Appendix A

ACR Questions

Assigned Commissioner's Ruling Questions

Tests for Determining Compliance with Parity

1. A standardized Z-test is proposed for purposes of determining compliance with parity. Explain why this standard textbook statistical test cannot serve as a measurement tool at least for the duration of the six-month trial pilot test period? Keep in mind that the incentive phase of the model can calibrate for measurement outcomes through various incentive plan structures and amounts.
2. Benchmark measures without any statistical tests are proposed for purposes of determining a performance failure. Explain why this simple approach cannot serve as a measurement tool at least for the duration of the six-month trial pilot test period? Keep in mind that the incentive phase of the model can incorporate information on underlying data values and distributions.

Minimum Sample Sizes

1. A minimum sample size of thirty, aggregated in up to three-month time periods, is proposed. Explain why this standard textbook statistical proposal cannot serve as a minimum sample size rule at least for the duration of the six-month trial test period? Keep in mind that the test would still be performed using whatever sample size is achieved at the end of three months.

Alpha Levels/Critical Values

Ten percent Type I alpha level for parity tests is proposed. Explain why this standard textbook statistical proposal cannot serve as an alpha level/critical value rule at least for the duration of the six-month trial pilot test period? Again, keep in mind that the penalty phase of the plan can calibrate the size of the payments as a function of the critical values.

Appendix B

References

Bartz, A. (1988). *Basic statistical concepts*. 3rd ed. New York: Macmillan.

Bickel, P. & Doksum, K. (1977). *Mathematical statistics: Basic ideas and selected topics*. San Francisco: Holden-Day.

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

Brubaker, K. & McCuen, R. (1990) Level of significance selection in engineering analysis. *Journal of Professional Issues in Engineering, 116,* 375-387.

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

Gold, D. (1969). Statistical tests and substantive significance. *The American Sociologist*, *4*, 42 – 46.

Good, P. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. 2nd Ed. New York: Springer Verlag.

Hays, W. (1994). *Statistics*. 5th ed. Fort Worth: Harcourt Brace.

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

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

Johnstone, D. & Lindley, D. (1995). Bayesian inference given data “significant at α”: Tests of point hypothesis. *Theory & Decision*, *38,* 51 – 60.

Khazanie, R. (1997). *Statistics in a world of applications.* 4th ed. Harper Collins.

McNemar, Q. (1962). *Psychological statistics*. New York: John Wiley & Sons.

Raiffa, H. (1970). *Decision analysis*. Reading, Mass.: Addison-Wesley

Sheskin, D. (1997). *Handbook of parametric and nonparametric statistical procedures.* Boca Raton: CRC Press.

Skipper, J., Guenther, A., & Nass, G. (1970). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist*, *2*, 16-18.

Verma, R. & Goodale, J. (1995) Statistical power in operations management research. *Journal of Operations Management*, *13*, 139-152.

Welsh, A.H. (1996). *Aspects of statistical inference*. New York: Wiley & Sons.

Winer, B.J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill

Appendix C

Decision Model

I. Parity measures

A. Statistical Tests

All statistical tests will be one-tailed tests.

1. Average-based Parity Measures

The Modified *t*-test will be used for all average-based parity measures as specified in:

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

The Modified *t*-test for the difference in means (averages) between the ILEC and the CLEC populations is:

$$t = (M_i - M_c) / [S_i * sqrt(1 / N_c + 1 / N_i)]$$

Where:

M_c = the CLEC mean result
M_i = the ILEC mean result
S_i = the standard deviation of the results for the ILEC
N_c = the CLEC sample size
N_i = the ILEC sample size
sqrt = square root

For measures of time intervals, except for data where "zeros" are not possible, the raw score distribution will be normalized by taking the natural log of each score after a constant of 0.4 of the smallest unit of measurement is added to each score. For example, if the smallest unit of measurement is an integer, then the added constant would be 0.4:

$$x_{tran} = \ln(x + 0.4)$$

Similarly, if the smallest unit of measurement is 0.01, then the added constant would be 0.004:

$$x_{tran} = \ln(x + 0.004)$$

Results that are not measures of time intervals (e.g., Measure 34) will not be transformed.

The Modified *t*-test calculation for average parity measures will be structured so that a negative sign indicates "worst" performance. Specifically, when a lower value represents better performance, such as time to provision a service, the CLEC mean will be subtracted from the ILEC mean. Different performance measures may require reversing the means in the equation to have a negative sign indicate poorer performance.

The *t*-statistic will be converted to an α (Type I error) probability using a *t*-distribution table or calculation. Degrees of freedom (*df*) will be based only on the ILEC sample size consistent with Brownie, et al. If the obtained α value is less than the critical α value, then the result will be deemed not in parity.

2. Proportion Parity Measures

Except for performance results that have numbers too large to calculate with the exact test, the Fisher's Exact Test will be used for all percentage or proportion parity measures as specified in:

Sheskin, D. (1997). *Handbook of parametric and nonparametric statistical procedures.* Boca Raton: CRC Press, pp. 221-225.

If the obtained α value is less than the critical α value, then the result will be deemed out-of-parity.

Performance results that are too large to calculate with the Fisher's exact test are those measures that exceed the following values:

1. For percentage-based measures where low values signal good service, Fisher's Exact Test shall be applied to all problems for which the CLEC numerator is less than 1000 "hits." The Z-test shall be applied to larger results.

2. For percentage-based measures where high values signal good service, the analysis is the same but is applied to the "misses" as opposed to the "hits." The Fisher's Exact Test shall be applied whenever the denominator minus the

numerator is less than 1000 for the CLEC result. The Z-test shall be applied to larger results.

Such results will be calculated using the Modified Z-test for proportions as follows:

$$Z = (P_i - P_c) / \text{sqrt}[P_i(1-P_i)*(1/N_c + 1/N_i)]$$

Where:
- P_c = the CLEC proportion
- P_i = the ILEC proportion
- N_c = the CLEC sample size
- N_i = the ILEC sample size
- sqrt = square root

The Modified Z-test calculation for proportion parity measures will be structured so that a negative sign indicates "worst" performance. Specifically, when a higher value represents better performance, such as percent on-time tasks, the ILEC proportion will be subtracted from the CLEC proportion. Different performance measures may require reversing the means in the equation to have a negative sign indicate poorer performance.

The Z-statistic will be converted to an α (Type I error) probability using a Z-distribution table or calculation. If the obtained α value is less than the critical α value, then the result will be deemed not in parity.

3. Rate-based Parity Measures

The Binomial Exact Test will be used for all rate parity measures. The Binomial Exact Test is specified in GTECs Exhibit C, Section 3, "Permutation Test for Rates", Equations 3.1 and 3.2 (Deliverable #7, Facilitated Work Group, April 2000).

4. Indexed-based Parity Measures

Measure 42 provides an index of parity performance that will be assessed by comparing ILEC and CLEC performance as follows:

Non-parity will be identified when the ILEC percentage minus the CLEC percentage exceeds 0.05 percentage points.

B. Critical Alpha Level for Parity Tests

The Type I error probabilities (alphas, α) obtained from the parity statistical tests will be compared to a critical alpha value of 0.10.

A performance result with α equal to or less than 0.10 will be deemed a performance failure with no additional conditions.

A performance result with α equal to or less than 0.20 and greater than 0.10 will be deemed a conditional failure. Additional conditions to determine failures will be specified in the final remedies plan.

C. Sample Sizes and Aggregation Rules

Statistical tests will be applied to the monthly performance results specified in D.99-08-020.

1. Average-based measures

For average-based performance results the following aggregation rules will be used:

(1) For each submeasure, the performance results for all samples with one to four cases will be aggregated with each other to form a single performance result.
(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.

2. Proportion and rate-based measures

All samples will be analyzed as they are reported without aggregation.

D. Measures without Retail Analogues.

In months where there are no retail analogue performance data, 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. If the aggregate does not produce sufficient ILEC data, the submeasure not be evaluated for the month.

II. Benchmark Measures

For large samples, the actual performance will be compared to the benchmark nominal percentage according to the percentage set in the Joint Partial Settlement Agreement approved by the Commission. For small samples, maximum permitted “misses” shall be determined by small sample adjustment tables. Small samples are defined as follows:

90 percent benchmarks - 50 cases or less
95 percent benchmarks - 100 cases or less
99 percent benchmarks - 500 cases or less

Adjustment tables:

90% Benchmark		95% Benchmark		99% Benchmark	
Sample size	Maximum permitted misses	Sample size	Maximum permitted misses	Sample size	Maximum permitted misses
1	0	1 to 3	0	1 to 19	0
2 to 9	1	4 to 19	1	20 to 97	1
10 to 20	2	20 to 40	2	98 to 202	2
21 to 31	3	41 to 63	3	203 to 319	3
32 to 44	4	64 to 88	4	320 to 445	4
45 to 50	5	89 to 100	5	446 to 500	5

The small sample adjustment tables shall be used in the following steps:

1. The number of performance "misses" for the CLEC industry-wide aggregate for each remedy plan benchmark submeasure will be compared to the number of permitted misses for all sample sizes covered by the related adjustment table. Industry aggregate performance will be identified as passing if the number of actual misses is less than or equal to the number of permitted misses, and identified as failing if otherwise.

2. For CLEC industry-wide aggregate sample sizes not covered by the related adjustment table, the actual performance percentage result will be compared to the benchmark nominal percentage value. Industry aggregate performance will be identified as passing if the actual performance percentage result is greater than or equal to the benchmark nominal percentage value, and identified as failing if otherwise.

3. For CLEC-specific analysis, results with sample sizes of four or less will be aggregated into a "small sample CLEC aggregate" for each submeasure. Each small sample CLEC aggregate performance result and all remaining non-aggregated CLEC performance results will be assessed.

4. For each submeasure where the CLEC industry-wide aggregate performance *fails* the benchmark, the actual performance percentage result for each small sample CLEC aggregate and each remaining non-aggregated CLEC result will

be compared to the benchmark nominal percentage value. Each individual or aggregate performance result will be identified as passing if the actual performance percentage result is greater than or equal to the benchmark nominal percentage value, and identified as failing if otherwise.

5. For sample sizes *covered* by the related adjustment table where the CLEC industry-wide aggregate performance *passes* the benchmark, the following shall apply for each submeasure. For each benchmark submeasure, the number of performance “misses” for each small sample CLEC aggregate and each remaining non-aggregated CLEC will be compared to the number of permitted misses. CLEC performance will be identified as passing if the number of actual misses is less than or equal to the number of permitted misses, and identified as failing if otherwise.

6. For sample sizes *not covered* by the related adjustment table where the CLEC industry-wide aggregate performance *passes* the benchmark, the following shall apply. The actual performance percentage result for each small sample CLEC aggregate and each remaining non-aggregated CLEC result will be compared to the benchmark nominal percentage value. Each individual or aggregate performance result will be identified as passing if the actual performance percentage result is greater than or equal to the benchmark nominal percentage value, and identified as failing if otherwise.

Appendix D

Fisher's Exact Test

Fisher's Exact Test

This appendix documents Fisher's Exact Test (FET) calculation methods and presents staff's comparison of Z-test and FET results.

Calculation methods

Calculation methods and examples for percentage measures where lower values represent better performance are presented in Attachment 1. Calculation methods and examples for percentage measures where higher values represent better performance are presented in Attachment 2.

Convergence of Z-test and FET results

Staff compared Type I error values (alpha probabilities) produced by the Z-test with those produced by the FET for one "lower is better" submeasure and one "higher is better" submeasure. Staff found that the results from the two tests converge for large sample sizes. Specifically, the size of the difference between the alphas calculated for each test was highly negatively correlated with the natural log of the CLEC sample size as listed in Table 1. "Highly negatively correlated" means that as sample size increases, the difference between the Z-test alpha and the FET alpha decreases in a close and predictable relationship.

Table 1

Measure type	Sample sizes	N	Correlation coefficient	p
High is better	1 to 100	102	-0.89	0.00
High is better	All	204	-0.74	0.00
Low is better	All	167	-0.94	0.00

The correlation for the whole sample for the "high is better" measure is artifactually smaller than for the half-sample because the difference between the alphas for the two tests reduced to zero and could not diminish further for very large sample sizes. Thus though the convergence was perfect for very large samples, since there was no variation, the correlation was zero for this part of the bivariate distribution.

Table 2 lists the extent of the differences between the alphas for the two tests and illustrates the convergence of the results as sample sizes increase.

Table 2

Measure type	Sample sizes	N	Mean difference	Median difference
High is better	1 to 30	63	0.12	0.09
	31 to 100	39	0.009	0.00
	101 +	102	0.0006	0.00
Low is better	1 to 100	102	0.40	0.44
	101 to 500	27	0.12	0.11
	501 to 1500	21	0.05	0.06
	1500 +	17	0.015	0.02

Mathcad worksheet: Hypothetical data example calculations for Fisher's Exact test. Measures for which low values represent good service.

Data :=

Measure	Cnum1	NC	Cval	Inum	NI	Ival
5	0	21	0.0%	6	598	1.0%
5	1	1	100.0%	231	598	38.6%
5	5	10	50.0%	321	743	43.2%
5	2	21	9.5%	234	598	39.1%
5	21	32	65.6%	321	743	43.2%
5	12	43	27.9%	345	743	46.4%
5	23	76	30.3%	321	743	43.2%
5	21	98	21.4%	210	598	35.1%
11	3	21	14.3%	32	298	10.7%
11	2	32	6.3%	98	678	14.5%
11	2	43	4.7%	76	876	8.7%
11	21	132	15.9%	98	688	14.2%
11	23	210	11.0%	101	678	14.9%
11	1	4	25.0%	8	289	2.8%
11	5	54	9.3%	6	321	1.9%
11	5	123	4.1%	32	832	3.8%
11	12	398	3.0%	34	876	3.9%
11	0	5	0.0%	0	17	0.0%
11	3	54	5.6%	7	65	10.8%
20	2	3	66.7%	65	432	15.0%
20	1	1	100.0%	210	748	28.1%
20	19	32	59.4%	154	746	20.6%
20	21	76	27.6%	111	1231	9.0%
20	3	9	33.3%	110	765	14.4%
20	5	19	26.3%	101	789	12.8%
23	0	1	0.0%	154	987	15.6%
23	1	1	100.0%	54	543	9.9%
23	3	9	33.3%	87	567	15.3%
23	2	10	20.0%	210	1122	18.7%
23	2	5	40.0%	132	876	15.1%

rows(Data) = 30
cols(Data) = 7

$HC := Data^{<1>}$ Numerator for CLEC

$NC := Data^{<2>}$ Denominator (sample size) for CLEC

$HI := Data^{<4>}$ Numerator for ILEC

$NI := Data^{<5>}$ Denominator for ILEC

The following function calculates Fisher's exact test using the above four parameters. If the CLEC numerator (HC) is zero, the probability is 1 regardless of the other parameters
.

$$FE(hc, nc, hi, ni) := \left| \begin{array}{l} x \leftarrow 1 \quad \text{if} \quad hc = 0 \\ x \leftarrow 1 - \text{phypergeom}(hc - 1, nc, ni, hi + hc) \quad \text{otherwise} \\ \text{return } x \end{array} \right.$$

$J := rows(Data) - 1$

$j := 0 .. J$

$p_j := FE\left(HC_j, NC_j, HI_j, NI_j\right)$

$Y := augment(Data, p)$

Measure	Cnum1	NC	Cval	Inum	NI	Ival	Prob
5	0	21	0.0%	6	598	1.0%	100.0%
5	1	1	100.0%	231	598	38.6%	38.7%
5	5	10	50.0%	321	743	43.2%	45.1%
5	2	21	9.5%	234	598	39.1%	100.0%
5	21	32	65.6%	321	743	43.2%	1.0%
5	12	43	27.9%	345	743	46.4%	99.5%
5	23	76	30.3%	321	743	43.2%	99.0%
5	21	98	21.4%	210	598	35.1%	99.8%
11	3	21	14.3%	32	298	10.7%	41.1%
11	2	32	6.3%	98	678	14.5%	95.5%
11	2	43	4.7%	76	876	8.7%	89.6%
11	21	132	15.9%	98	688	14.2%	35.2%
11	23	210	11.0%	101	678	14.9%	94.3%
11	1	4	25.0%	8	289	2.8%	11.8%
11	5	54	9.3%	6	321	1.9%	1.2%
11	5	123	4.1%	32	832	3.8%	53.0%
11	12	398	3.0%	34	876	3.9%	82.3%
11	0	5	0.0%	0	17	0.0%	100.0%
11	3	54	5.6%	7	65	10.8%	91.4%
20	2	3	66.7%	65	432	15.0%	6.3%
20	1	1	100.0%	210	748	28.1%	28.2%
20	19	32	59.4%	154	746	20.6%	0.0%
20	21	76	27.6%	111	1231	9.0%	0.0%
20	3	9	33.3%	110	765	14.4%	13.1%
20	5	19	26.3%	101	789	12.8%	9.1%
23	0	1	0.0%	154	987	15.6%	100.0%
23	1	1	100.0%	54	543	9.9%	10.1%
23	3	9	33.3%	87	567	15.3%	15.3%
23	2	10	20.0%	210	1122	18.7%	58.6%
23	2	5	40.0%	132	876	15.1%	16.8%

Y

Mathcad worksheet: Hypothetical data example calculations for Fisher's Exact test. Measures for which high values represent good service.

Data :=

Measure	Cnum1	NC	Cval	Inum	NI	Ival
9	1	1	100.0%	9	14	64.3%
9	3	3	100.0%	123	145	84.8%
9	6	7	85.7%	78	98	79.6%
9	9	11	81.8%	76	98	77.6%
9	14	15	93.3%	9	14	64.3%
9	17	19	89.5%	77	98	78.6%
9	17	21	81.0%	121	145	83.4%
9	23	24	95.8%	9	14	64.3%
9	24	24	100.0%	120	145	82.8%
26	145	154	94.2%	454	456	99.6%
26	276	287	96.2%	454	456	99.6%
26	321	323	99.4%	454	456	99.6%

rows(Data) = 12
cols(Data) = 7

$HC := Data^{<2>} - Data^{<1>}$ Numerator for CLEC. This value is converted from "hits" to "misses".

$NC := Data^{<2>}$ Denominator (sample size) for CLEC

$HI := Data^{<5>} - Data^{<4>}$ Numerator for ILEC, also converted from "hits" to "misses."

$NI := Data^{<5>}$ Denominator for ILEC

The following function calculates Fisher's exact test using the above four parameters. If the CLEC numerator (HC) is zero, the probability is 1 regardless of the other parameters.

```
FE(hc, nc, hi, ni) := | x ← 1   if  hc = 0
                      | x ← 1 − phypergeom(hc − 1, nc, ni, hi + hc)   otherwise
                      | return x
J := rows(Data) − 1
```

$j := 0 .. J$

$p_j := FE\left(HC_j, NC_j, HI_j, NI_j\right)$

$Y := augment(Data, p)$

Measure	Cnum1	NC	Cval	Inum	NI	Ival	Prob
9	1	1	100.0%	9	14	64.3%	100.0%
9	3	3	100.0%	123	145	84.8%	100.0%
9	6	7	85.7%	78	98	79.6%	80.1%
9	9	11	81.8%	76	98	77.6%	74.9%
9	14	15	93.3%	9	14	64.3%	99.4%
9	17	19	89.5%	77	98	78.6%	93.0%
9	17	21	81.0%	121	145	83.4%	48.9%
9	23	24	95.8%	9	14	64.3%	99.9%
9	24	24	100.0%	120	145	82.8%	100.0%
26	145	154	94.2%	454	456	99.6%	0.0%
26	276	287	96.2%	454	456	99.6%	0.1%
26	321	323	99.4%	454	456	99.6%	55.0%

Y

Appendix E

Binomial Exact Test

Binomial Exact Test

This appendix documents binomial exact test calculation methods and presents staff's comparison of Z-test and binomial test results. Calculation methods and examples for rate measures are presented in Attachment 1.

Convergence of Z-test and binomial exact test results

Staff compared Type I error values (alpha probabilities) produced by the Z-test with those produced by the binomial test for submeasure. As with the Fisher's Exact Test, staff found that the results from the two tests converge for large sample sizes. Specifically, the size of the difference between the alphas calculated for each test was highly negatively correlated with the natural log of the CLEC sample size as listed in Table 1. "Highly negatively correlated" means that as sample size increases, the difference between the Z-test alpha and the binomial test alpha decreases in a close and predictable relationship.

Table 1

N	Correlation coefficient	p
117	-0.93	0.00

Table 2 lists the extent of the differences between the alphas for the two tests and illustrates the convergence of the results for the two tests.

Table 2

Sample sizes	N	Mean difference	Median difference
1 to 100	61	0.32	0.38
101 to 300	37	0.05	0.05
300 +	19	0.008	0.00

Excell spreadsheet formula for binomial exact test calculations

The Excell© worksheet cell entry that calculates alpha for the binomial exact test is as follows:

=1-IF(B2=0,0,BINOMDIST(B2-1, B2+E2, C2/(C2+F2),TRUE))

Using trouble report rates as an example rate performance measure, column B contains “Cnum1,” the number of CLEC troubles; column C contains “N_c” the number of lines, column E contains “Inum,” the number of ILEC troubles; and column F contains N_i, the number of ILEC lines. The above formula is the cell entry for the first row of performance results in the spreadsheet (row 2) presented on the next page. The data is hypothetical data for demonstration purposes only.

Excell spreadsheet: Hypothetical data example of binomial exact test calculations.

	A	B	C	D	E	F	G	H
1	Measure	Cnum1	N_c	Cval	Inum	N_i	Ival	α
2	15	0	143	0.00%	987	1876543	0.05%	1.00
3	15	3	343	0.86%	4321	2012345	0.20%	0.04
4	15	1	432	0.22%	1321	2012345	0.07%	0.25
5	15	4	876	0.45%	4321	2012345	0.20%	0.12
6	15	2	2987	0.07%	3210	2101234	0.15%	0.94
7	15	6	4321	0.14%	2432	2101234	0.11%	0.38
8	15	5	5432	0.08%	2765	1876543	0.15%	0.90
9	15	7	13210	0.05%	1765	2012345	0.09%	0.94
10	15	8	13210	0.06%	4321	2012345	0.20%	1.00
11	16	0	4	0.00%	32	14321	0.21%	1.00
12	16	3	12	25.00%	876	7654	10.66%	0.16
13	16	2	13	15.38%	987	43210	2.20%	0.04
14	16	8	21	40.00%	876	7654	10.66%	0.00
15	16	1	21	4.55%	1231	48765	2.56%	0.41
16	16	3	76	3.90%	876	7654	10.66%	0.99
17	16	9	98	9.38%	543	21012	2.65%	0.00
18	16	6	132	4.62%	12101	543210	2.32%	0.08
19	16	7	187	3.83%	8987	432101	1.96%	0.10
20	16	4	198	2.06%	10123	498765	2.12%	0.57
21	16	5	365	1.39%	11012	454321	2.45%	0.94
22	19	0	1	0.00%	2799	54321	4.91%	1.00
23	19	2	18	11.11%	1012	321012	0.35%	0.00
24	19	1	54	1.82%	1012	321012	0.34%	0.16
25	19	8	54	13.56%	2987	65432	4.89%	0.00
26	19	7	87	7.95%	2987	65432	4.86%	0.11
27	19	0	87	0.00%	26543	3432101	0.72%	1.00
28	19	5	321	1.61%	987	301234	0.31%	0.00
29	19	9	876	1.09%	1876	210123	0.90%	0.38
30	19	4	987	0.44%	26543	3654321	0.72%	0.93
31	19	6	1210	0.47%	143210	12345678	1.34%	0.99

Appendix F

Beta Error

Beta Error Levels

This appendix documents staff's analyses of beta error levels for various performance. Staff prepared two analyses. The first analysis examined betas for all possible parity measures using Modified Z-test calculations for all measure types. While these are not the test applications that the Commission will implement, using these tests allows some comparisons that are otherwise difficult. These values are calculated from May 2000 performance data. The alternative hypothesis posed for all estimates was that the CLEC's results were at least 50 percent worse than the ILEC's results. The formula used is based on Hays, *supra* at 284-289 (1994) except that the ILEC and CLEC sample sizes are used:

$$t_{\beta} = (H_0 - H_{alt}) / SD_m$$

The second analysis examined beta error levels for all parity measures as implemented by the Commission in this decision with the exception that log transformations were not performed. These values are calculated from July through September, 2000 performance data. The above formula was used for the average-based parity measures. Pacific's Dr. Gleason calculated the betas for the percentage and rate measures using the hypergeometric and binomial distributions, respectively.

Table 1 lists beta values calculated from Pacific's May, 2000 performance data as described above. Calculations are presented for four different critical alpha levels and for two alternative hypotheses. The alternative hypotheses represents performance provided to CLECs that is 50 percent worse (150%) and 100 percent worse (200%) than performance the ILEC provides itself. For example, the mean beta value for a critical alpha level of 0.10, given an alternative hypothesis of 50 percent worse performance, is 0.63. This should be interpreted as: If we keep Type I error to a maximum of 10 percent (α_{crit} = 0.10), on average we will experience a 63-percent error rate when trying to detect performance for the CLEC that is at least 50 percent worse than performance for the ILEC.

Table 1

Average Beta values for Pacific May, 2000, parity measures				
	Alternative hypothesis			
	150%		200%	
Critical α	Mean	Median	Mean	Median
0.05	0.70	0.88	0.58	0.77
0.10	0.63	0.79	0.51	0.64
0.15	0.57	0.72	0.45	0.55
0.20	0.52	0.65	0.40	0.47

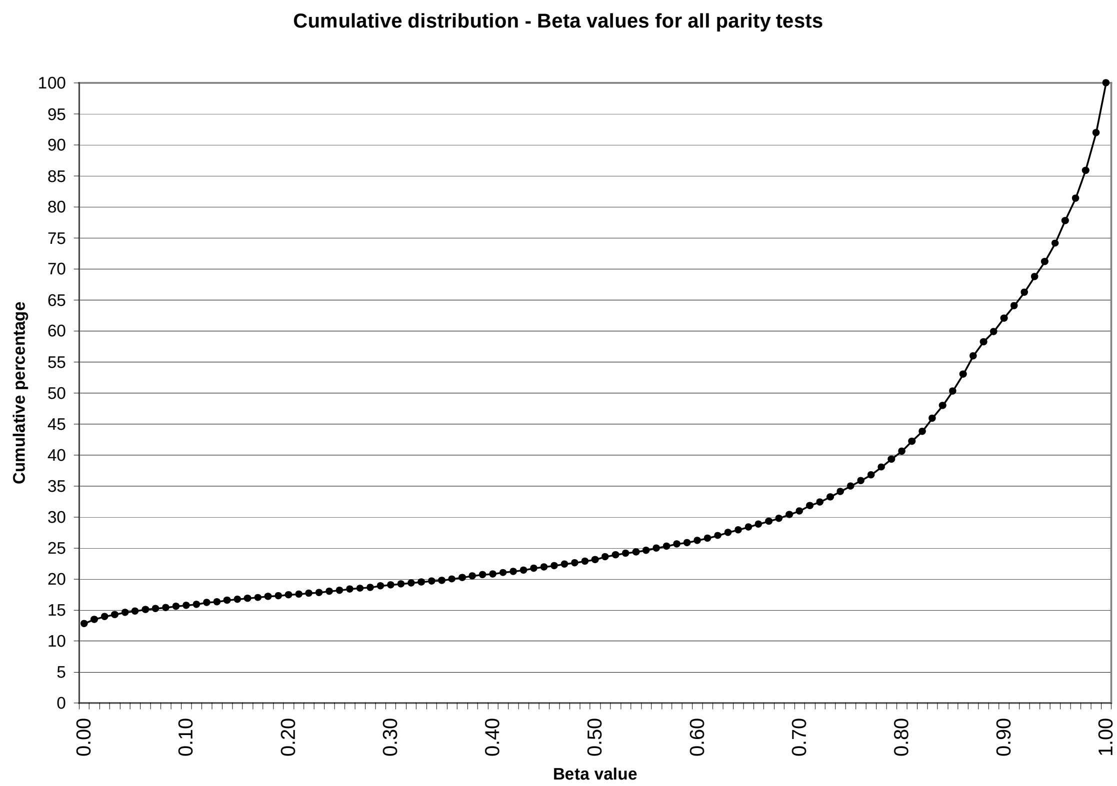
Table 2 presents beta values for parity measures by measurement type for Pacific's performance in July through September, 2000. All beta calculations are based on a 0.10 critical alpha and an alternative hypothesis of 50 percent worse performance (150%) for CLECs.

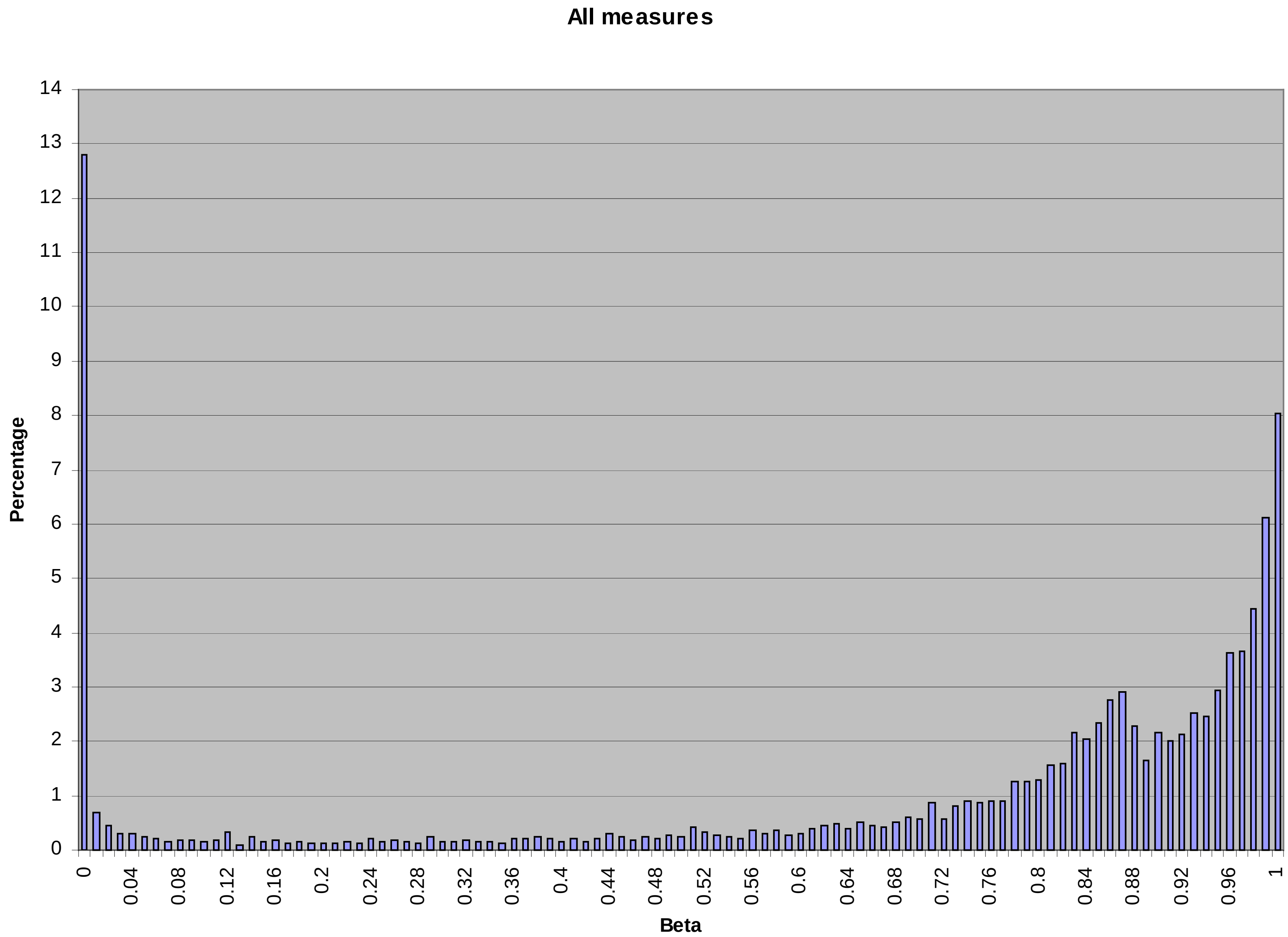
Table 2

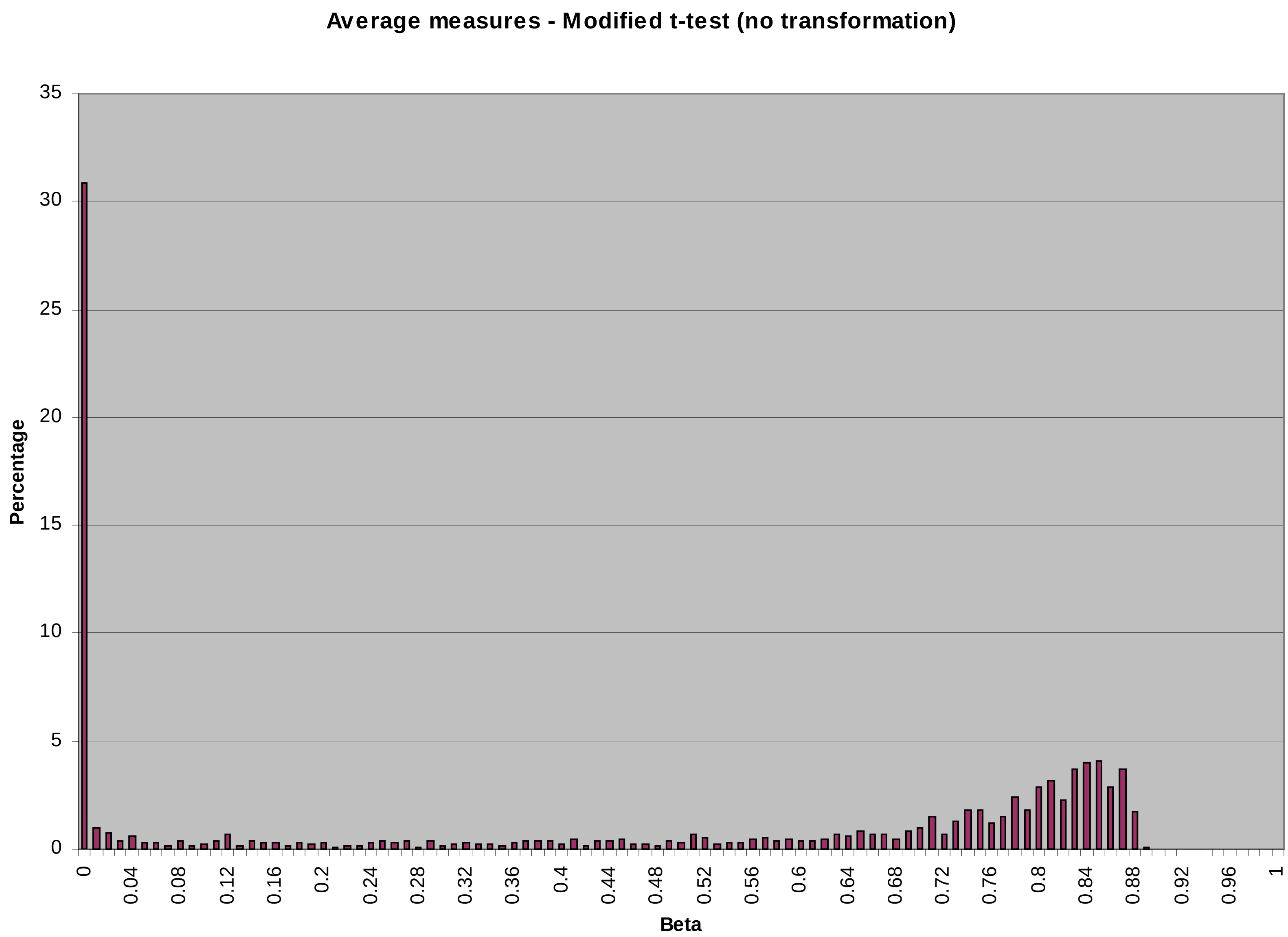
Average beta value by parity test type - Pacific performance July through August, 2000					
	All	Average	Percent (Hi)	Percent (Lo)	Rate
N	9909	2768	928	3558	2655
Percentage	100%	28%	9%	36%	27%
Mean	**0.70**	**0.45**	**0.42**	**0.87**	**0.83**
Median	**0.85**	**0.58**	**0.36**	**0.94**	**0.92**
SD	0.35	0.37	0.41	0.20	0.24
Skewness	-1.15	-0.20	0.18	-2.52	-2.18
Kurtosis	-0.24	-1.76	-1.75	6.58	4.16
Minimum	0.00	0.00	0.00	0.00	0.00
Maximum	1.00	0.89	1.00	1.00	1.00

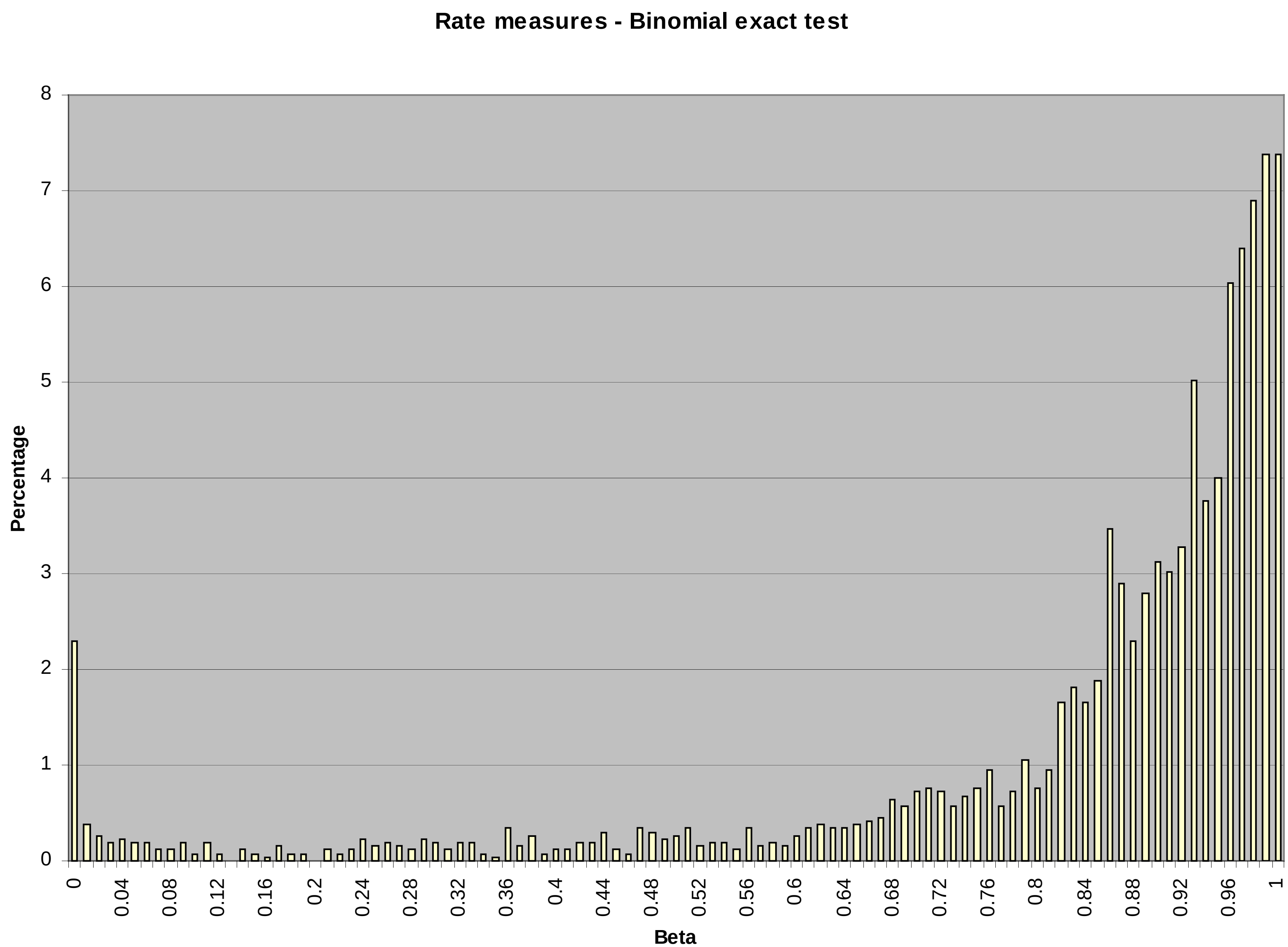
Attachment 1 presents the beta value cumulative distribution for all parity measures as presented in Table 2. For example, about 16 percent of all CLEC submeasure parity test results have beta values of 0.10 or less. In other words, when Type I error rate is held to 0.10 or less for all results, only 16 percent of all parity test results will have a Type II error rate of 0.10 or less.

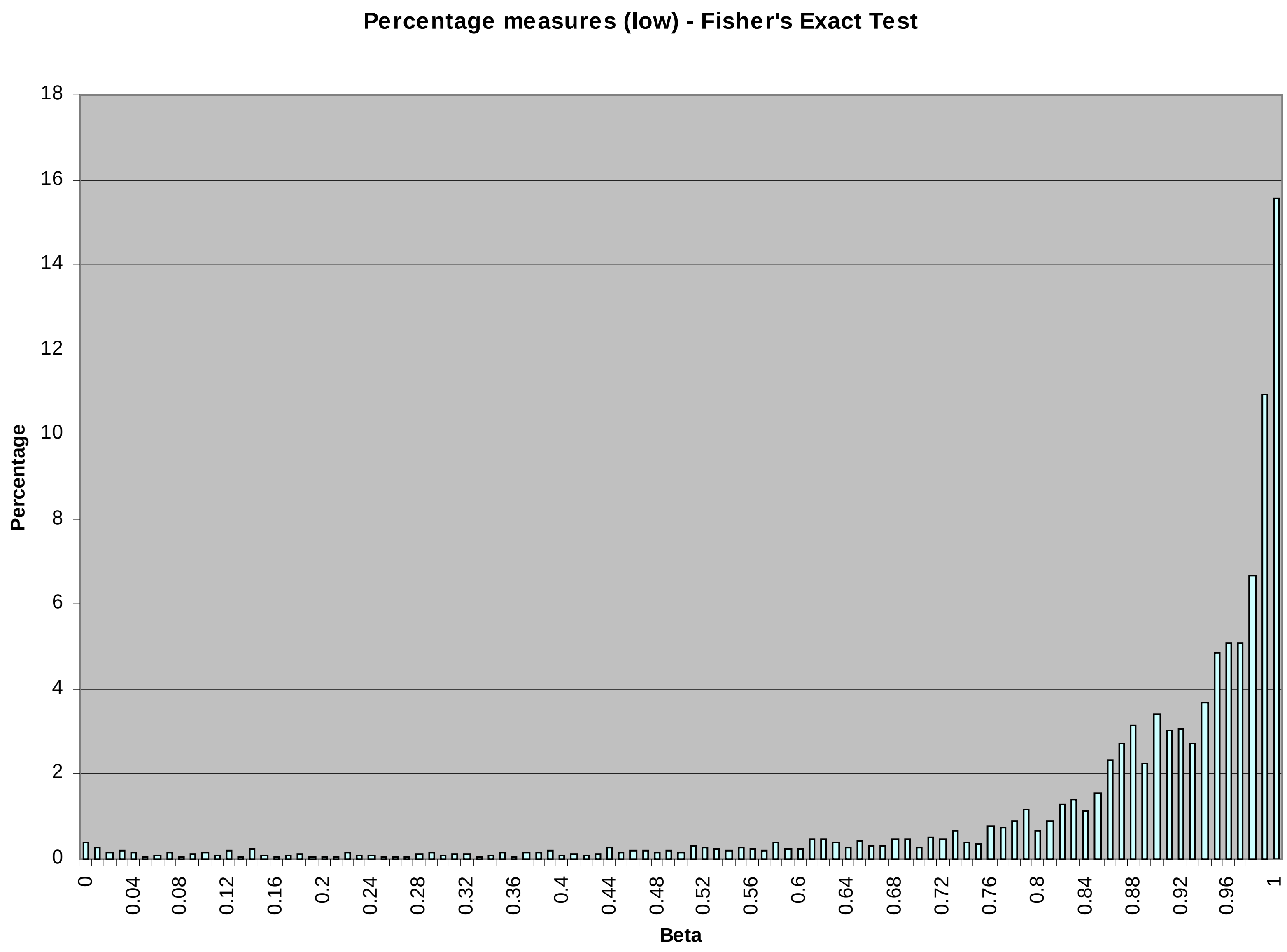
Attachment 2 presents the beta value frequency distributions for all parity measures combined and each measure type as presented in Table 2. For example, for all parity measures, two percent of beta values equal 0.84 (page 1).

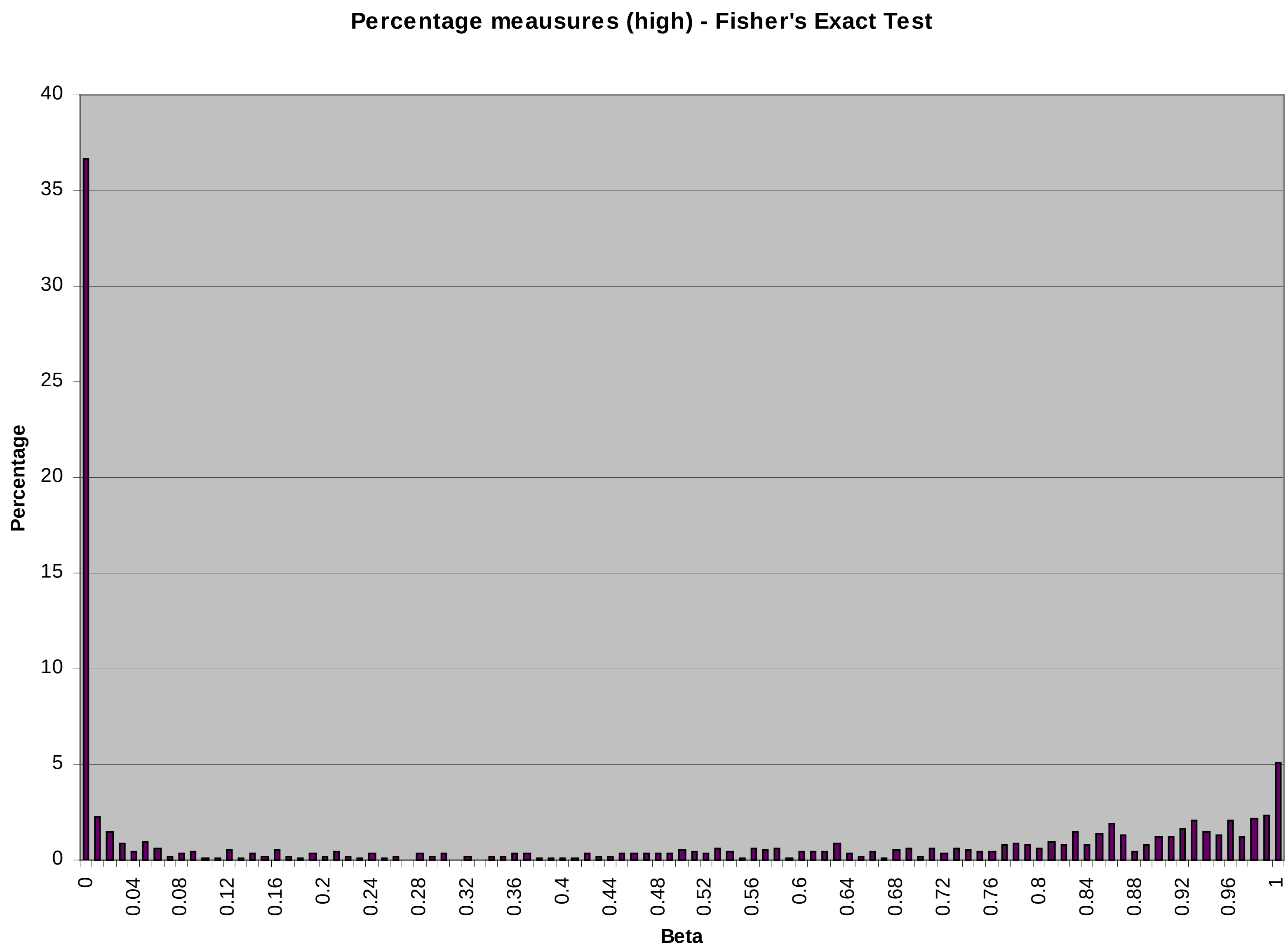

Cumulative distribution - Beta values for all parity tests
Cumulative percentage
Beta value
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00


All measures
Percentage
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
0.44
0.48
0.52
0.56
0.6
0.64
0.68
0.72
0.76
0.8
0.84
0.88
0.92
0.96
1
Beta


Average measures - Modified t-test (no transformation)
Percentage
35
30
25
20
15
10
5
0
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
0.44
0.48
0.52
0.56
0.6
0.64
0.68
0.72
0.76
0.8
0.84
0.88
0.92
0.96
1
Beta


Rate measures - Binomial exact test
Percentage
Beta
0
1
2
3
4
5
6
7
8
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
0.44
0.48
0.52
0.56
0.6
0.64
0.68
0.72
0.76
0.8
0.84
0.88
0.92
0.96
1


Percentage measures (low) - Fisher's Exact Test
Percentage
Beta
0
2
4
6
8
10
12
14
16
18
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
0.44
0.48
0.52
0.56
0.6
0.64
0.68
0.72
0.76
0.8
0.84
0.88
0.92
0.96
1


Percentage meausures (high) - Fisher's Exact Test
Percentage
40
35
30
25
20
15
10
5
0
0
0.04
0.08
0.12
0.16
0.2
0.24
0.28
0.32
0.36
0.4
0.44
0.48
0.52
0.56
0.6
0.64
0.68
0.72
0.76
0.8
0.84
0.88
0.92
0.96
1
Beta

Appendix G

Balancing Alpha and Beta Error

Balancing Alpha and Beta Error

This appendix documents staff's efforts to balance alpha and beta error levels for performance result assessment. To calculate the "balance point" for alpha and beta error staff adapted a balancing formula presented in Das (1994). Staff treated this formula as a "equal error" formula by assuming equal consequences for the two types of error. The formula was also adapted by including the "N" for both ILEC and CLEC samples. The formula as used was:

$$Z_\beta = ((H_0 - H_{alt}) / SD_i * sqrt(1/N_c + 1/N_i)) / 2$$

Where:
H_0 = Null hypothesis (ILEC mean)
H_{alt} = Alternate hypothesis
SD_I = ILEC standard deviation
N_c = CLEC sample size
N_I = ILEC sample size

Staff analyzed Pacific's May, 2000, performance results to estimate the effects of setting critical alpha levels equal to beta error for each result. An alternate hypothesis of 50-percent worse performance was assumed for the calculations. In other words, staff estimated the critical alpha level that would result in equal error (beta) in detecting performance at least 50% worse for the CLEC as for the ILEC. On the average, alpha balanced with beta at a value of 0.33. In other words, if alpha error was held to a maximum of 33 percent, beta error would also be 33 percent. Table 1 presents the summary statistics.

Table 1

	Alpha balanced with beta
N	3481
Mean	0.33
Median	0.41
Minimum	0
Maximum	0.5

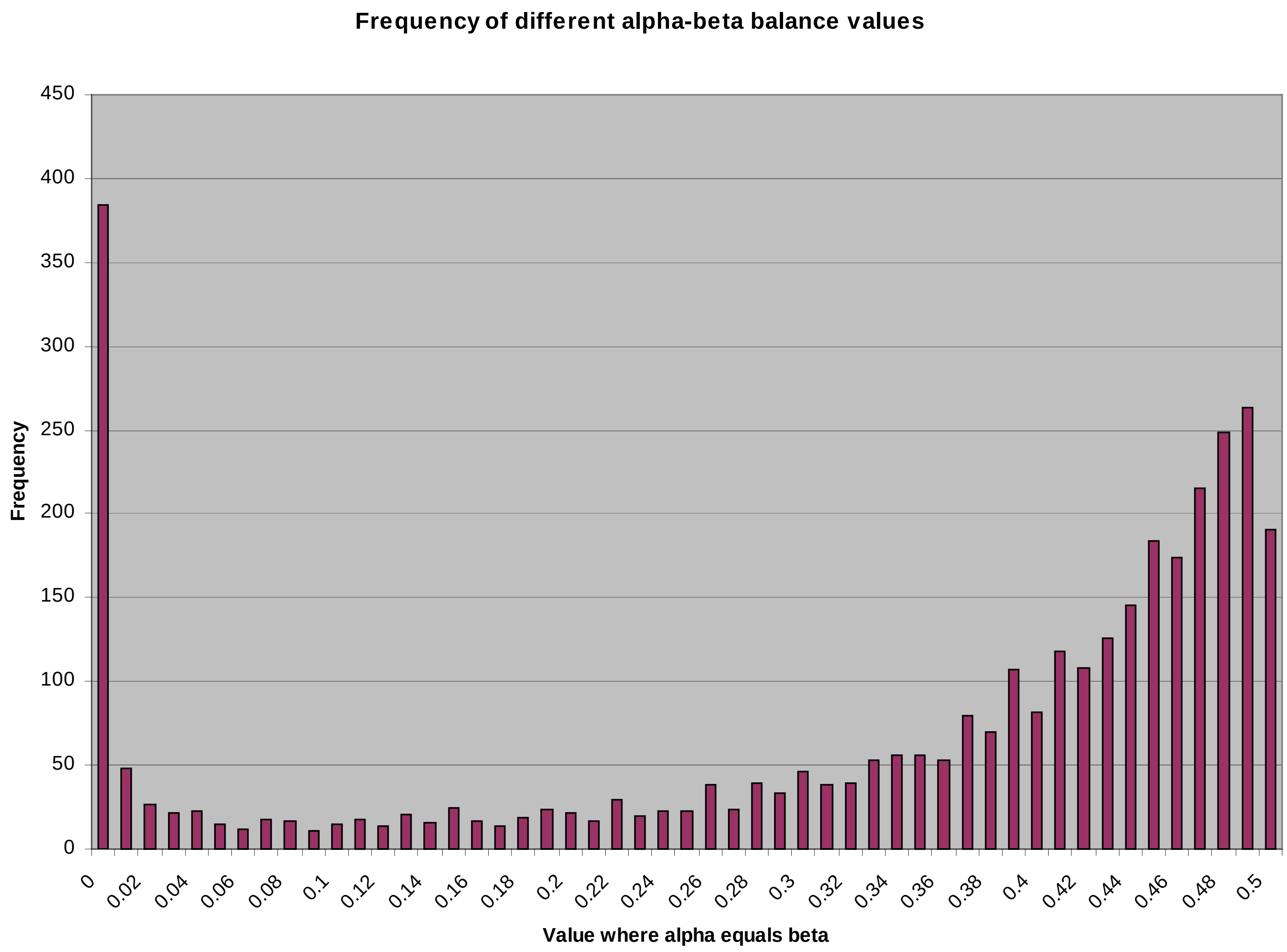
Attachment 1 presents a frequency distribution of the balancing values.

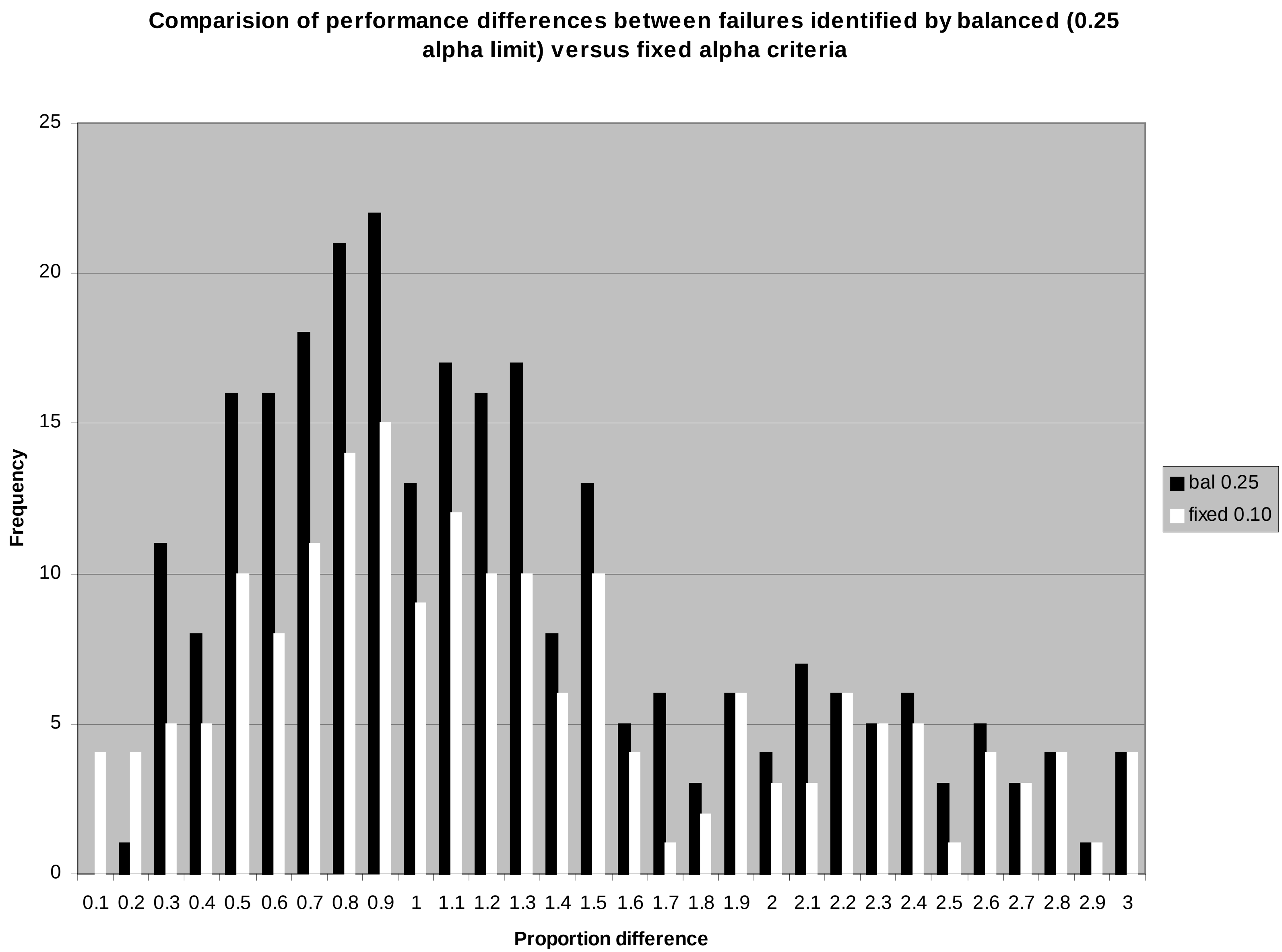
Staff also calculated the resulting error rates with an alpha error rate "ceiling." Table 2 presents summary statistics for those calculations.

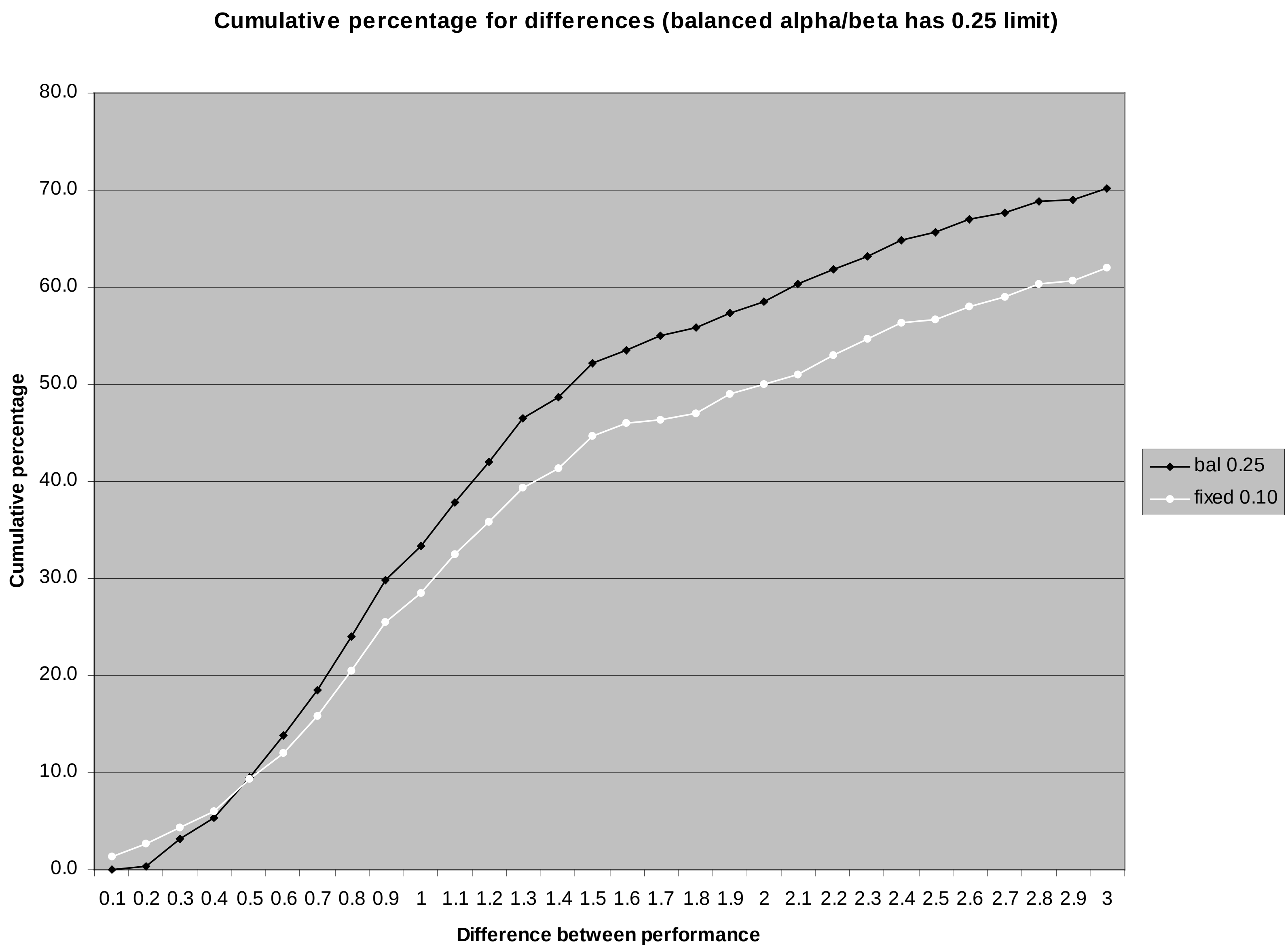
Table 2

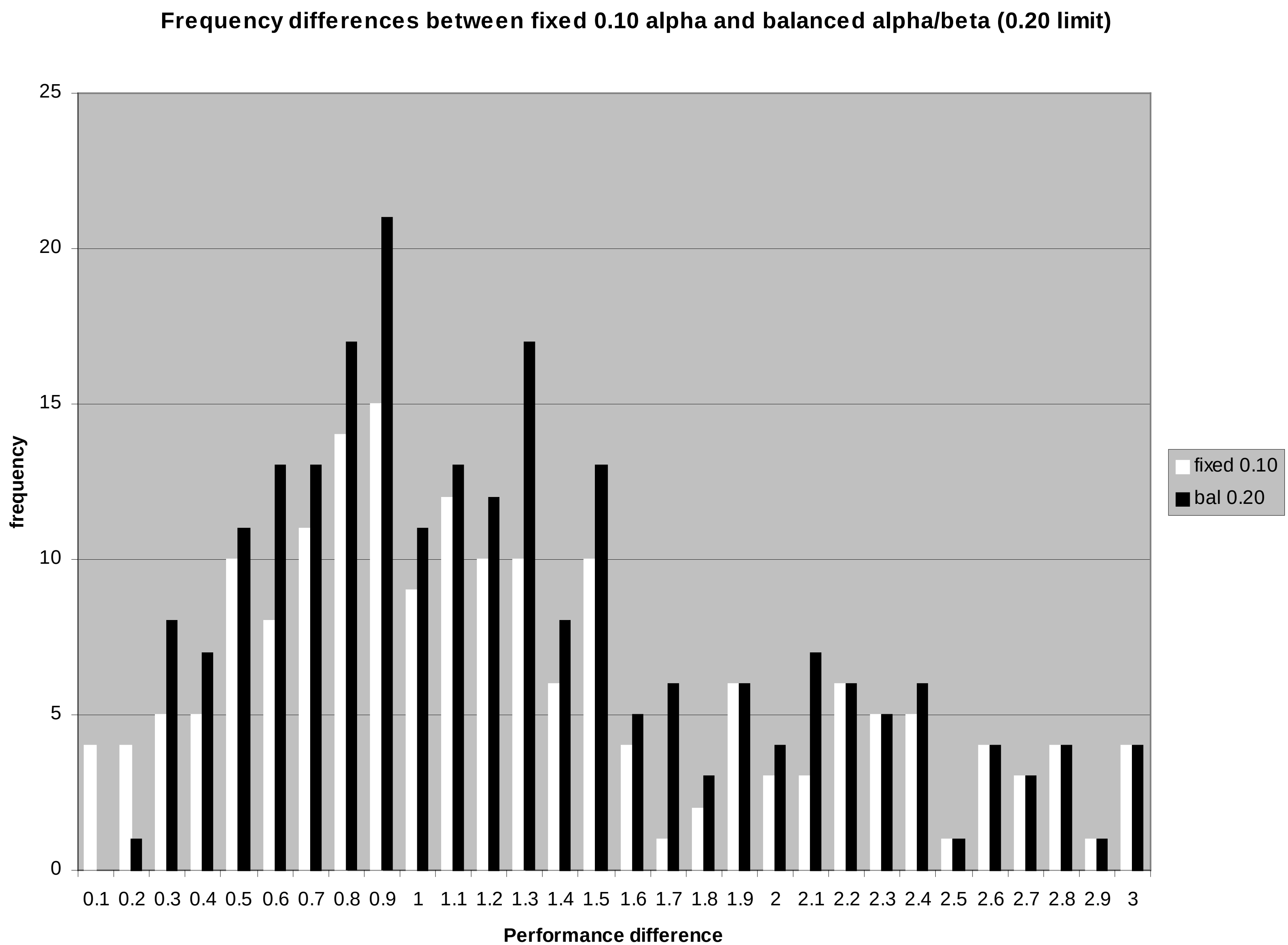
Critical alpha levels resulting from different alpha/beta balance limits			
	Alpha limit		
	0.33	0.25	0.2
N	1204	894	782
Mean	**0.131**	**0.0726**	**0.0499**
Median	**0.11**	**0.02**	**0.01**
Mode	0	0	0
Minimum	0	0	0
Maximum	0.33	0.25	0.2

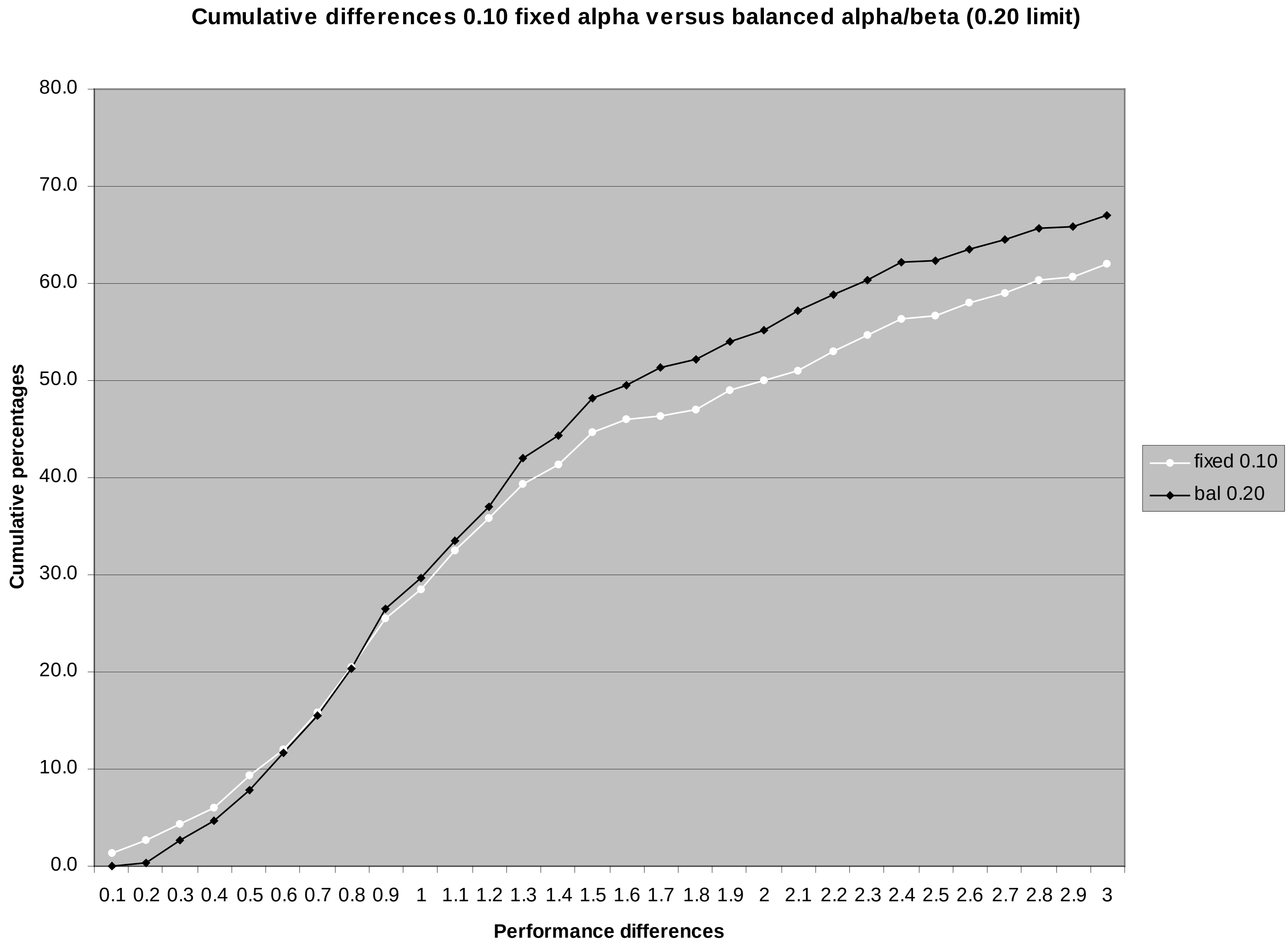
Staff also examined the net effect on the size of the difference between ILEC and CLEC performance that would be identified as a performance failure. Theoretically, balancing alpha and beta should result in an increase in larger differences being detected and a decrease in smaller differences being detected. Attachment 2 shows that this in fact would occur. Limiting alpha to 0.25, for example, results in a lower proportion of failure identifications where performance to a CLEC is zero to 50 percent worse than ILEC performance to itself (Attachment 2, page 2), relative to a fixed 0.10 alpha criterion. Conversely, this limit results in a greater proportion of failure identifications where performance to a CLEC is *at least* 50 percent worse than ILEC performance to itself. These charts only display results up to the point where performance to a CLEC is three times worse than performance for the ILEC. At this point, however, there are no further differences between a fixed 0.10 alpha criterion and either the 0.20 or 0.25 alpha / beta balance limited criteria.


Frequency of different alpha-beta balance values
450
400
350
300
250
200
150
100
50
0
Frequency
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
0.22
0.24
0.26
0.28
0.3
0.32
0.34
0.36
0.38
0.4
0.42
0.44
0.46
0.48
0.5
Value where alpha equals beta


Comparision of performance differences between failures identified by balanced (0.25 alpha limit) versus fixed alpha criteria
Frequency
25
20
15
10
5
0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
Proportion difference
bal 0.25
fixed 0.10


Cumulative percentage for differences (balanced alpha/beta has 0.25 limit)
Cumulative percentage
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
Difference between performance
bal 0.25
fixed 0.10


Frequency differences between fixed 0.10 alpha and balanced alpha/beta (0.20 limit)
frequency
0
5
10
15
20
25
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
Performance difference
fixed 0.10
bal 0.20


Cumulative differences 0.10 fixed alpha versus balanced alpha/beta (0.20 limit)
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
Cumulative percentages
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3
Performance differences
fixed 0.10
bal 0.20

Appendix H

Pacific’s Proposed Aggregation Rules

Effects of Pacific's proposed aggregation rules

Month	Total number of parity results	Number of results with samples larger than 10		Number of results with samples smaller than 10	Number of small samples that aggregate to samples 5 or larger		Number of small samples that aggregate to samples smaller than 5	Number of industry aggregate samples 5 or larger		Number of industry aggregate samples smaller than 5 (discarded)	
January	3062	1719	56.1%	1343	1221	39.9%	122	36	1.2%	86	2.8%
February	3138	1795	57.2%	1343	1257	40.1%	86	30	1.0%	56	1.8%
March	3425	1950	56.9%	1475	1348	39.4%	127	45	1.3%	82	2.4%
Total	9625	5464	56.8%	4161	3826	39.8%	335	111	1.2%	224	2.3%

Note: This table presents aggregation rules results for parity measures only.

Appendix H, Attachment 1

SAMPLE SIZE AGREEMENT

Submitted to ALJ Reed on 4/25/00

RULES: Only applicable to sub-measures that would normally have small sample sizes for all CLECs, ie., the process etc., being measured isn't something that is generally ordered a lot in a month.

The following measures and sub-measures are not subject to minimum sample size. Data for the following will not be discarded, but rather incentives will apply, once incentives are ordered.

What is agreed to in this memo is subject to appropriate incentives review, when ordered.

Measure 30: Agreed to by Pacific Bell and GTEC

Measure 40: Agreed to by Pacific Bell and GTEC

Measure 41: Agreed to by Pacific Bell and GTEC

UNE Loop DS-3: (Disaggregated as an Service Group Type) Agreed to by Pacific Bell, GTEC checking, but no GTEC commitment yet.

UNE-Transport DS-1: (Disaggregated within UNE-Transport) Agreed to by Pacific Bell, GTEC checking, but no GTEC commitment yet.

UNE-Transport DS-3: (Disaggregated within UNE-Transport) Agreed to by Pacific Bell, GTEC checking, but no GTEC commitment yet.

Interconnection Trunks: Agreed to by Pacific Bell and GTEC

Note: OC level services will also be added to this list if agreed to as a service group type as part of the JPSA performance measurements.

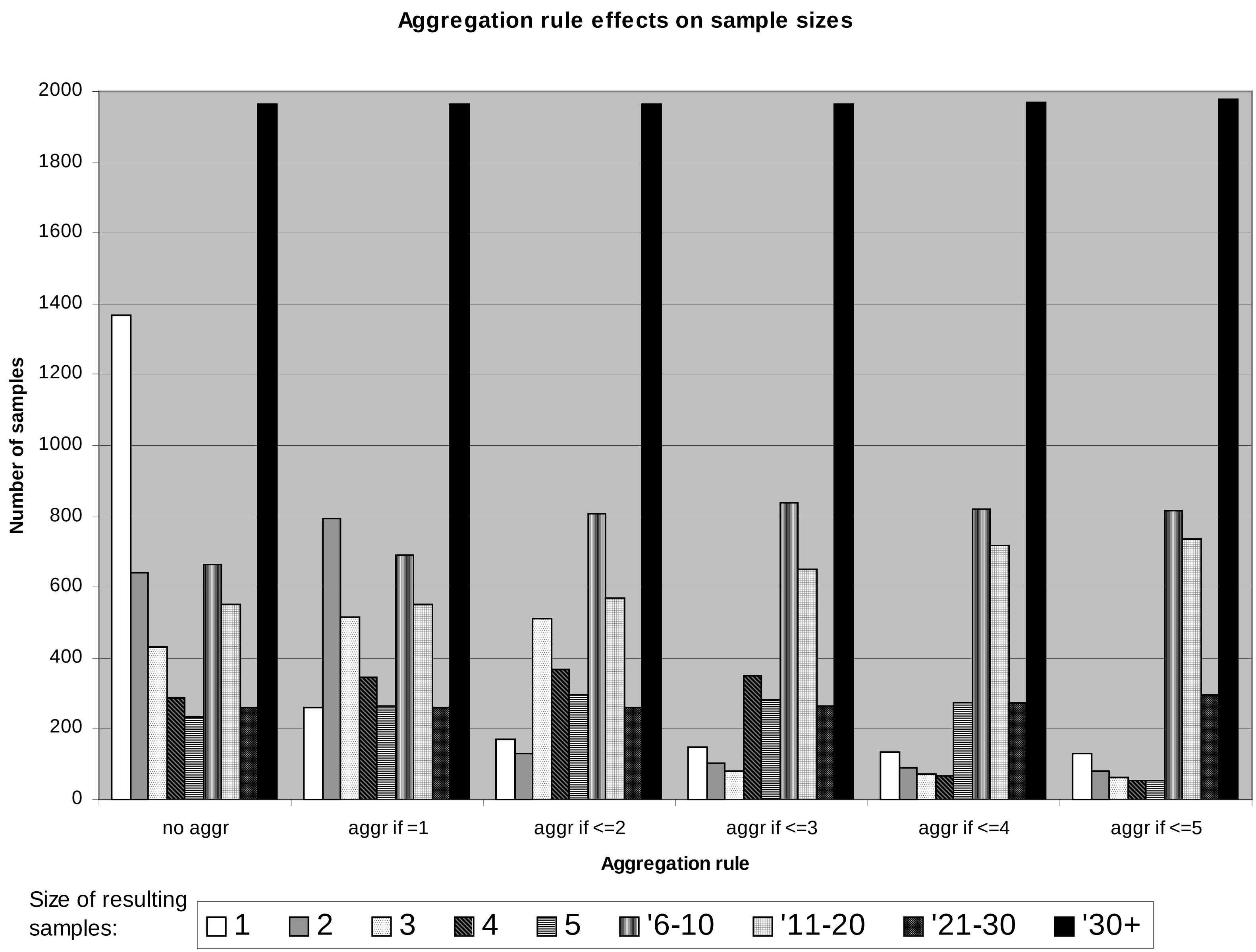
Appendix I

Implemented Aggregation Rule Results

Aggregation Rule Proposal Results

This appendix illustrates the effects of the aggregation rules implemented by this Decision. The attached chart shows how the frequency of small samples is diminished as increasingly larger sample sizes are subject to aggregation. For example, all possible aggregations of sample sizes of one reduces the number of sample sizes of one from nearly 1400 samples (see the white bar for the bar cluster labeled “no aggr”) to approximately 250 samples (see the white bar for the bar cluster labeled “aggr if = 1”). The bar cluster labeled “aggr if <= 4” represent the final results of the implemented aggregation rules.

These aggregation rules avoid some of the potential pitfalls of the previously proposed rules as discussed in the body of the Decision. First, small sample results are only aggregated with like small sample results, thus minimizing the likelihood that larger sample results will “mask” the small sample results. Second, while small sample failures may still cause non-failing samples to fail when aggregated, the incentive phase of the proceeding can address the problem of potential spurious allocation of incentive payments. Third, unnecessary aggregation is minimized. No small samples are aggregated with large samples as only like-size samples are aggregated. Fourth, since only like-sized small samples are aggregated, there is no ambiguity about which results, aggregated or non-aggregated, should determine results.


Aggregation rule effects on sample sizes
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Number of samples
no aggr
aggr if =1
aggr if <=2
aggr if <=3
aggr if <=4
aggr if <=5
Aggregation rule
Size of resulting samples:
1
2
3
4
5
'6-10
'11-20
'21-30
'30+