BELLSOUTH TELECOMMUNICATIONS, INC.

DIRECT TESTIMONY OF EDWARD J. MULROW, PH.D.

BEFORE THE FLORIDA PUBLIC SERVICE COMMISSION

DOCKET NO. 000121-TP

MARCH 21, 2001

Q. PLEASE STATE YOUR NAME, AND BUSINESS NAME AND ADDRESS.

A. My name is Edward J. Mulrow. I am employed by Ernst & Young LLP as a Senior Manager in the Quantitative Economics and Statistics Group. I have been retained by BellSouth as a statistical advisor. My business address is 1225 Connecticut Ave., NW, Washington, DC 20036.

Q. ARE YOU THE SAME EDWARD J. MULROW THAT FILED DIRECT TESTIMONY IN THIS DOCKET?

A. Yes. I filed direct testimony in this docket on March 1, 2001.

Q. WHAT IS THE PURPOSE OF YOUR REBUTTAL TESTIMONY?

A. The purpose of my rebuttal testimony is to respond to portions of the direct testimonies of Dr. Robert M. Bell representing the ALEC Coalition, and Dr. George S. Ford representing Z-Tel Communications. In responding to the direct testimony of these witnesses, I address the following issues:

- The appropriate statistical methodology for making performance measure parity comparisons.
- Dr. Bell's analysis of the impact of "delta."
- The use of a floor for the balancing critical value.

1. The appropriate statistical methodology for making performance measure parity comparisons.

Q. THE ALEC COALITION, REPRESENTED BY DR. ROBERT BELL, PROPOSES THAT THE FLORIDA COMMISION ORDER THE MODIFIED Z AS A COMPONENT OF THE STATISTICAL METHODOLOGY. PLEASE RESPOND.

A. As I said in my direct testimony, the appropriate methodology to use in situations where transaction level data is available and a BellSouth retail analog exists is the Truncated Z with Error Probability Balancing. This methodology is described in the Louisiana PSC "statistician's report" which is attached to my direct testimony as Exhibit EJM-1. One of the more interesting things about Dr. Bell's position is that it was another AT&T witness, now retired, who basically created the truncated Z formulas that BellSouth is now offering. Indeed, as I mentioned in my direct testimony, the methodology was developed in a joint effort between AT&T's statistical expert Dr. Colin Mallows (the AT&T witness who is now retired), and the Ernst & Young statistical team. I find it difficult to understand how AT&T and BellSouth could have expended such effort to reach a methodology that was satisfactory to the experts representing each party, only to have AT&T seemingly walk

away from that methodology.

Q. CAN YOU EXPLAIN THE FUNDAMENTAL DIFFERENCES BETWEEN WHAT DR. BELL PROPOSES AND WHAT BELLSOUTH IS PROPOSING?

A. As with many things involving statistics, the explanation is a bit complex, but I will try to explain in as clear a fashion as possible. You will recall that in my direct testimony, I discussed how BellSouth's methodology took various measures down to what I called the individual "cell" level. The purpose of creating "cells" was to break each comparison down it its most basic components, so that we could be relatively sure that we were comparing "apples-to-apples." For instance one of the cells would be a new residential provisioning order that is non-dispatched with less than 10 circuits that occurred in the first part of the month in a particular wire center. We would compare BellSouth's transactions that met those criteria as well as the ALEC transactions that met the criteria. We would determine the mean for both samples and would calculate a modified Z statistic for that cell. After doing this, we would roll this cell up with other cells related to plain old telephone service and would essentially aggregate all of the individual modified Z statistics into a single statistic. As I explained in my direct, when we rolled these individual statistics into a single statistic, we assign a value of zero to all of the statistics that have a positive value, so that we do not mask the impact of any negative values. This changing of positive values to zero is why we call the resulting statistic a truncated Z statistic.

On the other hand, Dr. Bell's process essentially stops with the calculation of the modified Z statistic for each of his "sub-measures" which can generally be thought of as being conceptually the same as BelllSouth's cells, except that Dr. Bell has disaggregated his sub-measures in a different way, and has not disaggregated them to the levels that BellSouth has proposed. For instance, and this is taking an extreme example, Dr. Bell's proposal would essentially have BellSouth stop at the cell level discussed above, and make the final comparison about whether parity is being provided right there.

Q. SO, BASED ON THIS DISCRIPTION, WHAT IS CONCEPTUALLY WRONG WITH DR. BELL'S APPROACH?

A. I can explain this most clearly by looking again at what BellSouth has done. Lets assume that there were 2000 "cells" associated with the provisioning of plain old telephone service. If we looked at the individual cells, we might find 75 that revealed an apparent discrepancy between BellSouth's performance and that provided to the ALEC. Based on these failures, relying solely on the modified Z statistic, BellSouth would be expected to pay a penalty. The problem is that we know that we are going to get some Type 1 errors in a statistical analysis like this. For instance, if we were willing to accept that 5% of the observations were going to be Type I errors, you would expect to see 100 failures. Viewed in this light, 75 failures would be well within the expected parameters. The point is, if you looked at the individual "cells" you would conclude there was a problem, but when you look at the whole picture

you see that there is not. Dr. Bell's approach, relying solely on the modified Z statistic for individual "cells" or "sub-measures" doesn't allow this to happen.

Q. IS THE IDEA OF "AGGREGATING" MANY STATISTICAL RESULTS TO MAKE AN OVERALL DETERMINATION OF PARITY SOMETHING THAT WAS DEVELOPED SPECIFICALLY FOR THE LOUISIANA STATISTICIAN'S REPORT?

A. No. The concept has been around for quite some time. It is sometimes referred to as a "multiple testing problem."

Q. WILL YOU BRIEFLY EXPLAIN WHAT A "MULTIPLE TESTING PROBLEM" IS?

A. Certainly. As I pointed out in my example above, whenever one performs numerous statistical tests at one time, it needs to be recognized that some of the tests will provide results indicating a problem even when there is no problem. By looking at all the results in a more global way, one can determine if the "failed" tests that are observed represent a true problem, or just random chance.

Q. IS BELLSOUTH THE ONLY COMPANY SUGGESTING THAT THIS FORM OF AGGREGATING SHOULD BE DONE?

A. No. The four states where the FCC has granted an RBOC the right to market long distance services have performance comparison plans that aggregate the

results of many comparisons into an overall result that determines parity/disparity.

In New York, Verizon uses a weighted average of performance scores to make parity judgments. In Texas, Oklahoma, and Kansas, Southwestern Bell uses the "K-value" method. This "K-value" methodology was described by AT&T's Dr. Mallows in an affidavit to the FCC in May 1998[1]. Thus, both of the methods of aggregation that AT&T's expert has suggested have been adopted by former Bell Companies for use in their performance plans. AT&T however, appears reluctant to accept either of these methodologies.

Q. WHY IS THE TRUNCATED Z STATISTIC A BETTER AGGREGATE STATISTIC THAN THE OTHERS THAT ARE IN USE, SAY IN TEXAS?

A. As I explained in my direct testimony, the truncated Z statistic was created so that it possesses five important properties.

1. It is a single, overall index on a standard scale; that is, you can use a probability distribution to make judgments.
2. If transaction counts for BellSouth and the ALEC across comparison cells (classifications) are exactly proportional, the aggregate index should be very nearly the same as if we had not disaggregated. This means that

[1] Affidavit of Dr. Colin L. Mallows before the Federal Communications Commissi
1998.

granular disaggregation I have discussed really wasn't necessary, you will still get the same results.

3. The contribution of each cell depends on the number of transactions in the cell.
4. As far as possible, systematic discriminatory performance in some cells is not masked by good performance in other cells.
5. The final result does not depend critically on minor details in the data; that is, small changes in transaction values only induce small changes in the final result.

In addition to these important properties, the error probabilities of the truncated Z test can be balanced. The other statistics used when aggregating results do not meet all of the criteria that I have outlined above, but I would note that any of them would be better than what Dr. Bell is proposing, which is that no aggregation be done at all.

Q. ARE THERE OTHER DIFFERENCES BETWEEN WHAT DR. BELL HAS PROPOSED AND THE BELLSOUTH METHODOLOGY?

A. Yes, and there is one in particular that should be considered. The formulae that Dr. Bell proposes for testing proportion and rate measures do not easily lend themselves to balancing Type I and II error probabilities. This creates a methodological inconsistency between the test Z statistics he recommends and

the balancing critical value. I will discuss this in more detail later in my testimony. In order to explain the root of the problem, however, I need to tell you something more about statistics. What we have been discussing in the examples above is a comparison of "means," that is, we take the average of the BellSouth transactions in the "cell" and compare that "average" or "mean" to the comparable "mean" of the ALEC transactions. Not all observations lend themselves to the calculation of "means," however. For instance, consider "missed appointments." With "missed appointments" you are looking at the percentage of the total number of scheduled appointments that were missed. As a result, you end up with a proportion, such as a tenth of a percent or 5 percent or whatever figure is appropriate. You do not have a mean per se. Another example is what we call a "rate" such as the "customer trouble report rate", where you are looking at the number of troubles BellSouth or the ALEC has per the number of available lines. Unlike the "proportional" measures described above, which would always have to be less than 1, the measurement of a "rate" could exceed 100 percent. For instance, if you had ten access lines and 12 reported troubles (that is some lines have more than a single trouble during the reporting period) you can get more than a figure of 100 percent. Again, these two special categories are simply different measures than the "means" calculation that we have been talking about.

The root of the problem is that Dr. Bell uses the modified Z concept irrespective of whether the measure is one based on "means," "proportions," or

"rates." The difficulty from a statistical perspective is that the concept the modified Z statistic is based on should not be applied across the board to all measure types. Specifically, the basis for the modified Z statistic is that you take the difference between the two "means" in the particular "cell" or sub-measure, and divide the result by the standard deviation of BellSouth's mean. This is done to make the test sensitive to changes in the ALEC standard deviation (compared to the BellSouth standard deviation) that would be harmful to the ALEC. In other words, BellSouth could try to give the same average service to ALEC customers as to it own customers, but do so in a way that some ALEC customers receive longer completion times. For example, suppose that BellSouth always services its own customers in 2 days. BellSouth could service one-third of the ALEC customers in 1 day, one-third in two days, and the remaining third in 3 days. On the average, the ALEC service times are the same as BellSouth's, but one-third of the ALEC customers received service that was "below" average. Dividing the difference between the means by only BellSouth's standard deviation avoids masking this problem.

The same situations cannot occur for "proportion" or "rate" measures. In the case of a proportion, such as "missed appointments" that is stated as a percentage of total appointments scheduled, you only have one parameter to consider, the proportionality. As a result, BellSouth cannot separately control the proportion value and the variability about that value.

Q. ARE THERE DIFFERENT FORMULAS THAT DR. BELL COULD HAVE USED TO ADDRESS THESE ISSUES?

A. Yes, but he did not do so, which is another reason why BellSouth's approach makes more sense.

II. Dr. Bell's analysis of the impact of 'delta."

Q. IN ORDER TO SUPPORT HIS CHOICE OF A "DELTA" VALUE OF 0.25, DR. BELL PROVIDES A TABLE SHOWING THE PERCENTAGE OF ALEC CUSTOMERS RECEIVING BAD SERVICE, BY BELLSOUTH PERCENT AND DELTA. CAN YOU COMMENT ON THIS TABLE?

A. Well, there are a couple of interesting points I can make. First, the methodology that Dr. Bell advocates in his testimony and exhibit is not the method that he used to calculate the numbers in the table. Second, the table does not accurately represent the way that BellSouth proposes to carry out balancing for proportion measures.

Q. WOULD YOU EXPLAIN YOUR FIRST POINT MORE FULLY?

A. The overall concepts for balancing error probabilities were first developed for mean performance measures. As previously described, these are measures that represent the average of a measured amount, for example the average time it takes to complete an order. In this case, "delta" represents the difference between the ILEC and ALEC averages in terms of an ILEC standard deviation.

When solving for the balancing critical value, it turns out to be mathematically convenient to define the alternative hypothesis this way, given the form of the modified z statistic for a mean measure.

A proportion measure, on the other hand, measures the fraction of transactions that possess a certain quality or attribute out of all transactions. For example, percent missed installations measures the fraction of all provisioning orders in a month where service was not completed on or before the assigned due date. Often in statistics, methods that are worked out for means can be used on proportions because a proportion can be considered as a special type of mean. However, a proportion is a fraction of the whole, so it can only be a number between 0 and 1 (or equivalently between 0 percent and 100 percent).

Now, if we want to describe "delta" for a proportion measure as the difference between the ILEC and ALEC proportions in terms of an ILEC standard deviation we have to be careful. The mathematical convenience present in the mean measure case is not present with a proportion measure. Thus a different method is needed.

Q. WHAT METHODS ARE AVAILABLE FOR BALANCING A PROPORTION MEASURE?

A. We have identified two ways to approach balancing for proportion measures. One way is to transform the proportion using the arcsine square root transformation. This is what Dr. Bell used to create Table 1 on page 13 of his

direct testimony. The other way is to use a concept called the "odds" ratio.

Q. WHAT IS THE REASONING BEHIND USING THE ARCSINE SQUARE ROOT METHOD?

A. Because "delta" for a proportion measure cannot be defined using a straightforward analogy with the definition for a mean measure, a transformation is used. This allows us to use the same formula to compute the balancing critical value as was used in the mean measure case. However, two problems arise: 1) the interpretation of "delta" is related to the transformed measure, and 2) the z statistic that is used in the test should also use the transformed measurement.

Q. WHAT IS THE INTERPRETATION OF "DELTA" WHEN THE ARCSINE SQUARE ROOT TRANSFORMATION IS USED?

A. "Delta" becomes twice the difference of the transformed ILEC proportion with the transformed ALEC proportion. It is no longer the difference between the performance measures in terms of an ILEC standard deviation.

Q. YOU STATED THAT THE Z STATISTIC USED IN THE TEST SHOULD ALSO USE THE TRANSFORMED MEASUREMENTS. COULD YOU EXPLAIN THIS?

A. Yes. In order to arrive at the same balancing critical value fo

proportion measure as that of a mean measure, you must redefine the basic Z statistic. When using the arcsine square root transformation the Z statistic should be

$$Z = \frac{2\left(\arcsin\left(\sqrt{\hat{p}_{ILEC}}\right) - \arcsin\left(\sqrt{\hat{p}_{ALEC}}\right)\right)}{\sqrt{\frac{1}{n_{ILEC}} + \frac{1}{n_{CLEC}}}}.$$

Q. IS THIS WHAT DR. BELL IS RECOMMENDING TO USE FOR PROPORTION MEASURES?

A. No. The formula he recommends is given in Exhibit RMB-1, page 14, of his direct testimony. This formula is a direct analog of the mean measure formula, but as I have already explained, we need to be cautious in directly applying mean measure formulae to other types of measures when we are using a balancing methodology.

Q. WHAT ARE THE CONSEQUENSES OF USING A BALANCING CRITICAL VALUE BASED ON THE ARCSINE SQUARE ROOT TRANSFORMATION WITH A Z STATISTIC THAT IS NOT BASED ON THE TRANSFORMATION?

A. There are many scenarios where the use of the wrong type of Z statistic would find BellSouth to be out of parity when the use of the proper Z statistic would find them in parity. Consider a simple example. Let's suppose that there are

1000 BellSouth provisioning orders, and that BellSouth "missed" 214 of the appointments, that is, the orders where not completed on or before the due date. Thus, BellSouth "missed" 21.4 percent of their orders. For the same time period, suppose there were 30 comparable ALEC provisioning orders, and that 8 of these were "missed." So, BellSouth "missed" 26.7 percent of the ALEC's orders. Now the balancing critical value based on the arcsine square root transformation and the "delta" of 0.25 that Dr. Bell uses is -0.675. If we use the modified Z formula given in Dr. Bell's direct testimony, we will get a Z score of -0.693. Since this is less than the critical value (further from zero on the negative side), we would conclude that there is a lack of parity, and BellSouth would pay a penalty. On the other hand, if we use the Z formula given above, which is based on the arcsine square root transformation, we get a Z value of –0.666. In this case, we would say that BellSouth is compliant, and there would be not a penalty assessment.

Q. SO YOU ARE SAYING THAT THE BASIC METHODOLOGY THAT IS USED TO CALCULATE THE BALANCING CRITICAL VALUE NEEDS TO BE MATCHED WITH THE SAME BASIC METHODOLOGY THAT IS USE TO CALCULATE THE Z TEST STATISTIC.

A. Yes, and there appears to be an inconsistency in what Dr. Bell is recommending for proportion measures as well as rate measures.

Q. YOU SAID THERE IS ANOTHER METHOD FOR BALANCING PROPORTION MEASURES THAT IS BASED ON THE "ODDS" RATIO.

WHAT IS AN "ODDS" RATIO?

A. The "odds" ratio is what BellSouth has used when the information in the "cells" involves proportions, which I have been discussing, rather than "means." The "odds" methodology is relatively straightforward. First we need to define the odds of an event such as a missed installation occurring. Odds are the ratio of the probability of an event occurring to the probability that the event won't occur. So, if BellSouth "missed" 21.4 percent of the installations to their own customers, then the odds of a customer experiencing a "miss" is found by dividing the probability of a "miss," 0.214, by the probability of an "on-time" installation, 0.786 (= 1 – 0.214). This gives the odds of a "miss" as 0.276. In odds terminology, we might say that the odds of a BellSouth customer experiencing a "miss" are approximately 1 to 2.6.

The odds ratio for "missed" provisioning installations is the ALEC customer's odds of a "miss" divided by the BellSouth customer's odds of a "miss." When this odds ratio is one or less, BellSouth is delivering parity or better service to the ALEC's customers. When this odds ratio is greater than one, then BellSouth is not necessarily delivering parity service. Under a balancing approach, we need to determine an odds ratio greater than one to use for the balancing alternative hypothesis.

Q. IS THE ODDS RATIO EASIER TO INTERPRET THAN THE ARCSINE SQUARE ROOT METHOD?

A. Not necessarily. Many people have trouble interpreting odds, and relating the value back to the probability of an event occurring. However, the interpretation in terms of odds is straightforward. If the odds ratio for "missed' installations is set at 3, then we know that an ALEC customer's odds of a "miss" is three times greater than that of a BellSouth customer. We would still need a table, such as Dr. Bell's Table 1, to interpret the actual difference in the performance. I want to say, however, that setting the "odds" ratio at 3, which is what the Louisiana Commission has done for Tier 1 measures, does not necessarily mean that the probability of actually having a disparity is that great.

Q. CAN YOU PROVIDE US WITH SUCH A TABLE?

A. Certainly. Figure 1 below will help one interpret the actual difference between the BellSouth proportion and the ALEC proportion for a given "odds" ratio. The table shows the percentage of the time an ALEC customer will experience a miss by the BellSouth percentage "missed," for two values of the odds ratio: 2 and 3.

Figure 1
ALEC Percentage of "Missed" Installations
By BST Percentage and
The Odds Ratio of the Alternative Hypothesis

BST PERCENTAGE MISSED	Odds Ratio 2	3
1	2	3
5	10	14
10	18	25
20	33	43

We see from the first row of this table that for an alternative hypothesis with an odds ratio of 3, the ALEC percentage of "missed" installations is about 3 percent when the BST percentage is 1 percent. However, the ALEC percentage is about 43 percent when the BST percentage is 20 percent. So when the BST percentage is close to 0, the ALEC percentage is about 3 times larger at the balancing alternative hypothesis. As the BST percentage get larger, the ratio of the ALEC percentage to the BST percentage gets smaller; converging to 1 as the BST percentage approaches 100 percent.

Q. THIS SEEMS TO SUGGEST THAT IF BELLSOUTH HAS A MISS OF 20 PECENT, THAT A MISS OF UP TO 43 PERCENT WOULD BE ACCEPTABLE FOR THE ALECS. IS THIS CORRECT?

A. No, that misses the point completely and that is what is wrong, in large measure with Dr. Ford's analysis, which I will discuss in more detail below. However, to put point on this, with numbers like that, with a very small sample size the methodology would show BellSouth out of parity almost 60 percent of the time and as the sample size approached a thousand transactions for BellSouth and only fifty for the ALEC, the probability that parity will not be concluded approaches 100 percent (see Table 3 below). I realize this is not intuitive, and I will discuss it more below, but it would be a mistake to conclude that the odds ratio balancing test allows the ALECs to experience significantly worse performance than BellSouth without detecting a failure to provide parity on BellSouth's part. I would also note that the same holds true for Dr. Bell's calculations using the arcsine square root method where he

shows a similar disparity. Once the sample size gets to the levels that I have just mentioned, the probability of finding a disparity at those levels approaches 100 percent.

Q. IF THE ODDS RATIO METHOD IS USED FOR DEFINING THE BALANCING CRITICAL VALUE, HOW DOES THAT EFFECT THE FORMULA THAT IS USED TO CALCULATE THE CRITICAL VALUE?

A. The balancing critical value for a proportion measure is based on a different formula than that of a mean measure when an odds ratio approach is used. The formula is more complicated than the mean measure formula, and it is given in Appendix C of the Louisiana "Statistician's Report" (Exhibit EJM-1 of my direct testimony).

Q. DOES THE Z STATISTIC USED TO COMPARE THE PERFORMANCE MEASURES NEED TO BE MODIFIED WHEN USING THE ODDS RATIO APPROACH?

A. I was able to derive the balancing critical value formula based on the odds ratio because it "fit in" with the method used to calculate the cell level Z statistic. This Z statistic that I refer to is given in Appendix A of the Louisiana Statistician's Report. As previously alluded to, it differs from the Z statistic given by Dr. Bell in his testimony.

Q. SO THE Z STATISTIC FOR PROPORTIONS PROFFERED BY DR. BELL

NEEDS TO BE MODIFIED, REGUARDLESS OF THE BALANCING APPROACH, IN ORDER TO HAVE THE BALANCING METHODOLOGY CONSISTENT WITH THE BASIC Z STATISTIC METHODOLGY.

A. Yes. I believe that we should try to be consistent with the Z statistic methodology when developing the methods for balancing. That's not to say that a balancing methodology cannot be worked out for the proportion Z statistic in Dr. Bell's testimony, but I think it would make a complex problem messier. Dr. Bell may also be able to show that a balancing critical value based on a method different from the one used to create the LCUG proportion Z statistic is a reasonable approximation under certain circumstances. The data that we have examined so far exhibit many different characteristics, so it is easy to find cases when the approximations break down. In fact all of the balancing methods break down when both BellSouth and ALEC transaction counts get very small. So, none of the methods we've looked at are perfect. I do believe that we should do our best to avoid problems that we can identify, and consistency between Z statistic methods and balancing methods helps.

Q. YOU SAID THAT YOU DID NOT THINK THAT DR. BELL'S TABLE 1 REPRESENTS THE WAY IN WHICH BELLSOUTH WILL CARRY OUT BALANCING FOR PROPORTION MEASURES. WILL YOU EXPLAIN THIS?

A. BellSouth has chosen to use the "odds" ratio approach to balancing. In fact, the Louisiana Public Service Commission has ordered BellSouth to use an

odds ratio of 3 for Tier I testing of proportion measures, and an odds ratio of 2 for Tier II testing. So Figure 1 above shows the impact of the choice of an "odds" ratio based on BellSouth's proportion measure balancing position.

Q. IS THERE ANY WAY TO TRANSLATE BETWEEN THE TWO METHODS?

A. Yes, the Louisiana "Statistician's Report" provides equations that can be used to translate between the two methods. Things are not that straightforward however. You must have an idea of what the BellSouth proportion is in order to translate between methods. For a proportion measure, we can determine what the largest "delta" value will be for a fixed odds ratio over the whole range of proportion values. For instance, with an odds ratio of 3, the largest value of "delta" based on the arcsine square root method is about 0.54. This occurs when the BellSouth percentage of "misses" is about 37 percent. For percentages smaller or larger than 37 percent, the equivalent delta for an odds ratio of 3 is smaller than 0.54. The equivalent delta gets very close to zero when the BellSouth percentage of "misses" is close to 0 or 100 percent.

III. The use of a floor for the balancing critical value.

Q. DR. FORD STATES IN HIS TESTIMONY THAT HE BELIEVES THERE IS A SERIOUS FLAW IN THE ERROR PROBABILITY BALANCING METHODOLOGY, AND THAT A LIMIT ON THE BALANCING CRITICAL VALUE NEEDS TO BE ESTABLISHED TO CORRECT THE FLAW. PLEASE RESPOND.

A. In reading through his arguments I sense that he is confusing hypothesis testing issues. The key issue that he is confused about is that there is a difference between the probability of a Type I error and the probability of detecting disparity. I also do not belief that Dr. Ford appreciates the problems imposed by the observational nature of a monthly performance incentive plan. I will briefly address these issues, and discuss an error in one of Dr. Ford's graphics.

When all of the statistical issues are properly understood and considered as a whole, I believe that there are no serious flaws in the balancing methodology. Therefore, there is no need for the "fix" that Dr. Ford suggests, namely, a floor on the balancing critical value.

Q. YOU SAY THAT DR. FORD IS CONFUSING THE PROBABILITY OF A TYPE I ERROR WITH THE PROBABILITY OF DETECTING DISPARITY. PLEASE EXPLAIN?

A. Dr. Ford makes several comments in his testimony that suggest that a statistical test with a small Type I error probability has very little power to detect discrimination. For instance, on page 21, lines 15 – 17, he states "At significance levels less than 0.0001 (assuming no more than 500 tests are conducted), balancing performs no function other than to make it nearly impossible to detect discrimination (i.e., reject the null hypothesis)." This is simple not true.

First, it needs to be understood that the significance level, i.e. the probability of a Type I error, is the probability of rejecting the null hypothesis (concluding that disparity exists) when, in fact, the null hypothesis is true (BellSouth is providing parity service). This is not the probability that the null hypothesis will be rejected when there is truly a certain amount of disparity in the system. Statisticians refer to that probability as the power of a test because it allows us to know how well a test can detect departures from parity. We can evaluate the power of statistical test based on a balancing methodology, and we can show that the power to detect discrimination beyond the materiality level defined by one-half "delta" is above 50 percent.

Q. WOULD YOU DISCUSS MATERIALITY AGAIN IN THE CONTEXT THAT WE ARE USING THE TERM IN THIS PROCEEDING?

A. Certainly. Recall from my direct testimony that as long as the average time taken to provide the relevant service to an ALEC does not exceed the BellSouth mean plus one-half "delta" times the BellSouth standard deviation, then the apparent difference in mean service times would not be material. That is, we would not conclude that BellSouth is providing discriminatory service. To state this another way, one-half delta, the parameter that defines the alternative hypothesis for balancing, is a materiality threshold for the disparity in the service system when a balancing method is used for a mean measure test.

Q. WOULD YOU PROVIDE US AN EXAMPLE OF THIS?

A. Yes. Figure 2 shows the probability that a mean measure statistical test will detect a difference in the mean performance of BellSouth and an ALEC when the balancing alternative hypothesis uses a "delta" of 1. To calculate these we assume that the true disparity is 0, 0.2, 0.45, etc. For the purpose of this example I am defining the "true disparity" as the numbers indicated across the top of the chart. This is not an observable figure; I am assuming the disparity to exist to illustrate what I am talking about. If we have used a delta of 1, this chart would tell us that any "true discrepancy" below 0.5 is immaterial and any "true discrepancy" above 0.5 is material. The chart shows the probability of detecting this condition. Using an example from the chart, assume a very small sample size, which is always going to be problematic. In the first line, even if the "true disparity" was zero, that is there was no disparity, the statistical analysis is going to show that there is disparity 32 percent of the time. On the other end of the scale, at 1, the analysis is only going to show a material difference 68 percent of the time, when we know that the disparity actually exists and is material. These are essentially examples of Type 1 and Type II errors, where the Type II error at the 1 disparity level is 32 percent (the complement of the probability of detection). Importantly, as the sample size increases, the analysis rapidly approaches an accuracy level of 100 percent, meaning that the Type I and Type II errors are essentially eliminated.

Figure 2: The Probability of Detecting Disparity
Mean Measure Test with Delta = 1

BST Sample Size	ALEC Sample Size	Balancing Critical Value	True Disparity Level						
			0	0.2	0.45	0.5	.55	0.8	1
10	1	-0.477	0.317	0.387	0.481	0.5	0.519	0.613	0.683
100	5	-1.091	0.138	0.256	0.457	0.5	0.543	0.744	0.862
1000	50	-3.45	0	0.019	0.365	0.5	0.635	0.981	1
12000	800	-13.693	0	0	0.085	0.5	0.915	1	1
100000	2500	-24.693	0	0	0.007	0.5	0.993	1	1

Q. IT SEEMS THEN THAT A MEAN MEASURE TEST BASED ON A BALANCING METHODOLOGY DOES MAKE IT POSSIBLE TO DETECT DISCRIMINATION AS LONG AS THE TRUE DISPARITY IS BEYOND THE MATERIALITY THRESHOLD. IS THAT TRUE?

A. Yes, a mean measure test based on balancing and large sample sizes has a high likelihood of detecting disparity beyond the materiality threshold, but a low probability of detecting disparity that falls under the threshold.

Q. ISN'T IT TRUE THAT THESE CONDITIONS ARE THE SAME ONES THAT LEAD TO BALANCING CRITICAL VALUES THAT ARE FURTHER FROM ZERO THAN THOSE THAT ARE CONVENTIONALLY USED?

A. Yes. Large sample sizes lead to critical values that are further from zero than those that are used in many applications. Such critical values, in turn, lead to small significance levels. But, as I have shown, those small significance levels (which are the probabilities corresponding to a true disparity of 0 in Figure 2)

do not imply that BellSouth will get away with any amount of discrimination. Those levels of disparity that are lower than the materiality threshold, which is defined by the choice of delta, will not be considered discriminatory. However, levels of disparity beyond the materiality threshold will be detected as discriminatory with a high likelihood.

Q. IS THE SAME THING TRUE FOR PROPORTION MEASURES?

A. A similar statement can be made for a proportion measure test. When using an odds ratio approach to balancing, the materiality threshold in not one-half of the odds ratio used in the balancing alternative hypothesis, but the threshold is at a point close to this. Figure 3 below illustrates this by showing the probability that the testing procedure will determine disparity (reject the null hypothesis), for a range of disparity levels and BST/ALEC sample sizes when the BellSouth proportion of missed installations is 0.20 and balancing is done for the alternative hypothesis with an odds ratio of 3.

Notice that for a balancing alternative with odds ratio of 3 (BST proportion of 0.20 and CLEC proportion of 0.43), there is a significant probability of determining disparity for odds ratio levels less than 3. For example, with a CLEC proportion of misses of 0.30 there is at least a 50% chance, regardless of sample size, that disparity will be determined and a remedy paid. Here we have an odds ratio of 1.75, much less than the balancing alternative of 3.

Figure 3: The Probability Of Determining Disparity When the BellSouth Proportion of Missed Installations is 0.20 and the Balancing Critical Value is Determined at an Odds Ratio of 3

Number of Transactions		Level of Disparity in Terms of Odds Ratio *Level of Disparity in Terms of CLEC Proportion*						
BST	ALEC	1*	1.25	1.75	2	2.25	2.75	3**
		0.20	*0.24*	*0.30*	*0.33*	*0.36*	*0.41*	*0 43*
10	1	0.4110	0.4440	0.5000	0.5220	0.5410	0.5750	0.5890
100	5	0.2920	0.3730	0.5040	0.5570	0.6030	0.6790	0.7080
1000	50	0.0410	0.1530	0.5130	0.6750	0.7960	0.9300	0.9590
12000	800	0.0000	0.0000	0.5520	0.9640	0.9990	1.0000	1.0000
100000	2500	0.0000	0.0000	0.5930	0.9990	1.0000	1.0000	1.0000

Q PLEASE RECAP YOUR POINT REGARDING DR. FORD'S TESTIMONY THAT YOU HAVE BEEN DISCUSSING.

A. Dr. Ford seems to believe that low significance levels means that actual and material disparities will not be discovered, particularly with large sample sizes. That is simply not true, as I have demonstrated above.

Q. LET'S MOVE ON TO THE SECOND ISSUE YOU BELIEVE DR. FORD IS CONFUSED ABOUT. CAN YOU DESCRIBE YOUR POINT WITH MORE SPECIFICITY?

A. Dr. Ford seems concerned about the large critical values that can result from the analysis that is proposed in BellSouth's plan. He believes that some sort of "standard analysis" would preclude the use of significance levels below one percent. For example, on page 20 of his direct testimony, lines 11 – 13, he states, "Recall that standard significance levels of a means-difference test are

* An odds ratio of one assumes that there is parity. Thus, the probability of determining disparity in this situation is the probability of a Type I error.

** The probability of determining disparity increases as the level of disparity goes beyond an odds ration of three.

5%, or in some cases as low as 1%. A 1% significance level is considered quite small. Rarely are significance levels chosen below this value." Basically, he is suggesting that large critical values in and of themselves suggest some sort of problem and that there ought to be a floor on critical values to eliminate any such problems.

The problem with appealing to the "standard," or "conventional" testing approach that is described by most introductory statistical textbooks, and even more advanced textbooks, is that there is almost always an assumption that the data in a study are collected according to a designed plan and that there is more than ample time to evaluate critically the data that is being used. In the simplest of cases, the assumption is that a simple random sample has been collected. In more complex cases, such as agricultural experiments or clinical trials, the sampling plans call for collecting data in specific ways. In all these cases, the sample size of the data collected is usually under the control of the data collector.

Most statistical textbooks also warn users of statistics to think about the results that they are observing. Just because a test results in a statistically significant difference between two means or proportions, one should also make sure that the observed difference makes sense from a practical point of view. This is especially true when sample sizes are very large. In these cases, Z statistics may have a large magnitude even when the actual difference between the performance measures is quite small.

Q. WHY ARE THESE POINTS IMPORTANT IN THE CONTEXT OF THIS PROCEEDING?

A. There are two reasons. First, the performance assessment plans that we are dealing with involve observational studies. This is a process where the subjects select themselves into one of the groups that are being compared. In our case customers select the telephone company that they want. We have very little control over this, and unlike the situations that textbooks usually cover; we have no control over the sample sizes that will be used every month.

Q. WHAT IS THE SECOND ISSUE?

A. The analysis of this data must be completed in a short amount of time, for many measures, every month. Normally, a good statistician would explore the data, and try to answer many questions about the data. This is particularly true when seemingly large Z values are calculated, which seems to be Dr. Ford's concern. That is, normally you should try to discover why such large Z values occurred. Was it due to a large discrepancy in the performance measure? Or, maybe it is the case that, from a practical point of view, there is very little difference in performance and the large Z value was simply caused by large sample sizes.

Q. DOES THE FACT THAT THESE PLANS REQUIRE VERY SPEEDY REPORTING AND PROVIDE ALMOST NO TIME FOR ANY CRITICAL ANALYSIS MEAN THAT THE STATISTICAL METHOD PROPOSED IS

SIMPLY INAPPROPRIATE TO USE?

A. Absolutely not. Indeed, the speed with which the data is to be reported and penalties paid if owed is one of the reasons why the Ernst & Young statistical team felt a balancing method was valuable. Large Z values that go beyond a balancing critical value are most likely caused by truly disparate treatment. But Z values that don't go beyond the balancing critical value are immaterial in terms of the difference in performance.

Q. IS THERE ANYTHING ELSE IN DR. FORD'S TESTIMONY THAT YOU FEEL IS IMPORTANT TO DISCUSS?

A. Yes. I think it is important to discuss the opposite of the small significance level issue that Dr. Ford raises. That is, the use of significance levels that are much larger than what is conventionally used when sample sizes are small. I would also like to discuss a graph in Dr. Ford's testimony that is very misleading.

Q. WHAT HAPPENS TO THE SIGNIFICANCE LEVEL OF A BALANCED STATISTICAL TEST WHEN SAMPLES ARE SMALL?

A. The significance levels can be 3, 5, or even up to almost 10 times larger than a conventional value of 5 percent. For example, with a BellSouth sample of 1000, an ALEC sample of 30, and a "delta" value of 0.25, the balancing critical value of a mean measure test is –0.675. This gives a significance level for the

test of about 25 percent. This means that BellSouth would be found to be out of parity 25 percent of the time.

Q. DOES THIS MEAN THAT YOU HAVE AN OBJECTION TO BALANCING FOR SMALL SAMPLES?

A. No. This is what balancing is supposed to do. When sample sizes are small it gives the benefit of the doubt to the ALEC. On the flip side, the data must show that there is a material difference, not just a conventionally significant difference, in the performance measure when the sample sizes are large.

Q. ISN'T IT MORE LIKELY THAT SAMPLE SIZES WILL BE LARGE?

A. On the contrary, in the performance measure data that I have looked at sample sizes tend to be small enough in such areas as UNE services and other special types of services that the balancing critical value of a Tier I test tends to be between 0 and –1. In the example I give above, samples of 1000 (BellSouth) and 30 (ALEC) lead to a balancing critical value of –0.675 for a "delta" of 0.25, and -1.35 for a "delta" value of 0.5. While a sample of size 30 for the ALEC is not huge, many would not consider it to be overly small.

Q. DOES DR. FORD RECOGNIZE THIS FACT ABOUT LARGE SIGNIFICANCE LEVELS FOR SMALL SAMPLE SIZES?

A. I believe he does, since he suggests using a "delta function" to choose "delta" based on the ALEC sample size. But I do not believe that this is the correct concept. Balancing error probabilities is not about searching for critical values that in some sense makes the two sides happy. When one adopts a balancing approach it has to be understood that you are really trying to determine what type of difference in performance truly has a material impact on an ALEC's business.

Q. YOU HAVEN'T MENTIONED HOW DR. BELL FEELS ABOUT THE EFFECTS OF BALANCING OF THE SIGNIFICANCE LEVEL OF THE TEST. DOES HE THINK THERE NEEDS TO BE A "FIX" FOR THE METHOD?

A. In discussing large negative Z scores that do not trigger a test failure because the balancing critical value is larger (further from zero than the Z score), Dr. Bell states on page 14, lines 16-17, "Such an outcome would be justified only if one could be certain that delta has not been set too large." He goes on to say that he feels no floor is warranted if the "delta" he advocates, 0.25, is used.

From this statement, I infer that Dr. Bell understands that a balanced test has sufficient power to detect truly discriminatory performance on the part of BellSouth. However, this will only be true if "delta" is chosen so that it effectively defines the materiality threshold.

Q. DO YOU AGREE WITH DR. BELL?

A. In principal yes. I am not convinced, however, that a "delta" of 0.25 is correct. The Louisiana Public Service Commission has ordered BellSouth to use a "delta" of 1. The Georgia Public Service Commission has ordered that a "delta" of 0.5 be used. In both situations, there will be periodic reviews of the effectiveness of the methodology. I assume that if these commissions find that "delta" was set too large, they will lower the value. It's also possible that a review will find that the values are too low. Only time will tell.

Q. YOU STATED THAT DR. FORD HAS INCLUDED A GRAPH IS HIS DIRECT TESTIMONY THAT IS MISSLEADING. PLEASE EXPLAIN THIS TO US.

A. Exhibit No.___(GSF-3) of Dr. Ford's direct testimony is supposed to be a graph that shows the alternative distribution with different "delta" values. Dr. Ford does not identify the exact distribution he is using, but based on the bell-shapes he uses, and the language in his testimony, I assume that he is using a normal distribution. Given that, there is no way his graph illustrates distributions that are shifted 0.25, 0.5 and 1 standard deviations from the BellSouth distribution.

Q. HOW CAN YOU TELL THAT?

A. The normal distribution has certain properties about it that indicate to you the size of its standard deviation based on the spread of the bell-curve. Figure 4 below illustrates this.

FIGURE 4: The Normal Distribution
Illustration of the Relationship Between
The Spread of the Bell-Curve and the Standard Deviation

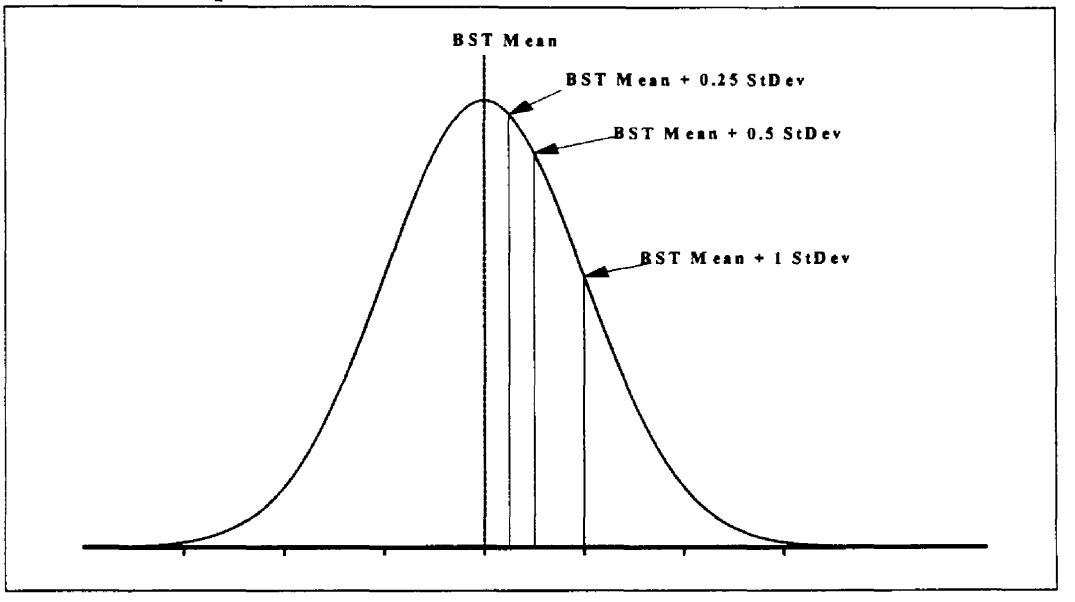


Figure 4 shows the location of the points that are 0.25, 0.5 and 1 standard deviations (StDev) from the mean of the distribution (BST Mean is at the center of the bell-curve). We can also look up the area under a normal bell-curve to the left of each of these values. These areas are approximately 60, 70, and 84 percent of the total area under the curve for the points 0.25, 0.5 and 1 standard deviation from the mean, respectively. A visual inspection of Figure 4 will indicate that this graph exhibits these area features.

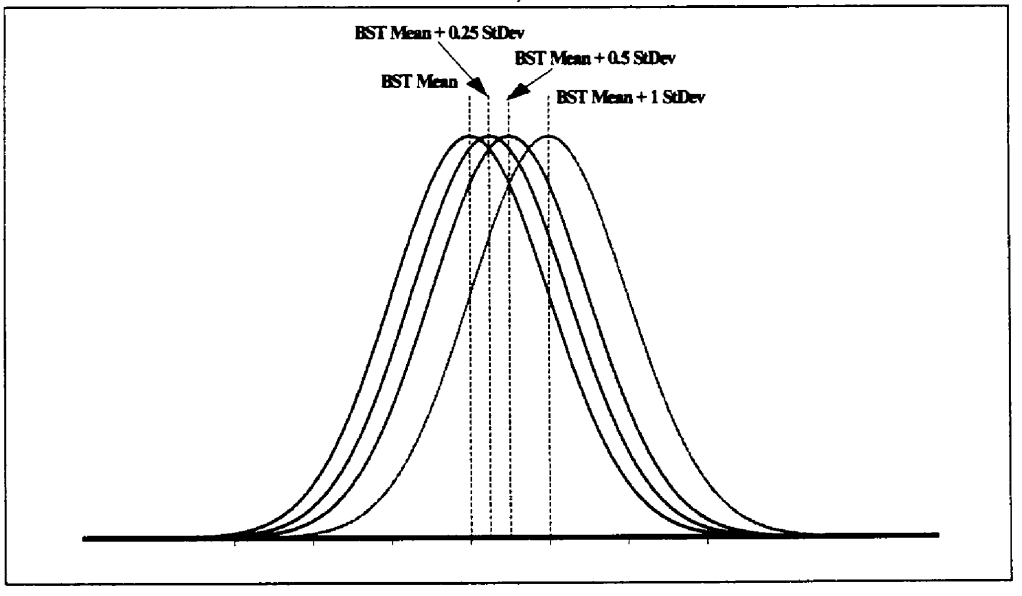
Looking at Dr. Ford's graph in Exhibit No.___(GSF-3), he does not place these points correctly on his graph. The point where he places the mean plus 0.25

standard deviation appears to really be about 2 standard deviations from the mean. I can only guess that the point that is supposed to be 1 standard deviation from the mean is located about 8 standard deviations from the mean.

Q. WHAT SHOULD THE CONCEPT DR. FORD IS ATTEMPTING TO ILLUSTRATE REALLY LOOK LIKE?

A. Figure 5 shows 4 bell-curves. The first one on the left represents the BellSouth service time distribution. The second one represents the alternative hypothesis distribution for an ALEC that has the same standard deviation as the BellSouth distribution, but its mean is larger than BellSouth's by 0.25 standard deviations. The third and fourth bell-curve are similar, representing ALEC means that are 0.5 and 1 standard deviations larger than the BellSouth means.

FIGURE 5: Location of Alternative Normal Distributions With Respect to the BellSouth Distribution
Delta = 0.25, 0.5 and 1



Q. THIS CERTAINLY GIVES A MUCH DIFFERENT VISUAL REPRESENTATION THAN DR. FORD'S GRAPH. WHY IS DR. FORD'S GRAPH SO DIFFERENT?

A. I am not sure. Perhaps he is not using a normal distribution. But his curves are symmetric bell-shapes, and while there are other distributions with similar shapes, the relationship between the curve and the point that is one standard deviation from the mean is not that much different from where it is located based on the normal distribution. I can only conclude that Dr. Ford either doesn't understand what he is doing, or he is deliberately trying to be misleading.

Q. DOES THIS CONCLUDE YOUR REBUTTAL TESTIMONY?

A. Yes.