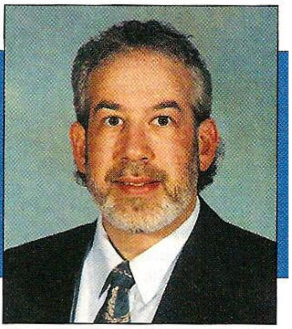
Electromedicine

CES in the Treatment of Anxiety Disorders - Part 2

Statistical Considerations in the Meta-Analysis of Cranial Electrotherapy Stimulation (CES) treatment of Anxiety Disorders

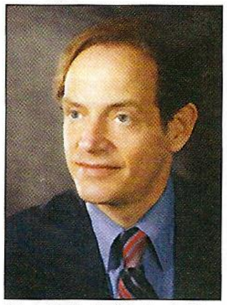
By Daniel L. Kirsch, PhD, DAAPM, FAIS, and Marshall F. Gilula, MD



Daniel L. Kirsch, PhD, DAAPM, FAIS
Department Head

While anecdotal results of CES treatment for anxiety disorders are invariably positive, a rigorous, scientific approach is required for analyzing, collating, and reporting results from the vast body of research done on CES. Due to varying methodologies and measures, the myriad of studies do not lend themselves to a simple consolidation of results. Therefore, a statistical method called 'meta-analysis' is used to combine results in a meaningful way and allow an objective measure of the efficacy of CES.

— Daniel L. Kirsch, PhD
— Marshall F. Gilula, MD



Marshall F. Gilula, MD

Part 2 continues from the March 2007 issue of *Practical Pain Management*. Meta-analysis is a statistical method of combining the results of several studies that address a set of related research hypotheses. Because the results from different studies investigating different independent variables are measured on different scales, the dependent variables in a meta-analysis are some standardized measure of effect size. The usual effect size indicator is either the standardized mean difference or an odds ratio in experiments with outcomes of dichotomous variables (success versus failure).

In this case, a meta-analysis of CES calculates the percent of patients improving versus the percent not improving to yield the treatment effect size r, which is equal to the amount of patient improvement given as percentage.[32] In the previous issue, it was reported that results of 500 patients produced an effect size r=.62. When the smaller groups of patients with specific types of anxiety related disorders were broken out, the effect size among those suffering from panic disorder was r=.45, OCD patients, r=.68, bi-polar disorder r=.71, PTSD (r =.55) ADHD (r =.62), and phobias (r =.49). The overall mean effect size for the combined smaller groups was r=.64. These results can be compared with the accepted standardized ratings of r=.10 for small effect, r=.30 for medium effect and r=.50 for large effect.[33] Thus it can be seen that the overall effect of CES for anxiety disorders is large and that there is a notable effect of duration of use that enhances such outcomes.

Statistical Significance

When any given study is published, the authors analyze the data and report whether or not the treatment utilized in their study had a discernable effect. They may report that the treatment had a significant effect at the .05 or .01, or .001 level of probability. In the first instance, the .05 indicates that if the study were to be repeated 100 times, the changes found might have occurred *by chance alone* only 5 times out of 100. Or in the case of .01 or .001 level of probability, the result would be expected to have occurred *by chance alone* only one time out of 100 or one time out of 1,000, respectively.

This form of data analysis and reporting are the hallmarks of contemporary science. Most health care professionals invest meaning in such reporting and deduce that we can have confidence in such data. We can know that the treatment effect is almost certainly genuine and effective, especially if we see one with a .001 probability utilized, as one can assume that a study yielding a probability of p<.001 had a really strong clinical effect.

Such considerations are called statistical significance. However, statistical significance does not always tell us anything regarding the actual improvement or efficacy of the treatment studied. For example, what if the study were designed to discern the effect of painting hospital room walls sunlight yellow for severe pain patients? In this hypothetical study, researchers might measure the patients' feelings of well being on a 100 point scale. Suppose most of the patients began at 3 on the scale, with a scoring range from 1 to 5, indicating very low feelings of well being, and went up to 4 on the 100 point scale after their room was painted. Although the average score increased by one-third, and that change was found to be significant at the p<.01 level, we are compelled to ask how important is such a finding to the total well being of pain patients, and by extension, what do such results imply in terms of cost and time impact (supposing that one were to use these results to justify painting the walls of hospital wards yellow)?

So how important is statistical significance? The answer can depend on many things, such as how much treatment effect, or patient improvement the treatment yielded, and the significance figure does not provide this. One could also consider what treatment costs are involved in the process of effecting that change, and whether there are other treatments available that can make the same, or even greater changes at less cost. In our hypothetical scenario such factors might involve the cost of scraping the old paint off, removing the mold, repairing and repainting the walls, and comparing that to other treatments that are available for the same amount of money that might provide equal or greater benefits for pain patients. Most of these questions are not statistical, but are questions of clinical relevance, cost and benefit, and can involve personal values as well.

There is also a second use of the term "significant" in medical literature. Most pharmaceutical companies state that any improvement of 25% or better is significant. Is a 25% improvement also statistically significant? Not necessarily. For example, if we compare the results of a treated group with the results of a sham treated group, the difference may not be statistically significant in that both groups may improve. Both may have improved 25% or more during the course of the study. Such results may not be reported in the journal article or advertisement for a given product. Instead, a statement such as, "40% of the treated group improved significantly at 25% or above" may be all that is provided. So when such studies are read, one needs to always look for any comparison between treated and sham treated subjects. For example, some follow-up studies by public interest groups have shown that several major antidepressant medications were later found to be no better than placebo treatment.[34] Also, seasoned neurological researchers opine that many new anticonvulsant drug studies routinely exclude from the treatment group any patients who show an initial intolerance to the drug, and this percentage of a selected population may routinely fluctuate between 12-25% of the population that is selected for testing.[35]

Effect Size

So while in the past there has been a focus on significant results in scientific studies, we now understand that the term, "significant" can be used in at least two ways. It can be inferred that "significance" alone is no longer an exclusive hallmark of sufficient information. A clinician needs to know how effective a treatment is in terms of the actual amount of improvement it produces in order to make an informed decision about which intervention to use. Certainly there might be less interest in a highly significant statistical result if the reduction of a given symptom is only 3%, and if that were clearly stated in the results section of a journal article or in an advertisement.

If two different studies report the results of two different types of treatments, and the results of both were found to be significant at the .05 level of significance, one would clearly be more interested in the one that resulted in a symptom reduction of 80% over the one that resulted in a reduction of 15%. This difference is known as the effect size. In advertisements and much of the scientific literature, this is ordinarily not disclosed. The reader cannot know the effect size from a study unless the published results are carefully evaluated for percent improvement pre- to post-treatment, above and beyond that of the controls.

Another problem is that when a treatment is used in studies of various groups in different parts of the country, or with groups showing slight differences in their diagnostic profile (or with groups studied at different times of the year), studies may all report significant improvement of the patients at the .05 level of confidence, but the effect sizes, when these can be ascertained, may vary considerably across the studies. A physician who wants to know what to expect if a medication or device is used in practice cannot accurately derive this knowledge from this type of reporting, and may thus not be able to reproduce the reported effects in actual patients. The best way to determine the overall effect from diverse and numerous studies is through the use of meta-analysis.

Meta-analysis is a statistical technique in which all the effect sizes found in a group of studies of the same treatment can be summarized into an overall average effect size. The derived mean effect size is what one can expect to see in most treated patients, most of the time. If meta-analysis of studies yields an average effect size of 15%, this will be of less interest to the practicing physician than a meta-analytic finding from another treatment which treats the same problem, but results in an average effect size of 60%.

Simply stated, the r effect size represents the percentage improvement to be expected on a scale of 0 to 100. An r effect size of .15 means that there was an average of only 15% improvement among patients when measured across combined studies, while $r=.75$ means that there was an average of 75% improvement in patients found in the combined studies, etc. In this scale, an r effect size of .10 is small, while r of .30 is moderate, and r of .50 or above is considered to be high.

Many early statistical meta-analyses were confined to studies that specifically reported the pre- and post-study means and standard deviations. All other studies had to be ignored, no matter how rigorous the scientific protocols. That left out the results of some well designed, well conducted double-blind placebo-controlled studies. Current use of meta-analysis tends to statistically transform whatever statistic the author reports into an effect size statistic and then proceed with subsequent analysis from that data set of the collected studies (See Appendix A for an example).

What one might gain from this discussion is that the effect sizes obtained by meta-analytical procedures of CES studies is very robust and holds up to scrutiny very well given the reasonably large number of studies available to work with. The effect size of CES — as derived from Tables 5 and 6 — was seen to stabilize in the high 50s or low 60s, with the expected effect size in 99 out of 100 times in a future meta-analysis of studies to range from $r=.40$s to $r=.70$s. That range is considered to represent a moderate to very strong clinical improvement.

Discussion of CES Meta-Analysis Results

The most pristine analysis of the re-ordered data yielded an effect size of $r=.57$ (as opposed to the un-ordered meta-analysis table that yielded $r=.58$). Analysis of the studies that used only the double-blind method provided $r=.53$.

After removing extraneous measures of anxiety and only analyzing for state anxiety or trait anxiety using the State/Trait Anxiety Inventory, $r=.60$ for state anxiety and $r=.68$ for trait anxiety. These involved a relatively small number of studies. When results were corrected for the number of subjects in each study, the r for state anxiety fell back to a more typical $r=.59$, while trait anxiety fell back to a more typical $r=.60$.

What this means is that while Klawansky at Harvard found an average effect size of $r=.53$ in earlier meta-analysis of eight CES studies, and O'Connor in Tulsa found an r effect size of $r=.51$ in the eight CES studies she chose, similar effect size results were obtained when more than five times that number of studies were meta-analyzed. If an additional 400 CES studies of anxiety were to be analyzed 50 years from now, the likelihood is almost certain that the effect size would still be within the

TABLE 5. PUBLISHED CES ANXIETY STUDIES

Author	Number of Patients			Statistic Reported*	Result	Zr Score
	CES	Controls	Total			
Bianco, 1994[24]	29	18	47	% Improvement, Beck AI	.77	1.02
				% Improvement, Hamilton AS	.71	.887
Feighner, 1973[36]	23	23	23	% Improvement	.31	.321
Flembaum, 1974[37]	28	Historic	25	% Pts Much, or Very Much Improved	.51	
Frankel, 1973[38]	17	17	17	% Improvement	.08	.080
Gibson, 1983[5]	16	16	32	% Improvement, EMG	.35	.365
				% Improvement, STAI	.43	.460
Gomez, 1974[23]	14	14	28	% Improvement, TMAS	.35	.365
Hearst, 1974[39]	14	14	28	% Pts. Asymptomatic	.71	
Heffernan, 1995[6]	10	10	20	t-score, EMG	2.35	.717
				t-score, Heart Rate	2.55	.784
				t-score, finger temperature	2.62	.717
				t-score, capacitance	2.14	.662
Heffernan, 1996[16]	10	20	30	% Increase in EEG Correlation Dimension	.54	.604
Jemelka, 1975[40]	14	14	28	P=<.05 Improvement, Hamilton AS	.51	.563
Kirsch, 2002[28]	298		298	% Improvement	.83	1.188
Krupitsky, 1991[41]	10	10	20	% Improvement, State Anxiety	.41	.436
				% Improvement, Trait Anxiety	.73	.929
				% Improvement, TMAS	.47	.510
Levitt, 1975[42]	5	6	11	% Improvement, TMAS	.80	1.099
McKenzie, 1976[43]	8	4	12	% Improvement, Skin Potential	.48	.523
Magora, 1967[44]	20		20	% Improvement	.75	
Matteson, 1986[19]	32	22	54	t-score, State Anxiety	4.63	.640
				t-score, Trait Anxiety	3.37	.523
				t-score, POMS Anxiety	5.43	.701
May, 1993[45]	14		14	% Improvement, MAACL	.75	.973
Moore, 1975[46]	17	17	17	% Improvement, Psychiatrist Ratings	.35	.365
Overcash, 1999[47]	182		182	% Change, EMG	.72	.908
				% Change, Electrodermal Response	.48	.523
				% Change, Temperature	.13	.131
				% Change, Self Rating Scale	.76	.996
Overcash, 1989[48]	16	16	32	% Change, EMG	.92	1.589
				% Change, 16PF, Planful Scale	.80	1.099
Passini, 1976[49]	30	30	60	% Improvement, MACL	.28	.288
				% Improvement, State Anx.	.30	.310
				% Improvement, Trait Anx.	.10	.100
Patterson, 1984[50]	186		186	% Improvement, Anxiety	.75	.973
Philip, 1991[51]	10	11	21	P = < .05 Improvement	.60	.693
Rosenthal, 1972[20]	11	11	22	% Improvement	.67	.811
Rosenthal, 1970[52]	9		9	% Improvement	.47	.510
Rosenthal, 1970a[53]	12		12	% Improvement	.54	.604
Ryan, 1976[18]	12	12	24	F statistic = 8.26	.65	.775
Ryan, 1977[17]	10	10	20	P=<.001 Improvement	.55	.618
Sausa, 1975[54]	40	40	80	% Improvement, TMAS	.35	.365
				% Improvement, HAS	.45	.485
				% Improvement, CRS	.35	.365
Schmitt, 1986[26]	30	30	60	P=<.05 STAI, State Anxiety	.35	.365
				P=<.05 STAI, Trait Anxiety	.35	.365
				P=<.05 IPAT	.35	.365
				P=<.05 POMS Anxiety	.35	.365
Smith, 1999[55]	23		23	t-score, State Anxiety	.74	.950
				t-score, Trait Anxiety	.81	1.127
Smith, 1975[27]	36	36	72	P=<.001, POMS Anxiety	.46	.497
Smith, 1992[56]	31		31	% Improvement	.41	.436
Smith, 1994[22]	10	11	21	P=<.05, POMS Anxiety	.60	.693
Smith, 2002[57]	146	107	253	P=<.03, POMS Anxiety	.19	.192
Taylor, 1991[58]	15	15	30	P=<.05, Diastolic BP	.50	.549
				P=<.05, Systolic, BP	.50	.549
				P=<.05, STAI, State	.50	.549
				P=<.05, pulse rate	.50	.549
Von Richthoven, 1980[59]	5	5	10	P=<.001, Psychiatric Rating	.97	2.092
				P=<.005, STAI, State	.92	1.589
				P=<.005, Self Rating	.92	1.589
Voris, 1995[7]	40	65	105	P=<.0001, STAI, State	.49	.536
				% Improvement, EMG	.63	.741
				P=<.01, Temperature Change	.40	.424
Voris, 1996[60]	8	7	15	P=<.01, STAI, Trait	.80	1.099
				% Improvement, EMG	.53	.590
Weingarten, 1981[61]	12	12	24	P=<.05, POMS Anxiety	.55	.618
Winick, 1999[8]	16	17	33	P=<.02, Dentist RS	.56	.633
				P=<.02, Patient SRS	.56	.633

*Beck AI is the Beck Anxiety Index; Hamilton AS is the Hamilton Anxiety Scale, also known as HAS or HAMA; EMG is the electromyogram; STAI is the State/Trait Anxiety Inventory; TMAS is the Taylor Manifest Anxiety Scale; EEG is the electroencephalograph; State Anxiety and Trait Anxiety are both from the STAI; POMS is the Profile of Mood States; MACL is the Modified Adjective Check List; 16PF is the 16 Personality Factor Scale; CRS is a Clinical Rating Scale; IPAT is the International Personality and Ability Test; BP is blood pressure; RS is rating scale; SRS is self rated response scale

TABLE 6. LIST AND DESCRIPTION OF ANXIETY STUDIES

Author	Diagnosis	Blinding			Study Design	Outcome Measure
		Patient	Therapist	Assessor		
Bianco, 1994	Polysubstance Abusers	Yes	Yes	Yes	Double-blind	Beck/Hamilton Anxiety Scale
Feighner, 1973	Psychiatric Inpatients	Yes	Yes	Yes	Double-blind Crossover	Global Rating Scale
Flemenbaum, 1974	Psychiatric Outpatients	No	No	No	Open Clinical, Historical Controls	Global Rating Scale
Frankel, I973	Insomniacs	Yes	Yes	Yes	Double-blind, Crossover	TMAS
Gibson, 1983	Outpatient Psychiatric	Yes	Yes	Yes	Double-blind	EMG, STAI
Gomez, 1974	Heroin Addicts	Yes	Yes	Yes	Double-blind	TMAS
Hearst, 1974	Outpatient Psychiatric	Yes	Yes	Yes	Double-blind	Self Ratings
Heffernan, 1996	Outpatient Pain Patients	Yes	Yes	Yes	Double-blind	4 Physiologic Measures
Heffernan, 1996a	Outpatient Pain Patients	Yes	Yes	Yes	Double-blind	EEG
Jamelka, 1975	Prisoners, Psychiatric Ward	Yes	Yes	Yes	Double-blind	Hamilton AS
Kirsch, 2002	Physicians' Report of Patient Response	No	No	No	Physician Survey	Physicians' Clinical Ratings
Krupitsky, 1991	Alcoholic Inpatients	Yes	Yes	Yes	Double-blind	STAI, TMAS
Levitt, 1975	Psychiatric Inpatients	Yes	Yes	Yes	Double-blind	TMAS
McKenzie, 1976	Psychiatric Outpatients	No	No	No	Open Clinical	Skin Potential
Magora, 1967	Psychiatric Inpatients	No	No	No	Open Clinical	Physician Clinical Rating
Matteson, 1986	Graduate Students, Business School	No	No	No	Open Clinical	STAI
May, 1993	Inpatient Drug Treatment	No	No	No	Open clinical	MAACL
Moore, 1975	Outpatient Psychiatry	Yes	No	No	Crossover	Psychiatrist Ratings
Overcash, 1999	Outpatient Psychiatry	No	No	No	Open Clinical	Physiological Measures, Self Rating Scale
Overcash, 1989	Marijuana Patients	No	No	No	Open Clnical/different therapies	EMG, 16PF
Passini, 1976	Inpatient Psychiatric	Yes	Yes	Yes	Double-blind	MACL, STAI
Patterson, 1984	Polydrug Abusers	No	No	No	Open Clinical	Abstinence Syndrome
Philip, 1991	Polydrug Withdrawal	Yes	Yes	Yes	Double-blind	Visual Analog Scale
Rosenthal, 1972	Psychiatric Outpatients	Yes	Yes	Yes	Double-blind	Psychiatrist Ratings
Rosenthal, 1970	Psychiatric Outpatients	No	No	No	Open Clinical	Psychiatrist Ratings
Rosenthal, 1970a	Psychiatric Outpatients	No	No	No	Open Clinical	Psychiatrist Ratings
Ryan, 55	Psychiatric Inpatients	Yes	Yes	Yes	Double-blind	STAI-State
Ryan, 1977	Psychiatric Inpatients	Yes	Yes	Yes	Double-blind	STAI-State
Sousa, 1975	Psychiatric Outpatients	Yes	Yes	Yes	Double-blind	TMAS, HAS, Clinical Rating Scale
Schmitt, 1986	Inpatient Polydrug	Yes	Yes	Yes	Double-blind	POMS, IPAT, STAI
Smith, 1999	Outpatient Psychiatry	No	No	No	Open Clinical	STAI
Smith, 1975	Inpatient Addiction	Yes	No	Yes	Single Blind	POMS
Smith, 1992	Outpatient Phobic	No	No	No	Open Clinical	Self Rating Scale
Smith, 1994	Closed Head Injured	Yes	Yes	Yes	Double-blind	POMS
Smith, 2002	Inpatient Polydrug	No	No	No	Retrospective	POMS
Taylor, 1991	Normal volunteers	Yes	Yes	Yes	Double-blind	BP, Pulse Rate, STAI
Von Richthofen, 1980	Anxiety Neurosis	Yes	Yes	Yes	Double-blind, Crossover	Psychiatrist RS, Self RS, STAI
Voris, 1995	Prison Parolees, Sex Offenders	Yes	Yes	Yes	Double-blind	STAI, EMG, Temperature
Voris, 1996	Prison Parolees, Sex Offenders	No	No	No	Open Clinical	STAI, EMG
Weingarten, 1981	Inpatient Alcoholics	Yes	Yes	Yes	Double-blind	POMS
Winick, 1999	Dental Patients	Yes	Yes	Yes	Double-blind	VAS

r=.44 to r=.70 range.

There are numerous statistical considerations that must be taken into account in performing meta-analysis and Appendix B illustrates the most important ones. The non-statistician may find it useful to consider these factors and gain personal confidence in this valuable technique.

Conclusion

There have now been roughly 50 years of experience in the U.S. using CES as a non-pharmaceutical treatment for anxiety although it has yet to achieve ubiquitous status as a therapeutic modality. This is most likely due to the fact that few U.S. medical schools teach CES treatment as part of their curricula, and none of the seven or eight CES companies in U.S. history have had sufficient staff to visit physicians' offices in the ubiquitous manner of today's pharmaceutical representatives.

Yet when physicians who use or prescribe CES are asked about its effectiveness, they are generally enthusiastic, as are the majority of CES patients themselves. Patient response on surveys are even more significant because some CES device distributors have a 30 day period during which a patient can return the device at little or no cost if it proves ineffective. Less than 2% of patients return the devices for this reason, and almost none are returned by patients who use them in the suggested manner for the treatment of their anxiety (e.g., 20 minutes to one hour a day for the first three weeks, then as needed to prevent symptoms from returning). The fact that such devices can cost over $1,000 makes the tendency to keep them even more impressive.

It is also noteworthy that among the more than 6,000 patients who have been involved in CES studies in the U.S., and from the thousands of patients who completed surveys, there have been no significant, negative side effects reported from the use of CES. The National Research Council evaluated the safety of CES for the FDA stating that, "...significant side effects or complications attributable to the procedure are virtually nonexistent."[67]

From the data available, one would assume that CES will continue to receive greater attention from clinicians as more become aware of the safety and efficacy of this treatment for anxiety and the myriad of anxiety related disorders, especially chronic pain. ■

Appendix A. Example of Meta-Analysis Probability Conversions

For example, if percent improvement is reported, that percent figure is converted directly into the effect size, *r*. Similarly, *Z* scores are converted to *r* by the formula $r=\frac{Z}{\sqrt{N}}$.

If student *t* scores are given, they are converted into *r* by the formula $r=\sqrt{\frac{t^2}{t^2+df}}$ where *df* is the degrees of freedom. If the author only gives the resulting probability figure, such as .05 or .01, one can convert that into the *t* score from published probability tables and compute r from the formula given just above.

When non-parametric statistics are reported, such as chi squared (X^2), the author ordinarily reports the probability estimate obtained (.05, .01, etc.). In these cases, the *r* can also be obtained by converting the probability estimate into a *t* score.

There are other considerations that a diligent statistician must keep in mind when conducting meta-analysis. For example, an author might report finger temperature as a physiologic correlate of anxiety and report that the patients' average finger temperature rose from 91 to 94 degrees farenheit.[7] For a clinical researcher in the field of biofeedback, that is a dramatic change, but how can it best be added to a meta-analysis? As important as it appears to a biofeedback therapist, it is in fact only a 3.3% improvement. That percent gain could be translated directly to an effect size r of .03 which would make it appear insignificant.

On the other hand, if a *t* score of 2.62 was derived from the patients before and after treatment data, from the formula given above (10 patients were treated, giving a df = 9) one can derive an effect size of *r*=.66. What makes that difference possible, and which effect size is the correct one? It is well known that temperature does not exist on a scale of one to 100 in humans. That is, it is not a 100 point scale. Therefore raw temperature scores must be adjusted accordingly.

One way to do that is to determine the criteria for the temperature range in humans. If the finger temperature range in subjects who would be well enough to be able to walk into a clinic to participate in a study is 95 to 101, or six points, then each temperature shift would be equal to 16.67 points on a 100 point scale and an *r* from this example, derived that way, would be 50. That is greater than the *r* of .03, but not as great as the *r* of .66.

In our example the *r* of .66 was derived from *t* scores, because the researcher had utilized actual temperature scores from the subjects who were in the study and compared both the subjects' actual mean finger temperature scores and the variance of all of those scores around the mean in arriving at the *t* score. That indicates that the range was not evenly divided from 95 to 101 among the research subjects, and was obviously much narrower if a change in three temperature points resulted in the 66% gain.

Therefore the total possible range of finger temperatures in normal people walking the streets is less important to the meta-analysis than those found among pain patients who are anxious. Pain related anxiety is known to restrict the finger temperature range considerably. That is what was found in the study used in this example, and that is why it was examined.

In one study, pre- and post-temperatures were given with no other information. Accordingly, the derived 8% improvement had to either be deleted from the meta-analysis or statistically dealt with after a very close reading of the original publication, since simply adding that *r* of .08 into the analysis would not only be in error but would unnecessarily skew or bias the results.

To summarize the problem, if a given data point measured does not ordinarily fall along a 100 point scale of variation, the percent change has to be adjusted for consistency before the number can be added to the analysis. T scores, F scores, probability scores, X^2 and the like are the calculations used to determine the actual range.