachment 6. The data represent performance measures where higher scores indicate worse performance For the raw data, the CLEC mean was worse than the ILEC mean, 9.94 and 8.29 respectively. The reverse was true for the transformed data. The CLEC mean was better than the ILEC mean, 1.81 and 2.03 respectively. The Modified Z-test score changed from the raw data Z of 1.39, to the transformed data Z of -1.89. The raw data alpha result was 0.083, whereas the transformed data alpha result was 0.97.

probabilities even for large samples where there should not be large differences change dramatically when scores are transformed.[169] While the data sets we reference may be unique examples, they raise questions that we should resolve, but are not in a position to entirely do so from the record in this proceeding to date. For the above reasons, we decline to order transformations of the data on a permanent basis unless the record is adequately developed in subsequent phases of this proceeding. Additionally, our preference is that more exact tests be used, if appropriate, which solve the small sample normality problems without transformations.

However, since we must still use the Modified Z-test, and since we must apply it to samples where normality can not be assumed, then we find that the log transformation is reasonable and appropriate, and is at least as an interim solution is necessary for application of the test to small to moderately large samples. We also find that the transformation improves normality for large samples.[170] Therefore, we shall order that log transformations be utilized for all average-based performance measures as specified in staff's analysis in Appendix J.

---

[169] We take official notice of a submeasure analysis for actual February 2000, OSS performance. With a CLEC sample size of approximately 500 and an ILEC sample size of 6,340, a Modified Z-test on raw scores produces an alpha of 0.85, whereas a Modified Z-test on transformed scores produces an alpha less than 0.0001. The difference is interpreted as follows: Raw score analysis indicates about seven to one odds that the result is due to random variation, whereas the transformed score analysis indicates there is virtually no chance that the result is due to random variation.

[170] App. J, Attachment 2.

This still leaves us with the issue of the meaning of outliers. If the impact of outliers should be minimized in our performance assessment, then the log transformations accomplish this and nothing further needs our attention. However, if outliers are meaningful in their own right, then we need to address the issue. As stated above, it is plausible that an outlier can have a disproportionate affect on competition when in the CLEC sample. Very long provisioning times could gain notoriety that could harm the reputation of a CLEC. On the other hand, outliers in the ILEC results could raise the mean and mask the fact that the ILEC is providing predominately superior service to its own customers. We believe this issue should be discussed in the incentives phase of this proceeding, and we will be open to proposals for a separate treatment of outliers in their own right. But even if parties do not propose a separate treatment of outliers or agree on their meaning, we are convinced that the log transformations provide a more appropriate Modified Z-test application. If further deliberations and negotiations of the parties do not result in adequate development of permutation testing or outlier treatment, at this point we accept the fact that log transformations may become the permanent solution.

Benchmark issues

In contrast to the parity standard for CLEC performance results with ILEC retail analogues, where there is no retail analogue, the standard is performance that allows a "meaningful opportunity to compete." In the performance measurement phase of this proceeding, the parties agreed to establish "benchmarks" which specify such performance levels.[171] Since there is no

[171] *See* D.99-08-020, *mimeo.* at 5-6 (August 5, 1999).

measure of an ILEC's internal performance (i.e., no retail analogue), there is no ILEC variability on which to base an expected performance parity standard. Consequently, parties negotiated measures with thresholds that would allow CLEC service access judged to allow a meaningful opportunity to compete.

The parties discussed two contentious issues regarding benchmarks. The parties discussed alleged problems of small sample sizes causing falsely missed benchmarks and random variation causing falsely missed benchmarks. Pacific proposed using adjustment tables to remedy the sample size issue and statistical testing to remedy the random variation issue.

Benchmark adjustment tables

Pacific contends that performance measures for small samples present problems in that some benchmarks would not be met even though an ILEC provided adequate service. For example, if a benchmark established that 90 percent of orders for a particular service must be complete within a certain timeframe, then for every 100 orders there could be 10 missed timeframes without failing the benchmark. Pacific points out that for small samples, one failure could drop performance below the 90 percent level. For example, if only five orders were made per CLEC, then across 20 CLECs (100 orders) there could be 10 missed timeframes (90 percent on time) and for this aggregate performance a "meaningful opportunity to compete" could be assumed by original agreement of the parties. However, at least two and at most ten CLECs in this example would have missed the benchmark. That is, if ten CLECs missed one timeframe each (for a total of 10 missed timeframes), then they each would have performance measure results of 80 percent. At least two CLECs would have to fail the performance measures (5 failures each for the total of 10 missed timeframes) even though performance was right at the benchmark.

Recognizing this problem, the CLECs have agreed to allow adjustments to the benchmark outcomes, although not to the extent desired by Pacific. Noting that benchmarks were created under the federal definition of performance allowing a "meaningful opportunity to compete,"[172] we are reluctant to allow less than the levels set by the benchmarks. To do so suggests less than a "meaningful opportunity to compete." However, in this case, because of the legitimacy of the small sample problem, and since the CLECs have agreed to some adjustments, we shall include an adjustment table in our decision model. Although the ILECs and the CLECs agree to use a benchmark adjustment table, they disagree on two aspects of such tables, sample sizes to which they will be applied and sample sizes from which they will be derived.

For the application of the adjustment tables to benchmarks results, the CLECs agree to the use of adjustment tables up to a performance result sample size of 30, and propose they be used down to a sample size of 1.[173] The ILECs propose using the tables for performance result sample sizes up to 100, down to 10 with no aggregation, and down to five with the aggregation rules they proposed for parity measures as discussed above.[174] The difference between the two proposals appears to be the type of problem they address. The CLEC table proposal appears to be addressing more closely the data "granularity"

---

[172] *Id.*

[173] CLECs' Reply Brief at 14-16 (May 5, 2000).

[174] Pacific's Reply Brief at 4-7 (May 5, 2000); Pacific's Opening Brief at 12 (April 28, 2000); Verizon CA's Reply Brief at 11 (May 5, 2000).

problem[175] as we have described above, whereas the Pacific table proposal appears to go beyond data granularity and address broader statistical applications to benchmarks as we discuss below.

The ILECs and the CLECs also differ on the second issue, the adjustment table derivation sample size. The CLECs argue that since the table will be used on small samples, the tables should not be derived from larger samples. While they wish to limit the table's application to samples of 30, as a compromise they offer to base the table's derivation on a sample size of 100. Pacific wishes to derive the table from a sample size of 1000, but offers a derivation sample size of 400 as an alternative. Pacific states that a derivation sample size of 400 or 1000 is appropriate because the "implied performance" resulting from these derivation sample sizes is closer to the benchmark and is not unreasonably larger as would be the case with the CLEC's proposed derivation sample sizes.

While the CLECs' position is intuitively attractive in terms of the *construction* of the table, we appreciate Pacific's analysis because it assesses at least one *net effect* of the table. However, just as we are concerned with inferential statistical testing issues, we are concerned that other essential net effects have not been considered, namely the net effect that adjustment tables have in lowering the effective benchmark levels. For example, Pacific's adjustment table would allow performance to drop well below the nominal benchmarks without any failures being identified. Where the adjustment tables

[175] CLECs' Reply Brief at 14 (May 5, 2000).

are applied, performance could average as low as 82 percent or lower across all performance results.[176]

Additionally, we are concerned that "one size fits all" application and derivation sample size specifications may not be appropriate. For example, we note that the smallest application sample size where a whole integer failure matches the nominal 90 percent benchmark limit is 10, yet the similar smallest sample size for the nominal 99 percent benchmark is 100.[177] We find it appropriate to set different application sample sizes for different benchmark percentage levels. In the same manner, we find that a fixed derivation sample size results in varying levels of implied performance relative to the benchmark limit. For example, a derivation sample size of 400 for the nominal 90 percent benchmark results in a 92.9 percent implied performance level, which is a 29 percent movement toward perfect performance.[178] In contrast, the same derivation sample size of 400 applied to the nominal 99 percent benchmark results in a 99.68 implied performance level, which is a 68 percent movement toward perfect performance.[179] We find that the appropriate application and derivation sample sizes vary with the benchmark level.

---

[176] *See* Appendix K.

[177] One failure in 10 equals 90 percent success. One failure in 100 equals 99 percent success.

[178] *See* Pacific Reply Brief. at 5 (May 5, 2000) A 92.9 level is 30 percent of the interval between 90 and 100 percent.

[179] *Id.* A 99.68 level is 68 percent of the interval between 99 and 100 percent.

Inseparable from the problem of the granularity of the data affecting the implied performance is the affect that any adjustment will have on the established benchmarks. For example if one miss is allowed for a nominal 90 percent benchmark when applied to a sample size of five, then the benchmark percentage is effectively changed to 80 percent. Using the example of 20 CLECs with samples of five cases each as discussed above, all 20 CLECs can experience 80 percent performance without failures being identified. The overall performance for the total submeasure would be 10 percent below the nominal benchmark.

Staff has summarized the net changes to the nominal benchmarks in Appendix K. It is clear that when the adjustment tables are used, the benchmarks are substantially lowered. Recognizing these potential changes, we conclude that the implied performance level should set to address what is analogous to a Type I error without disproportionately increasing what is analogous to a Type II error. In other words, the implied performance level allowance should be higher from the nominal benchmark to a similar degree as the adjusted benchmark is effectively lowered from the nominal benchmark. With this balance in mind, we find that the application and derivation sample sizes recommended by staff in Appendix K, are more appropriate than the parties' proposals. Consequently we shall order the ILECs to use the small sample adjustment tables presented in Appendix K.

In comments to the draft decision, the CLECs object to the size of the application and derivation sample sizes stating that they are larger than necessary to address granularity. AT&T, et. al. Comments at 6 (December 18, 2000). However, we point out that because of granularity (i.e., integers) without adjustment tables the net effective percentage criterion is always higher than the

nominal percentage except when the sample size is an exact multiple of the allowed missed percentage. (For example, sample sizes 10, 20, 30, 40… allow 90 percent net percentage results for the benchmark that allow 10 percent misses – the 90 percent benchmark. See Appendix K for a discussion.) We have made a judgement to address only some of that granularity, limiting our adjustment with the explicit criteria described in Appendix K.

The CLECs also object to the new tables fundamentally because they "harm CLECs by allowing more misses before finding a violation of the benchmark." AT&T, et. al. Comments at 20 (December 18, 2000). The CLECs fail to consider that compared to both the ILEC and CLEC proposals, our application of these tables is more restrictive. Any time the CLEC industry-wide aggregate fails the benchmark, the adjustment tables are not used for CLEC-specific assessment. Our application is tailored to address conditions where actual performance result information indicates granularity most likely is a problem. (See Appendix K.)

Benchmark statistical testing

Pacific and Verizon CA also favor complete statistical testing for all benchmarks. They assert that benchmarks are subject to the same random variation problems as are parity measures. However, Pacific only acknowledges the effect of random variation on alpha and only presents remedies for alpha. We are concerned that these adjustments increase beta, and since we are at least as concerned about effects on beta, we are reluctant to make the statistical adjustments recommended by Pacific. Additionally, we interpret benchmarks to be absolute performance limits that define a "meaningful opportunity to compete." Pacific argues that the benchmarks were created before statisticians were involved and before performance data was available, and thus the

"negotiators relied on their experience in telephony and the needs of the CLECs to arrive at plausible benchmarks," and "did not fully appreciate. . . or consider. . the potential effects of random variation. . . ."[180] Yet Pacific goes on to admit that benchmarks were set recognizing that "the process in question is not completely controllable." (Id.) Pacific's speculation about what was in the minds of the negotiators is contradictory and unpersuasive. We have no confidence in basing a new statistical overlay on such speculation, as we similarly have no confidence in rejecting telephony expertise for statistical expertise.

It is clear to us that the benchmarks already allow for some random variation – no benchmark requires all services to be completed within a certain time period, and no benchmark sets a limit on the degree of any one service's outcome. For example, if the benchmark is 90% of orders completed within 4 days, and 92 percent of the actual orders were completed in 4 days or less, then Pacific is not held accountable for the random or even non-random variation of the remaining 8 percent. It would make no difference in the remedies plan whether these orders were completed within 5 or 100 days.

We are concerned that adding any additional tolerance margin to existing tolerance margins would allow two or three bites at the same apple. We prefer that if the benchmarks are not consistent with their definition of performance that will allow "a meaningful opportunity to compete," that they be adjusted directly, rather than add all the complexities and ambiguities that a new statistical overlay would create. With the inclusion of the adjustment tables we

[180] Pacific's Reply Brief at 4 (May 5, 2000).

specify above, we shall order that benchmarks be treated as tolerance limits. This is an issue that may be re-examined in the incentive payment phase.

Benchmark modification

Closely related to the problems that the adjustment tables and statistical tables are intended to address is the benchmark levels themselves. One possible view is that instead of using adjustment tables that the benchmarks themselves be adjusted. However, since the adjustment depends on the sample size, different benchmarks would have to be set for different sample sizes. This would be virtually the same as using adjustment tables with the current benchmarks. Consequently, we will not order a review and revision of the benchmarks at this time.

Correlation analysis

All parties agree that performance measures that are correlated because they are redundant should be treated so that multiple payments are not made for the same failure. At the same time, parties recognize that a statistical correlation alone cannot distinguish between failure redundancy and multiple instances of independent discrimination. No party wishes to implement a self-executing statistical correlation component to reduce payment for discrimination. Since our immediate concern here is for the self-executing performance remedies plan, we do not order any statistical correlation component to our decision model at this time.

We also find that parties presented correlation analysis only as an abstract concept. No implementable plans were described or proposed. If any party wishes for us to consider a correlation plan we ask that they describe a plan down to the level of detail that will allow implementation. For example, it will be important to understand what data will be analyzed, what analyses will be

employed, what decision criteria will be used, and what follow-up will be used to distinguish redundancy from multiple discrimination. The plans should provide numerical examples so there is no misunderstanding about the necessary specificity of the plan.

Historical data

While our discussion here has necessarily focused on ILEC performance relative to CLEC performance at fixed time periods, ORA raises important issues about absolute performance levels. It is concerned that ILEC performance, and thus performance on behalf of the CLECs, could deteriorate over time, possibly because an ILEC's OSS systems were not constructed sufficient to handle the necessary CLEC business. Consequently, ORA is concerned that ratepayers would suffer poorer service overall, which could offset any gains that the new competitive market could provide. We agree that this is a legitimate concern, and in another phase of our review of Pacific's Section 271 application we have instituted volume testing to address this concern. However, we realize that even the best-designed test cannot anticipate all future variables. While we do not currently have anything in the record to support ordering any self-executing historical data-tracking incentives model component, we will be asking the parties to add monitoring capability to the overall plan. We shall order that at a minimum, certain performance data be monitored and analyzed for trends over time. We shall direct the parties to present proposals by the end of the trial period that would accomplish this monitoring and analysis.

Identical models for ILECs

The two ILECs, Pacific and Verizon CA, differ on an important component of our decision model. Pacific prefers to use the Modified Z-test for average-based measures whereas Verizon CA prefers to use permutation testing for these

measures. We considered creating two different versions of our model to accommodate these preferences, but have decided to require the same model for both ILECs.

We have carefully analyzed all proposed model elements and have made the selections most consistent with our selection criteria. As such, our model represents the best model we could specify from the information in this record.

Additionally, since Verizon CA will in effect be a CLEC seeking access to Pacific's OSS services, and Pacific will in the same manner be a CLEC seeking access to Verizon CA's OSS services, it would not fit our criterion of fairness to allow different performance assessment methods for the two ILECs. For competition to be optimal, the playing field must be as level as possible. The two ILECs must be held to the same standard. For example, it is likely that for some average-based measures, given the same results, the permutation test would show the results as a "pass" while the Modified Z-test would show the same result as a "failure." For the above reasons, we order the same decision model for both ILECs.

Payment retroactivity

Verizon CA asks that the Commission hold any performance remedies plan incentive payments in an escrow account until the end of the trial period. However, since we expect that Pacific will be making its Section 271 application on the basis of the trial period having a self-executing performance remedies plan, we do not wish to allow retroactive adjustments. To do so would in essence nullify the self-executing nature of the plan. In other words, a self-executing plan is one that will trigger incentive payments without any new decisions; the decision model automatically makes decisions. If retroactive changes are made after new consideration, debate, and decisions, then the plan is

not truly self-executing. We are also concerned that allowing retroactive payment alteration will make the already difficult decision model development task more cumbersome.

Some "calibration" with actual data will be helpful in assessing our decision model and its effects on the overall plan, and we will order a calibration period to occur simultaneously with the incentive payment setting phase of this proceeding before the trial period begins. We are concerned that retroactively allowing payment amounts to be adjusted at the end of the trial period will cause the parties' positions regarding the appropriateness of the decision model to be too influenced by their own corporate outcomes, relative to being influenced by the criteria we have described herein. For the above reasons, the trial incentive payments shall be made consistent with the self-executing function of the plan to be determined before the trial period begins. Incentive payment amounts shall not be altered retroactively unless we specifically provide for such alteration in the final plan.[181] In comments on the draft decision, Verizon raises legal questions that we intend to resolve before a final plan is adopted. Verizon Comments at 5, 16-19.

[181] Our discussion and decision on retroactivity does not address the issue of the correction of mistakes in the data or calculations necessary to arrive at incentive payments. This correction issue should be resolved in the incentives phase of these proceedings.

Other issues

Z-statistic negative/positive interpretation

The Modified Z-test statistic becomes a negative or positive value depending on whether the average CLEC performance measurement result (mean) is larger or smaller than the ILEC result (mean), and depending upon whether the CLEC mean is subtracted from the ILEC mean or vice-versa.[182] We note that potential[183] non-parity performance is represented by a negative Z-statistic in both the New York remedies plan and the Louisiana proposed remedies plan and by a positive Z-statistic in the Texas plan. While there would be some merit in constructing our decision model to be consistent with other states, given the already established inconsistency, we must base our decision on some other criterion. We prefer the convention that is most likely understood by those with little statistical sophistication. Because the typical connotations of the words "negative," "discrimination," and "failure," are similar, and the

---

[182] For the sake of this illustration, assume the average time taken for Pacific to provision a hypothetical service for its own customers is 7 days and the average time taken for Pacific to provision service for a CLEC customer is 14 days. In this case, a longer time is worse performance and could create a barrier to competition. If the ILEC mean is subtracted from the CLEC mean (14 – 7 = +7), then a positive Z-test statistic represents a potential non-parity condition. But if the CLEC mean is subtracted from the ILEC mean (7 – 14 = -7), then a negative Z-test statistic represents a potential non-parity condition. This would be reversed for measures where a larger number represents better performance. For consistency in the interpretation of the Z-statistic, the order of the means (i.e., which mean is the subtrahend) must be reversed for situations where larger numbers represent worse performance compared to situations where larger numbers represent better performance.

[183] We use the term "potential" here because non-parity identification will also depend on the magnitude of the Z-statistic (i.e., it must be either a larger positive value than a positive critical value or a larger negative value than a negative critical value).

connotation of "positive" is opposite from these other words, we prefer the Z-test be implemented with a negative Z-value representing potential discrimination. Reading "negative" values to represent negative outcomes is intuitively understandable whereas the reverse is not. Therefore, we shall order our decision model constructed so that negative Z-values represent potential discrimination.

Performance Measure 42

In comments to the draft decision, Pacific pointed out that Performance Measure 42 was unique, and that proposed statistical tests could not be appropriately applied. Pacific Comments at 3 (December 18, 2000). Pacific proposed that for the parity submeasures within Measure 42, "the ILEC percentage minus the CLEC percentage should not exceed 0.05 percentage points. (Reflected in proportions, this difference would be 0.0005)."[184] ORA agrees that Pacific's proposal is appropriate. ORA Reply Comments at 5 (December 22, 2000). As other parties are silent regarding Pacific's proposal, we assume no objections. As Pacific's proposal seems reasonable and has either explicit or implicit concurrence of other parties, we shall include it as part of the decision model we adopt today.

[184] For example, for "systems available 500 hours during a month, this difference translates into a total discrepancy of 15 minutes." Pacific Comments at 4 (December 18, 2000).

Parity performance measures without sufficient ILEC data

Parity comparisons cannot be made without ILEC performance data. Since there may be insufficient ILEC activity in some months for some measures, we need to specify alternative retail analogues. Tests that require standard deviation calculation require at least two observations and exact tests require at least one observation. Pacific proposes that the prior six months of ILEC data be aggregated (to the extent that such data exist) and used in place of the data-deficient month, and if the aggregate does not produce sufficient ILEC data, the submeasure not be evaluated for the month. Pacific Comments at 19 (December 18, 2000). The CLECs agree with the exception that they wish to use the prior three months CLEC data as a surrogate analogue instead of failing to evaluate the performance results. AT&T, et. al. Reply Comments at 3 (December 22, 2000). We agree with Pacific's proposal. Using historical CLEC data may confound discriminatory behavior with seasonal fluctuations. If there is no retail analogue for six months, parties should consider creating a benchmark to assess performance.

Interim and permanent models

As recommended by the ACR, the model we now adopt is an interim model that will generate incentive payments once we have added the incentive components in the next phase of this proceeding. After six month's experience with the model we will review its performance and adjust any component that we find needs changing. Implementing this model as a fully functioning and self-executing performance remedies plan will allow Pacific to

file its section 271 application for entry into the in-region interLATA long distance market. At the same time, this trial period will allow actual experience to guide future refinements. While any party can at any time petition us to change the model, we will remove that burden of persuasion by scheduling this review and adjustment opportunity. As discussed in detail above, there are many unresolved issues regarding what would be the best and most appropriate model. We find that we cannot resolve all these issues. Yet at the same time, we conclude that we can proceed with a fully implementable model while gaining the experience necessary for future development of a permanent model.

Comments on Draft Decision

The draft decision of ALJ Jacqueline A. Reed in this matter was mailed to parties in accordance with Pub. Util. Code § 311(g)(1) and Rule 77.7 of the Rules of Practice and Procedure. Comments were filed on December 18, 2000, and reply comments were filed on December 22, 2000. We have taken the comments into account, as appropriate, in finalizing this decision. As this is an interim decision, there will be an opportunity for us to consider and implement modifications before a final decision is adopted.

Findings of Fact

1. The cornerstone of any performance incentive structure is how parity is defined, since it is on those occasions when an ILEC is out of parity that incentive payments will be made.

2. This Commission's definition of parity incorporates the objectives of the TA96 and the FCC.

3. It will be helpful to rely on statistical testing and benchmarks to infer whether or not parity has been achieved.

4. In late fall 1999, the existent ILEC models and the CLECs' model were distinct and irreconcilable.

5. The parties revealed considerable misunderstanding and confusion about the two sets of respective model assumptions and calculations.

6. The outcomes of the two models were highly discrepant because both approaches were trying simultaneously to design and implement the total model (both the performance assessment model elements and the incentive plan elements) without the benefit of an implementation and data calibration structure.

7. It is unlikely that either model could be implemented as designed.

8. During the February 1999 technical workshop, each proposed plan produced dramatically different payments due to different input assumptions.

9. There is a need to have one common interim model framework of analyses for review and discussion in order to implement the performance remedies plan.

10. To achieve a common model framework, the performance assessment model elements and the incentive plan elements need to be separated.

11. Since the task of accurately assessing the state of competitive conditions must be self-executing, the decision model must be able to automatically identify performance result levels that reveal competition barriers and that will trigger incentive payments.

12. There are two fundamental categories of performance measures that must be assessed to determine the existence of competitive conditions: “parity” and “benchmark” measures.

13. In identifying parity or non-parity, accurate remedies-plan decision-making involves more than accurately calculating average ILEC and CLEC

performance and identifying non-parity if ILEC service to CLEC customers is significantly worse than ILEC service to ILEC customers.

14. Given that there is variability in ILEC performance in providing retail services to its own customers, a measurement showing inferior service to CLEC customers could be due either to this variability, or actual discrimination, or both.

15. Statistical testing allows estimation of decision accuracy, or in other words, calculation of the decision error probabilities.

16. These probabilities can then assist decision-making by quantifying the different error probabilities and comparing them to standards of confidence that the Commission wishes to apply.

17. Using measures of performance averages and variability, statistical analysis provides estimates of: (1) the probability that a result of a certain magnitude would be detected when it exists (test power and corresponding error beta) and (2) the probability that the result is due to random variation when in fact there are no differences (confidence level and corresponding error alpha).

18. Benchmarks have been constructed as tolerance limits.

19. The issues for statistical analysis accuracy of benchmarks are not the same as those for parity measures.

20. None of the presented models for parity assessment are acceptable in their entirety.

21. Four types of measurements have been developed for monitoring ILEC performance: averages, percentages, indexed and rates.

22. Each measurement type requires a different statistical test or a variant of the same test.

23. All parties have agreed that a one-tailed statistical test should be used.

24. In response to the CLECs' concerns that ILEC discrimination could increase the CLEC variance, and thus make it more difficult to detect any discrimination, all parties agreed to use a Modified Z-test instead of the standard Z-test.

25. According to the statistical literature, requiring normally distributed data in the use of any Z-test may be only partially correct.

26. The Central Limit Theorem states that for sufficiently large samples, non-normality in the data does not affect the test.

27. The permutation test has the potential for being a more accurate test that can handle small samples.

28. The Z-test relies on the resulting sampling distributions being approximately normal.

29. The few data examples we have available to us comparing permutation and Z-tests show the expected divergence for small samples, but not the expected convergence for larger samples, contrary to the theoretical expectation that the results should be the same for large sample sizes.

30. The results of the few available data examples raise doubts that the record is sufficiently developed to allow the Commission to confidently select the permutation test as a superior test for average-based measures.

31. In the interim, the Z-test is the most developed and accepted alternative to permutation testing.

32. The advantage of exact tests for the Commission's statistical model is two-fold: (1) calculations are made directly from the raw data, and (2) the exact tests have the potential to produce more accurate results for small samples.

33. Unlike for average-based permutation applications, outliers cannot affect the result of the Fisher Exact test, as the data consist only of "cell counts."

34. Additionally, unlike for average-based permutation applications, the results from the percentage-based Modified Z-test and the results from the Fisher's Exact Test converge towards equality as theoretically expected.

35. The Fisher's Exact Test generates computationally difficult numbers that unnecessarily drain computer resources for no benefit in accuracy for large samples.

36. The Fisher's Exact Test is appropriate and can be calculated up to a limit of 1000 CLEC performance "hits" or "misses," and the Modified Z-test for proportions is appropriate for performance results above this limit.

37. Like the percentage-based Fisher's Exact test applications, and unlike for average-based permutation applications, the results from the rate-based Modified Z-test and the results from the binomial exact test converge towards equality as theoretically expected.

38. Balancing alpha and beta to be equal only ensures that the most accurate decision is made, not what the relative consequences of those decisions will be.

39. The record is relatively silent on the actual beta values that various critical alpha levels might produce.

40. The record is relatively silent on the appropriate test power or beta error level.

41. The record is incomplete regarding what performance level deltas would be, because no party has submitted any proposal containing a comprehensive set of specific deltas.

42. A fixed alpha is not an adequate long-term solution.

43. Test power is very low for the small samples that represent the majority of the performance measure results.

44. Fixed alphas that provide better test power for small samples result in unnecessarily high test power for large samples.

45. A larger alpha level of 0.10, instead of the 0.05 level, enhances decision accuracy and avoids uncorrectable decision-making errors while still addressing correctable errors in the next phase of this proceeding.

46. A smaller alpha level than 0.15 is reasonable because of concerns about the effect on large-sample results.

47. An 80% confidence level (0.20 alpha) in the model for conditional failure identifications is warranted because of the high beta error still remaining when using the 0.10 alpha level.

48. Both record efforts to establish "material" thresholds have merit.

49. The "material difference" standard has merit and the potential to improve the decision model we specify.

50. Minimum sample size requirements vary depending upon the type of statistical test used.

51. Harmful ILEC performance in small new or innovative market niches, or harmful ILEC performance to smaller CLECs, could be masked by relying on assessments of larger market samples or larger CLEC samples when the results for CLECs are aggregated.

52. It is important to examine performance at the smaller market and smaller CLEC levels.

53. There are unresolved issues regarding minimum sample size and sample aggregation rules, and the rules for incentive payments are integrated with the aggregation rules.

54. Minimum sample size rules result in some data being discarded.

55. Our small sample aggregation rules avoid discarding any data and increase sample sizes for the very smallest samples with minimal impact on the actual results.

56. The previously proposed sample size rules are complicated and fall short of our goal of simplicity.

57. The fundamental problem with small sample sizes for parity measures is that they fail to satisfy the normality assumptions for the Modified Z or *t*-test.

58. Statistical texts indicate that the *t*-distribution is more appropriate than the Z-distribution for tests between two sample means, especially for small samples.

59. Using the t-distribution table would adjust for decreasing sample size.

60. Percentage and rate-based measures are assessed using exact tests, which do not depend on inferences or assumptions about underlying distributions.

61. A log transformation (1) brings the distributions much closer to normality, and (2) provides a reasonable interpretation of skewed data.

62. ILEC distribution normality is improved when log transformations are used.

63. Log transformations also change the effect of outliers.

64. Log transformation improves normality for large samples.

65. Log transformations provide a more appropriate Modified *t*-test application than an application using data that is not transformed.

66. Although the ILECs and the CLECs agree to use a benchmark adjustment table, they disagree on two aspects of such tables, sample sizes to which they will be applied and sample sizes from which they will be derived.

67. A fixed derivation sample size results in varying levels of increased implied performance relative to the benchmark limit.

68. The appropriate application and derivation sample sizes vary with the benchmark level.

69. When the adjustment tables are used, the benchmarks are substantially lowered.

70. The application and derivation sample sizes recommended by staff in Appendix K, are more appropriate than the parties' proposals.

71. Benchmarks are absolute performance limits that define a "meaningful opportunity to compete."

72. Benchmarks already allow for some random variation – no benchmark requires all services to be completed within a certain time period, and no benchmark sets an upper limit on any one service's outcome.

73. Performance measures that are correlated because they are redundant should be treated so that multiple payments are not made for the same failure.

74. No party wishes to implement a self-executing statistical correlation component to reduce payment for discrimination.

75. Parties presented correlation analysis only as an abstract concept; no implementable plans were described or proposed.

76. Allowing retroactive adjustments would nullify the self-executing nature of the performance remedies plan.

77. Reading "negative" values to represent negative outcomes is intuitively understandable whereas the reverse is not.

78. A special index must be created for performance measure 42 since the proposed parity statistical tests cannot be appropriately applied.

79. Parity comparisons cannot be made without ILEC performance data and alternative retail analogues must be created for months where there is insufficient ILEC data.

80. Tests that require standard deviation calculation require at least two observations and exact tests require at least one observation.

81. Using the prior six months of aggregated ILEC data be aggregated (to the extent that such data exist) is an appropriate alternative retail analogue.

82. Using historical CLEC data as a surrogate for a retail analogue may confound discriminatory behavior with seasonal fluctuations.

83. The present fully implementable model is an interim one that will generate incentive payments once we have added the incentive components in the next phase of this proceeding.

Conclusions of Law

1. Parity means that the ILEC is providing services to the CLECs in substantially the same period of time and manner (including quality) as it is providing to itself.

2. This Commission endeavors to ensure that the CLECs have OSS access that is at least equal to the ILECs' own access.

3. One interim performance remedies plan model and set of explicit assumptions would allow common quantitative analyses to be performed and estimates to be developed.

4. A single model approach would allow the Commission to make informed and fair policy decisions about the performance remedies plan.

5. A single model approach focuses on the goal of parity service by the ILECs, economic incentives paid by the ILECs, and/or a change in ILECs' operations support to the CLECs.

6. A single interim model and a single set of explicit assumptions should allow calibration of economic outcomes both before and after a six-month pilot test period using actual empirical data.

7. The interim pilot test period will assist the Commission in determining the appropriate levels of long-term economic incentives.

8. Long-term incentive impacts can be calibrated in relation to one model, one common set of assumptions, and actual test period empirical data.

9. Statistical testing should be used to assess the balance between finding and preventing actual barriers, and avoiding the identification of barriers when they do not exist, thus enabling greater decision quality and attainment of legislative goals.

10. A new "hybrid" of elements from each of the different models presented in this proceeding constitutes the most appropriate performance remedies statistical model.

11. Consistent with academic texts and with the FCC's view of the appropriate statistical application regarding the requirements of the Act, a one-tailed test is appropriate for situations where there is only interest in outcomes in one direction, in this case where the CLEC performance results are worse than the ILEC results.

12. The selection of the appropriate test for small samples should be based on the relative accuracy of the different tests.

13. It is reasonable for our sample aggregation rules to act as an interim solution and a "floor" for sample sizes.

14. Evidence in this proceeding is compelling that normality cannot be assumed for small samples since measures of time-delay are commonly skewed – the distribution is "bunched up" for shorter delays, and tapers off slowly for longer delays.

15. Until the Commission can determine which test is the more appropriate treatment of the data, including underlying issues such as "production output"

versus “larger process population sampling” and more specific issues regarding outlier treatment, it is not reasonable to either approve or order use of the permutation test.

16. There is a need to better understand what the appropriate sample sizes are for using the permutation test versus the Modified Z or *t*-test.

17. Since there are unresolved questions surrounding the potential of the permutation test, the active interested parties in this proceeding should collaboratively conduct or fund a research inquiry to answer these unresolved questions.

18. In the case of the percentage-based performance results data, the Fisher’s Exact test is appropriate.

19. The Fisher's Exact test should be used for percentage-based performance results because it provides accurate decision error probabilities, is consistent with theoretical assumptions, solves the Z-test application problems.

20. The binomial exact test should be used for rate-based performance results because it provides accurate decision error probabilities, is consistent with theoretical exceptions, solves the Z-test application problems, is preferred by most parties.

21. The question of relative risk is more appropriately addressed in this proceeding’s next phase, which will establish the “consequences” for the performance decisions made in the present phase.

22. To remedy the lack of critical record information, it is reasonable to direct the ILECs to calculate both alpha and beta values whenever a statistical test is applied.

23. As a general policy statement, it is reasonable to assume that a Type II error is at least as important as a Type I error. Apparent discrepancies can be adjusted in the incentive payment phase.

24. It is reasonable that the problems of insufficient test power for small samples (large beta) and "too much" test power for large samples can be better resolved through even approximate alpha / beta balancing techniques.

25. A fixed alpha critical value should only be used in the model as an interim decision-criterion solution.

26. The 90% confidence level (0.10 alpha, or 10% significance level) should be adopted in the statistical model to control the Type I error and to reduce the Type II error to more acceptable levels for the preponderance of the performance results.

27. Pending establishment of applicable conditions, the 80% confidence level (0.20 alpha) should be adopted in the statistical model for conditional failure identifications because of the low power of these tests.

28. The parties should be directed to devise and propose specific conditional failure identifications in the next phase of this proceeding.

29. One goal of the performance remedies plan is to assess each CLEC's performance results for each submeasure.

30. The smaller market and smaller CLEC levels may be critical for entry and innovation, which in turn are critical to a healthy competitive telecommunications infrastructure.

31. Consistent with the academic justification of the Modified Z-test, the test statistic should be compared to the *t*-distribution.

32. The small sample aggregation rules we have designed should be easily understood with the results easily reproduced.

33. To assess performance subject to the performance remedies plan, statistical analysis and decision rules should be applied to all data, including sample sizes as small as one case, after our small sample aggregation rules are applied.

34. How payments will be triggered or allocated under the aggregation rules should be addressed in the upcoming incentives phase.

35. All percentage and rate-based data at the submeasure level for each CLEC should be analyzed for parity regardless of small sample sizes since exact tests are accurate for all sample sizes.

36. Staff's analyses of several ILEC and CLEC distributions demonstrate that even in cases where the log transformation dramatically changes results from the non-transformed data, the transformed results are reasonable and appropriate treatments of the performance data.

37. Log transformations of the data should not be ordered on a permanent basis until the record is adequately developed in subsequent phases of this proceeding.

38. More exact tests should be used in addressing small sample size issues, if subsequent research shows them to be appropriate.

39. The log transformation is reasonable and appropriate, and is necessary at least as an interim solution for application of the Modified Z-test to small to moderately large samples.

40. Log transformations should be utilized for all average-based performance measures as specified in Appendix J.

41. The meaning of outliers should be discussed in the incentives phase of this proceeding.

42. Because of the legitimacy of the benchmark small sample problem, and since the CLECs have agreed to some adjustments, a benchmark small sample adjustment table should be ordered as part of the decision model.

43. It is appropriate to set different application sample sizes for different benchmark percentage levels.

44. The implied performance level should be set to address what is analogous to a Type I error without disproportionately increasing what is analogous to a Type II error.

45. The ILECs should use the small sample adjustment tables presented in Appendix K.

46. If any benchmark is inconsistent with the performance definition "a meaningful opportunity to compete," it should be adjusted directly rather than add all the complexities and ambiguities that a new statistical overlay would create.

47. Benchmarks should be treated as tolerance limits; however, the issue may be re-examined in the incentive payment phase.

48. A review and revision of the benchmarks should not be ordered at this time because it could be more cumbersome than using adjustment tables with the current benchmarks, and establishing benchmarks is the subject of a different proceeding.

49. Since parties recognize that a statistical correlation alone cannot distinguish between failure redundancy and multiple instances of independent discrimination, we should not order any statistical correlation component to our self-executing performance remedies plan model.

50. Any party seeking to have a correlation plan considered in the next phase of this proceeding should describe the plan down to the level of detail that will

allow implementation. Parties should provide numerical examples so there is no misunderstanding about the necessary specificity of the plan.

51. The parties should present proposals by the end of the trial period that would put into effect the monitoring and analysis of certain performance data for trends over time.

52. The same performance remedies model should be applied to both Pacific and Verizon CA in the interest of fairness.

53. Since some "calibration" with actual data will be helpful in assessing our decision model and its effects on the overall plan, a calibration period should be ordered to occur simultaneously with the incentive payment setting phase of this proceeding before the trial period begins.

54. Allowing retroactive payment alteration will make the already difficult decision model development task more cumbersome.

55. Incentive payment amounts should not be altered retroactively.

56. Following a six-month trial period, to be specified in the incentive payment phase of this proceeding, the performance of the remedies plan model should be reviewed and any component determined to need changing should be adjusted.

57. A fully implementable interim model should be utilized while gaining the experience necessary for future development of a permanent model.

58. This decision should become effective immediately so that the calibration process can begin and the incentive payment phase may proceed.

INTERIM ORDER

IT IS ORDERED that:

1. A performance remedies plan decision model, which identifies performance failures and non-failures, as specified in Appendix C incorporated by reference herein, shall be adopted for Pacific Bell (Pacific) and Verizon California Inc. (Verizon CA).

2. The performance remedies plan, comprised of the decision model adopted herein and an incentive payment component that will be determined in the next phase of this proceeding, shall be implemented for a trial period of six months.

3. Pacific and Verizon CA shall use the Modified *t*-test for average-based parity performance measures.

4. Log transformations shall be utilized for all average-based performance measures as specified in Appendix J.

5. Pacific, Verizon CA and the active interested competitive local exchange carriers (CLECs) in Rulemaking 97-10-016/Investigation 97-10-017 shall collectively conduct or fund a research inquiry into whether the permutation test or the Modified *t*-test is the more appropriate treatment of the data, including but not limited to underlying issues such as "production output" versus "larger process population sampling" and more specific issues regarding outlier treatment. The inquiry shall adopt a collaborative research approach so that all interested parties can collectively influence the research proposal.

6. The Fisher's Exact test shall be used for all percentage-based parity results except for those that cannot be computed because of large numbers. Results where the CLEC numerator exceeds 1000 shall be calculated with the Modified Z-test for proportions.

7. The binomial exact test shall be used for all rate-based tests.

8. The performance remedies plan model shall be constructed so that negative Z and *t*-values represent potential discrimination.

9. Pacific and Verizon CA shall calculate and report both Type I (alpha) and Type II (beta) error values whenever a statistical test is applied.

10. The parties shall collaboratively develop and implement an alpha / beta balancing procedure for the statistical model adopted herein and detailed in Appendix G no later than the end of the trial period, unless the parties reach agreement and jointly move to implement the components sooner.

11. If the parties are unable to agree on an alpha / beta balancing decision component for the model by the end of the trial period, the parties shall submit their individual models for Commission review and decision as directed by the assigned Commissioner and / or assigned Administrative Law Judge.

12. Until an alpha / beta balanced criterion is established, fixed alpha critical values shall be adopted for the interim.

13. A 90% confidence level (0.10 alpha, or 10% significance level) shall be adopted as the interim fixed critical value in the statistical model for failure identifications.

14. For the possible implementation of an 80% confidence level (0.20 alpha), the parties shall devise and propose specific conditional failure identifications for our consideration in the next phase of this proceeding.

15. Except for rare submeasures identified in Appendix H, Attachment 1, the following small sample aggregation rules shall be used for average-based parity performance measures: (1) For each submeasure, all samples with one to four cases shall be aggregated with each other; and (2) statistical analyses and decision rules shall be applied to determine performance subject to the

performance remedies plan for all samples after the aggregation in step (1), regardless of sample size.

16. Rare submeasures identified in Appendix H, Attachment 1, shall be analyzed without aggregation and regardless of sample size.

17. How payments will be triggered or allocated under the aggregation rules shall be addressed in the upcoming incentives phase.

18. All percentage and rate-based data at the submeasure level for each CLEC shall be analyzed for parity without aggregation and regardless of sample size.

19. Pacific and Verizon CA shall use the small sample adjustment tables presented in Appendix K.

20. Benchmarks shall be treated as tolerance limits; however, the issue may be re-examined in the incentive payment phase.

21. Pacific, Verizon CA and any interested parties shall present proposals by the end of the trial period that would put into effect the monitoring and analysis of certain performance data for trends over time.

22. The same performance remedies model shall be applied to Pacific and Verizon CA.

23. A calibration period shall occur simultaneously with the incentive payment setting phase of this proceeding before the trial period begins.

24. Following a six-month trial period, to be specified in the incentive payment phase of this proceeding, we shall review the performance of the remedies plan model and adjust any component that we determine needs changing.

This order is effective today.

Dated January 18, 2001, at San Francisco, California.

LORETTA M. LYNCH
President
HENRY M. DUQUE
CARL W. WOOD
Commissioners

Commissioner Richard A. Bilas, being necessarily absent, did not participate.