Reforming EdD Quantitative Analysis Courses from Statistical Significance and Mathematical Complexity to Analyzing Practical Significance

by

Stanley Pogrow
Professor of Leadership and Equity
San Francisco State U

stanpogrow@att.net

Rough draft. Please do not duplicate or cite without permission.

Comments and suggestions appreciated

This paper is abstracted from the recent NCPEA book:

Authentic Quantitative Analysis for Educational Leadership Decision- Making and EdD Dissertations:

A Practical, Intuitive, and Intelligible Approach to Critiquing and Applying Quantitative Research for (1) Improving Practice, and (2) Developing a Rigorous and Useful EdD Dissertation

Given the inadequacy of *statistical significance* as a basis for leaders to evaluate the state of evidence, this section will present an alternative approach to critique research and assess the quality and utility of evidence in published research and syntheses of research—specifically for leadership decision-making. Specifically, this chapter will present a way to determine the *practical significance* of evidence and how to use such data to determine whether an intervention is appropriate for your school(s). A lessor standard of *potential practical significance* will also be described. Both types of significance require a critical analysis of the nature of the sample used in the research in terms of its validity and resemblance to your situation.

The methodology presented for determining *practical significance* or *potential practical significance*, as well as the nature of the sample, provides a basis for critiquing the appropriateness, integrity, and generalizability of any formal research under consideration.

It must be emphasized that the methods described are specific for supporting leadership decision-making, and will differ substantially from the techniques described in other texts and from how researchers analyze research for their own purposes.

Issues in Defining Practical Significance

The first person to develop the idea of *practical significance* seems to be Thompson (2006) who defined it as"

> …how much difference an intervention makes or how related various variables are (e,g,, how much longer on average, will you tend to live if you do not smoke…) p. 134.

However, I much prefer the definition offered by a colleague, Christopher Tienken, associate professor of Educational Leadership at Seaton Hall University. He described *practical significance* in this way:

> *Did the students in a study benefit so substantially that it makes sense for you the leader to commit a substantial part of your discretionary monies and time, and other's time, to implement the intervention used in the study?*

Tienken's perspective is reflected in conversations I have with principals. When I ask them what they want to know from research, they invariably want to know how the intervention would do in their school, and whether it would make a BIG difference. (The only time leaders are not looking for a BIG difference is if their school(s) are already at the top. In that case any improvement is difficult to make, so a small improvement is important to them.)

Clearly, leaders are looking for evidence that they consider that if they adopt the practices they are likely to see BIG improvements. Aside from the question of how to measure "BIGNESS", there is also the real question of how to project the likely impact of research done at other sites to expected improvement in leaders' school(s). Both of these practitioner driven concerns form the basis of the techniques presented in this chapter for critiquing research in terms of its *practical significance*. Clearly, from the discussion in the previous chapter, there is a difference between *statistical significance* and BIG differences. In addition, as will be seen, conventional approaches to presenting quantitative methods as well as published quantitative research, do not adequately deal with the issue of projecting expected benefits of published research to one's own school(s).

Most quantitative studies in major journals focus to a far greater extent on internal validity, and have major *external validity* problems.[1]

Therefore, *practical significance* is defined as the likelihood that the results of a study will produce BIG improvements in <u>your</u> school(s).

Before on can figure out what 'BIG' improvement means, it clearly requires knowing what data in published research you want to compare to the current state of your school(s). However, it turns out that knowing what data to extract and finding it are problematic. To understand the problem, consider the following conversation between a husband and wife planning a January getaway to a warm beach.

> Wife: *I cannot wait for our vacation in January. Let's go somewhere warm.*
> Husband: *Definitely.*
> Wife: *Where should we go?*
> Husband: *I just read that Greenland is warmer than Antarctica in January, and that due to climate warming it will be warmer there this year than last. Plus, it has 27,394 miles of coastline, so it will be no problem finding beaches.*
> Wife*: That's great. It will be wonderful to go somewhere where we can leave our winter clothes behind.*

They are clearly in for a surprise. They have made the mistake of confusing a comparative/relative measure of something, in this case temperature, for knowing the actual, i.e., absolute, level of the temperature of the desired destination. On an absolute level the temperature in Greenland in January averages -8 degrees Celsius with 0 hours of sunshine. In other words, something can look good on a relative basis and actually be horrible on an actual/absolute basis.

Clearly this couple is not very bright. They are more likely to die from hypothermia in Greenland then get a tan. Obviously, we would never confuse <u>relative</u> outcomes for <u>actual/absolute</u> outcomes in the sophisticated world of quantitative educational research. Would we? Unfortunately we have—with equally devastating results.

Consider the following example. The program with the most positive research published to date is 'Success for All', a whole school reading model for grades k-5. Based on this extensive literature in the top journals, it received by far the most federal funding of any other program, and was the most widely adopted program in high poverty schools around the nation as a reform model over the past several decades. The most famous example of its 'success' involved an experiment in Baltimore Public Schools in the early 90's. The dataset from that work is still used in more recent research articles and proposals. That original research was also used in the program's recent successful i3 funding request in 2010 that netted it one of the 4 major $50 million awards from the U.S. Department of Education (ED).[2]

The findings of the research showed very large *Effect Sizes* (*ES*) favoring the experimental schools that used this program relative to a comparison schools; particularly for the lowest

[1] External validity is an even bigger problem with qualitative research which tends to use very small samples.

[2] i3 was then a new funding initiative from ED to support investing in innovation.

achieving students. However, a reanalysis of the data by Venezky (1998) found that not only did the experimental sample apparently drop all the special education students who had initially started the program, and all students who had not been at the school all 5 years of the experiment, this cherrypicked sample of experimental students had actually done poorly. My own calculations based on Venezky's work concluded that after 5 years in the program the Success for All students entered the 6th grade reading almost 3 years below grade level (Pogrow, 1998; Pogrow, 2000). Had all students in the Success for All schools been included, a reasonable expectation for a program that bills itself as a schoolwide model, the results would have been even worse.

Does that actual/absolute outcome for these experimental students indicate a success?

In other words, just like the couple in the first example where Greenland looked warm on a comparative basis when it was actually very cold, the fact that on a comparative basis the 'Success for All' students appeared to do better, the reality is that on an actual/absolute basis they did terribly. However, all the research and funding community, especially ED, looked at was the reported difference between the groups, and used this 'evidence' as the basis for providing extensive funding to disseminate its use. (Indeed, ED rated that program as the highest of any program that applied for the grant.) This is the equivalent to ED advocating that all educational leaders vacation in Greenland in January to get a tan.

There is another typical problem in looking at evidence on differences between groups of at-risk students in terms of the achievement gap. Classically, research studying the effectiveness of an intervention to help at-risk students compares the relative performance of at-risk students using the intervention and a similar group not using it. An example is research using the Math Pathways and Pitfalls project; an intervention that develops skills and understanding in the use of academic math language. Heller, Hansen, & Barnett-Clarke (2010) found that 30 hours of instruction with this math reform over 2 years in grades 4, 5 produced significant effect sizes for the standardized test scores of Latino and English learner students relative to those not receiving the instruction on proximal gains. This research is elegant, done with integrity, and the intervention is clearly substantive. These researchers then did the following three things that are often not done:

1. They also looked for distal gains in the form of standardized math test scores,
2. They also provided the intervention to the advantaged students and reported their distal gains as well,
3. They provided *Impact Scores* on the actual/absolute progress of all the group's distal scores.

If you look inside the impact estimates on the standardized tests section, the impact coefficients of the program were (a) close to zero for the experimental at-risk students, i.e., they made no progress, and (b) negative for the comparison at-risk students, i.e., they fell further behind. In other words, most of the advantage for the experimental group resulted from the comparison group declining. Interestingly, the impact coefficient for the advantaged students who also received the intervention was positive, i.e., they made progress. In other words, providing the same high quality intervention to all students actually widened the achievement gap, though less so for the experimental at-risk students. In other words, what appears to be effective based on the relative performance of two groups of at-risk students can mask the undesirable absolute outcomes that (a) the experimental group did not progress and their advantage resulted from the

comparison group's decline, and (b) the achievement gap widened. As a result, while a significant ES in such research will cause researchers to have an orgasm, and promote the use of the intervention, there is no way that educational leaders should adopt an intervention on that basis. This is not the rejection of the use of research—it is prudent decision-making.

The importance of rejecting relative data and seeking data on absolute performance was brought home to me when the State of New Jersey tried to pressure all the high poverty elementary schools that were receiving extra state funds to implement the 'Success for All' program. I asked one New Jersey principal whose school was rapidly improving why he had decided not to implement that program. His response was: "I looked at their research and concluded that the students in my school were already doing better than how students in that program ended up doing." In other words, the principal's criterion was to compare the absolute performance of the experimental students (only) to the current performance of his students.

This principal's insight also illustrates that there is a fundamental difference in the type of research questions asked by academicians and leaders as illustrated in Table 3.1 below:

Table 3.1
Differences in Key Questions Asked of Research Findings

Academicians/Researchers	Leaders
Did the experimental group in the research context do better than the comparison group, and was that difference statistically significant? What are the implications of the findings for theory?	If I adopt the approach used by the experimental group in the study is it likely to produce a BIG difference in my school(s)?

This difference in perspective is not trivial. Indeed, the problem of leaders extrapolating likely outcomes from the context used in published research to one's own school(s) is far more complex than is generally acknowledged in methodology textbooks (unless of course the research was conducted in the leader's own school/district).

In addition to highlighting this important difference of perspective in the type of insights leaders seek from research, this principal's conclusion also illustrated the following two points:

- It was possible for a non-statistician to determine how the experimental students did on an absolute basis—if you looked for the right numbers.
- The key issue was not how the experimental students did relative to the comparison group in the study, *but how the experimental students did on an absolute basis compared to what his students were already doing*.

The latter is an important, common sense insight that forms the major basis in this chapter for critiquing research evidence in terms of its *practical significance*. Specifically, the key criteria for critiquing research evidence in terms of its *practical significance* for helping improve your school(s) are to ask the following critical questions of the evidence presented in a study:

- How did the experimental students in the research actually do on an absolute non-weighted basis? [3]
- How big a comparative advantage did the performance of the experimental group provide as compared to what currently exists in your school(s), and is there a big difference?
- How was the sample in the study constructed and how similar was it to the demographics of your school/district?

Most of the more sophisticated statistics and data presented in quantitative research are used for comparing experimental and control groups while simultaneously controlling for several variables, does little to answer these 3 questions. Fortunately, determining the answer to these questions can be done with very basic statistics.

Determining the Practical Significance of Research Evidence

The typical quantitative research study focuses on the differences between the groups, with tons and tons of intimidating data massaging, statistics, tables, and technical jargon. How can the average leader/student deal with it? The good news is that you do not need to. Ignore it all—since the stuff you do not understand has little to do with answering the 3 key questions listed above for determining *practical significance*. Indeed, you can ignore virtually all the technical terms you do not already know in any quantitative study.

Instead, focus on the statistics that indicate the absolute performance of the experimental group. The most important statistics that typically indicate the absolute performance are the *Mean*/Average, the Standard Deviation/Variance, and the Median. (Hereafter the terms *Mean* and *Standard Deviation* will be used as they are more common in the literature.)

i. The relationship between the mean and standard deviation.

The *Mean* is in theory the statistic that is most representative of all the different scores in a distribution. For example, the *Mean* of how all the experimental students did is in theory the most representative value for that group, even though most of the scores will typically be different than the mean. Some individuals will have larger scores than the *Mean*, some will have lower scores, and some will have the exact same score. So, for example, if the *Mean* score on reading for an experimental group of 5th graders at the end of the year was 5.2, few of the actual scores would necessarily need to be 5.2 for it to be the most representative value for the group as a whole. However, in certain circumstances the *Mean* can be misleading.

For example, consider two communities, each of which has 10 families. Using Table 3.2 below, which of the following two communities is richer?

Table 3.2

A Comparison of the Average Family Income in Two Communities

	Community A	Community B
Mean Family Income	$111,925	$36,349

[3] The meaning of non-weighted will be explained shortly.

Clearly, the answer is Community A. But is it? We cannot know for sure till we look at the distribution of individual incomes in each community. Table 3.3 has a distribution, i.e., a listing, of each family's income in each community.

Table 3.3

A Distribution of All Incomes by Community

Income of Each Family in Community A	Income of Each Family in Community B
12,200.00	39,000.00
11,000.00	28,800.00
11,600.00	42,000.00
16,500.00	38,500.00
19,245.00	33,200.00
17,000.00	46,000.00
9,000.00	28,000.00
14,800.00	37,000.00
7,900.00	34,890.00
1,000,000.00	36,100.00

After eyeballing the two columns, do you still think Community A is the wealthier one? Probably not! What has happened is that Community A has an outlier, i.e., a very unusual case with an unusually high income, and this single datapoint *biased*, or *skewed*, the *Mean* sharply upwards. As a result, the *Mean* for Community A was not representative of that community as a whole since the community was in reality very poor. The statistic that would indicate that the *Mean* was not representative of Community A, or any situation, is the *Standard Deviation.* The *Standard Deviation* indicates how much variation there is in the distribution of all scores relative to the Mean. The larger the divergence of scores from the mean, i.e., the higher the *Standard Deviation is relative to the Mean,* the less representative and more misleading the *Mean* is.

Table 3.4 below adds the standard deviation to Table 3.2.

Table 3.4

The Mean and Standard Deviation of Family Income in Two Communities

	Community A	Community B
Mean Family Income	$111,925	$36,349
Standard Deviation	312,059	5,537

Table 3.4 shows that the Standard Deviation of the individual incomes in Community A is very large relative to the *Mean*, so much so that it is actually bigger than the *Mean*, while the *Standard Deviation* for Community B's scores are much less than the *Mean*. This indicates that the *Mean* is not a representative of family income in Community A—but that it is for Community B.

When the *Standard Deviation* is large relative to the *Mean* a different statistic needs to be used to represent the central tendency of the distribution. Usually that statistic is the *Median*, which is the middle score of the distribution. The *Median* is calculated by listing all the individual scores, in this case the income of each family, in chronological and selecting the one in the middle. Since there is an even number of scores in this case, the midpoint would be in the middle of the 5th and 6th scores. In the case of Community A, it would be the midpoint of 12,200 and 14,800, or 13,500. Table 3.5 shows the difference between the *Mean* and *Median* scores.

Table 3.5
The Mean and Standard Deviation of Family Income by Community

	Community A	Community B
Mean Family Income	$111,925	$36,349
Median Family Income	$13,500	$36,550

Clearly, for this *distribution* of incomes the *Median* is a more representative score and indicates clearly that in fact Community B is the wealthier one. This example illustrates that when the *Standard Deviation* is large in comparison to the *Mean*, then the *Median* is usually the more accurate statistic. (Note how different the *Mean* and *Median* are for Community A.)

Basic rule of Mean-Median differences: When the *Mean* and *Median* are very different, trust the *Median,* or at the very least explore the nature of the *distribution*.

This discussion also explains why the average home price in a geographic area is reported is the *Median* price. The *Median* is used because there are usually a few exceptionally expensive homes that are sold whose prices are vastly different then the price range of the vast majority of homes sold. This makes the *Mean* price too unrepresentative, and therefore the price of the home in the middle of the distribution, i.e., the *Median*, is used.

Now let's apply this knowledge of the relationship between *Mean, Standard Deviation*, and *Median* to education research. Consider a study to determine whether a new practice increases the mathematics performance of 5^{th} graders. Suppose that by the end of the year the *Mean* math score of the experimental group is 6.2, and that of the comparison Group A's end of year *Mean* math score is 5.8. Which group did better?

We cannot say which group of 5^{th} graders did better until we know the Standard Deviation for each group. Suppose the Standard Deviation for the comparison group is 5 or even 2. Such a large Standard Deviation relative to the *Mean* of 5.8 indicates that the *Mean* is not representative of the sample. For example, it may be that the comparison group as a whole did much better than the experimental group, but that a few of the students in the comparison group did not try and handed in blank tests which dramatically reduced that group's overall *Mean*. These *outlier* scores may have disproportionately pulled down the overall *Mean*, while all the others in the comparison group did very well.

The rule of comparative Standard Deviations. We can only be confident that the experimental group did better if the *Standard Deviations* of both groups are small relative to their *Mean*. Otherwise the comparison has to be done based on the *Median* score(s).

So given that one of the standard deviations is large in the above example, you cannot determine which group did better by examining the *Means*. Determining which group of 5th graders did better requires knowing: *Which group of 5th graders had the larger Median score?*

Ironically the most advanced forms of statistical analyses can only do calculations based on the *Mean*. Hmmm… Houston—We have a problem!

[**SIDENOTE**: another problem is that the most advanced widely used statistics in education currently assume that relationships are linear; while many clearly are not. In other words, advancements in the use of statistics in education research may be providing knowledge about a different universe than the real education world of practice. At the same time, this is not intended to be a screed condemning quantitative analysis or advanced statistics. Rather, it is to point out that in many cases the most basic statistics are the most informative for guiding real world leadership decision-making.]

Getting back to the issue of using the *Standard Deviation* as a tool to indicate whether the *Median* should be used as the most representative statistic, the key question becomes: What constitutes a large *Standard Deviation* relative to the *Mean*? There is no formal rule. You can use common sense. In the case of the math scores just discussed, given the *Mean* values of 5.8 and 6.2, I would want to see a *Standard Deviation* of 1.5 or less to have confidence in the *Mean*. The basic rule is that when in doubt, calculate and examine the *Median*, and apply the *basic rule of mean-median differences*. [4]

Of course, in this example it was easy to eyeball the data in the community wealth example and figure out what was going on. How does one handle a more normal situation where there are thousands of families in a community, or even hundreds of thousands? Instead of looking through a column of thousands of incomes, one uses a statistical package such as SPSS to generate a picture of the distribution of all the values. A *Histogram* indicates how often each score occurs using bars, while a *Frequency Polygon* produces a line graph.

Most achievement and attitude variables in education tend to be uni-modal, i.e., with the largest number of values for the variable generally clustered in one place, often somewhere near the middle of the range. In such cases either the *Mean* or *Median* are appropriate (depending on the *Standard Deviation*).

What this illustrates is that before using any statistic to determine what the most representative score for the group as a whole is, it is important to look at a picture of the distribution of values for each variable, and for each group, to make sure that there are no *bi-modal* distributions or distributions that are badly *skewed* in some other way. The most commonly used statistics, such

[4] As an alternative, there may be cases where it makes sense to drop an outlier score. For example, if you find out that a student who did not answer any questions on the test has a parent die the day before, it would make sense to drop that students' score. Clearly, whatever reason is given for dropping an outlier has to be applied equally to each group, and should be applied with great caution and disclosed. In addition, the researcher should inform the reader of the scores both before and after an outlier is dropped.

as the *Mean*, assume that the distribution of each variable is *normal*, i.e., the distribution is close to a bell shaped curve. However, the *Median* or *Mode* statistics do not require a *normal* distribution; and are used because the distribution is in fact not close to *normal*.

Understanding the overall nature of the distribution of scores in your school(s), as well as in key subgroups, is a key aspect of leadership. Similarly, knowing the nature of the distribution of scores of the different groups is a critical element in designing and analyzing quantitative research. Among other things, the nature of the distribution is critical for selecting the statistics to be used and for determining the *practical significance* of the research. *Indeed, at its heart quantitative research is about analyzing the distribution of differences/variations between and within groups.*

ii. How to determine the actual/absolute (unweighted) performance of the experimental group.

As just discussed, the most important statistics for determining the absolute performance of the experimental group are its *Mean*, *Standard Deviation*, and perhaps *Median* of the outcome measure(s). Usually, the *Mean* is some type of score expressed as a grade, scale score (e.g., attitude scale), or percentile. However, the *Mean* can also represent the average percentage of students achieving a level, e.g., basic or proficient, percentage of students/schools achieving success/failure, or the probability of students/schools achieving success, e.g., probability of students being suspended or retained. [5]

Given that the mass of statistical analyses and data presentation generally deal with the comparative analysis between groups, you may have to wade through the article and lots of technical jargon to find the data that indicate the actual performance of the experimental group. The data on how the experimental students actually did is usually found in just a single sentence, or in a single line in one of the many tables in the article. *The bottom line is to <u>ignore</u> all the other statistics, data, and technical jargon in order to find the Mean or Median, to determine how the experimental students actually did.* This is akin to wading through the data about Greenland to find the single piece of info that indicates the <u>actual</u> temperature in January.

In theory it should be easy to find out how the experimental group actually did in published research since this is the most relevant data for leadership decision-making. However, in reality it is often difficult to determine how the experimental group did. There are two problems. The first problem is that the *Mean* or *Median* may not be provided (this will be discussed later). The second problem is that contrary to popular belief, published research rarely conducts analyses with actual scores. The raw performance data often undergo a variety of transformations to make the comparisons between experimental and comparison groups fairer. The two most common types of transformations are calculating 'weighted scores' and 'z scores'.

<u>Weighted Scores</u>.
The first type of data transformation occurs because there are often initial differences between the experimental and comparison groups at the start of an experiment that can bias the outcome—particularly where the groups were not formed via *random assignment*. Suppose, for example, that the *Mean* of the starting scores for the experimental group was lower than for the comparison group. Or suppose that there were more low-income, or special education students in the experimental group. In each case it can be argued that all

[5] Probabilities range from -1 to +1.

things being equal the experimental group is at a greater disadvantaged to make progress than the comparison group. When such initial differences exist, the final scores of the experimental group are weighted upwards in accordance with the degree of initial disadvantage.

The statistical procedure typically used to do such weighting is an *analysis of covariance*. How this statistic is calculated is not important to this discussion. What is important to understand is that in the example just discussed, this statistical procedure will bump up the final *Mean* of the experimental group. This means that the final *Mean* reported in the research is not the actual *Mean performance of the experimental group.*

Of course, if a school administrator bumped up students' scores he/she would go to jail. At the same time, this weighting of scores is legal statistically, and makes sense if the goal is to have a fair comparison between the groups. (Paradoxically, states sometimes weight the state scores of schools based on key demographics such as percentage free and reduced lunch, percentage of minority students, etc. Such weighting gives a sense of how the performance of low-SES schools would compare to more advantaged schools if all things were equal.)

However, for the purpose of *practical significance* we need to know the actual, *Unweighted Means* of the experimental group to compare to how your school(s) is already doing. The basic convention is that when an *analysis of covariance* is conducted the researcher should indicate specifically which variables were weighted and why, and then present both the final *Weighted* and *Unweighted Means*. For example it is possible that the *Weighted Means* could show that the experimental group did better, while the actual or *Unweighted Means* might show that in reality the comparison group did better, or did the same.

Unfortunately, a lot of the research I have been seeing lately present only the *Weighted Means*, and in one recent case I was not even sure from the description whether a weighting of scores occurred. For the purpose of determining the *practical significance* of the performance of the experimental group, only the actual *Unweighted Means* are appropriate.

In the absence of clear disclosure of the *Unweighted Means* and the rationale for weighting the scores, it is too easy for researchers with an agenda to find some way to weigh scores to have the data come out the way they want. One can always find some initial differences between experimental and comparison groups, and it is easy with modern technology to selectively decide on what to weight by seeing how weighting different variables can affect which group does better. For example, the weighting methods used in research about the effects of the 'Success for All' never weighted for the fact that more money was spent for the experimental schools—which would have bumped up the *Means* for the comparison schools. The academic reputation of a researcher is often enhanced by findings that favor the experimental group. Even if the researcher does not have a preference as to which group does better, it is in the interest of all researchers to find some difference. Not finding differences between groups reduces the likelihood that an article will get published, a phenomenon known as *publication bias*. Finally, as long as quality quantitative research is associated with increasingly sophisticated statistical procedures there is little incentive for researchers and editors to present such mundane data as *Unweighted Means* and *Standard Deviations*.

At the same time, I do think that most researchers approach their data with integrity. There are, however, notable exceptions that damage the profession and diminish educational

opportunity for the students that are most at-risk by hyping practices that are supposedly research-based, but that in reality have no *practical significance*. The only safeguard is for leaders to be more knowledgeable consumers of research.

z scores.

There is also a second common type of statistically legal transformation of actual scores. This occurs when there are differences in the outcome measures used within both the experimental and comparison groups. Suppose for example the study includes schools from different states in both the experimental and comparison groups. Each state currently has different tests, standards, cutoff points, and scoring protocols. How do you equate the results from different states?

The problem of equating test scores from different states, or any other different measures such as different attitude scales, is akin to comparing apples and oranges. Contrary to common belief, there are actually many ways to compare them. For example, you can compare the sugar content of each. The statistical way of doing such a comparison is by calculating z scores. A z score calculates where the actual score is in relation to the characteristics of the distribution of all scores. For example, suppose you go into a store and want to buy either an apple or an orange, and want to buy whichever is the better value. Just because one is cheaper does not mean that it is a better value. How good the value of each is in that store depends on what everyone else is charging for apples and oranges. So you can see what is being charged for an apple in this store in relation to what all other stores are charging for apples, and how much oranges in this store are being sold in relation to what other stores are charging for oranges. Once you know the distribution of prices for each across all stores, you can then find the *Mean* price charged for both apples and oranges in all stores. The *Mean* prices will probably be different for each type of fruit, but these values can now serve as a basis of value comparison. You can now determine where the price of the orange at the store you are in relative to the *Mean* price of oranges at all stores, and do the same for the apple. If, for example, the price of the orange at this store is less than the *Mean* price for oranges, while the price of the apple is higher than the *Mean* price for apples, then the orange has a lower z score and is the better value—even if the actual price of the orange at this store is higher than the cost of the apple.

While this shopping example is a bit contrived, the issue of relating scores across different tests and scoring systems in different states is a real problem, one that justifies converting actual scores to z scores to equate the results. So let's say that the *Mean* scale score for 5th graders is 256 (whatever that means), and that the average difference from the Mean is 20 points, then the Standard Deviation is 20 points. So a student who gets a score of 246 (whatever that means) is .5 of a *Standard Deviation* below the *Mean*, and that student would have a z score of -.5. Similarly, every student in the study would have their actual score converted to z scores. The same would be done with the test scores in the other state. The *Mean* and *Standard Deviation* of all students in State B will be different, but then a z score of -.5 of a given student in that state will represent his/her performance relative to all students the same in that state as that of the student in the other state. The statistical analysis will then be conducted on the z scores of each student in each group.

Of course, if the initial *Mean* z score of students in the experimental group is lower than that of experimental group, the *Mean* of the post-test scores of students in the experimental group will be weighted upwards. So now we have weighted z score results. While we can say

which group did better than the other based on such a calculation, what sense can school leaders, or any average person, make of such highly transformed data? How can leaders tell whether the experimental group did better than their students are already doing? The answer is that they cannot, unless the researchers happen to present the data in some more understandable way as to how the experimental group actually did in terms of unweighted scores that are recognizable.

At best, research with this highly transformed data will attempt to make it intelligible by including a statement such as: "… the experimental group did .15 Standard Deviation units better than the comparison group." It may also further clarify with a statement such as: "… this is the equivalent on a standardized reading test of students moving from 45th to the 54st percentile." However, this is not an actual difference. Rather, that is hypothesized difference if the students' scores in each of the groups was a normal distribution, with similar standard deviation, and a *Mean* that approximates the national average. That is highly unlikely for most schools seeking a BIG improvement. As a result, the 'clarifications' provided still do not tell you how the students in each group actually did, or how much of the hypothetical performance was the result of weighting scores.

While all the statistical manipulations just described are considered statisticallyvalid, they render the results essentially unusable for leadership decision-making. There is no way to determine the *practical significance* of such results. *So the fundamental problem is not that leaders and Ed.D. students do not understand the sophisticated statistical methodology and results that are presented in published research, it is that the methodology and results do not provide the information they need.* It should not be such a low priority to let non-researchers know how the experimental group actually did on an absolute basis with an *Unweighted Mean*, *Standard Deviation*, and *Median*.

iii. How to determine practical significance from Unweighted Means?

If a study does provide the actual *Means*, *Standard Deviations*, and *Medians* for the experimental group, how do you determine whether the outcome is of *practical significan*ce for your school/district? *In this case, there is no mathematical rule.* Like beauty, a *BIG* difference is in the eye of the beholder. Leaders have to decide on a target improvement goal for their school(s) that in their view would be big enough to warrant changing course.

For example, suppose a district is trying to reduce suspensions. A study is found demonstrating a technique wherein an experimental group of schools had 25% fewer suspensions than the comparison group of schools. Such a result would probably be *statistically significant*. However, in order to determine whether the results were of *practical significance* you need to know the *Unweighted Mean* number of suspensions for the experimental schools, and then compare them to yours. It may be that the suspension rates in your school(s) are already lower or about the same as those for the experimental group. In that case there is no incentive to adopt the suspension reduction technique in the study. How much lower should the suspension rate in any study be before leaders should consider adopting the technique used in the study for their schools? That is up to the leaders to decide, and this should be done before reviewing any studies.

The first step before reviewing any research is to set a goal for how much improvement in a given outcome(s) is sought over the next 2 years that would cause the district to adopt a given new approach. Keep in mind that the goals should be ones of substantial improvement expressed in

quantitative terms. However, these goals should represent incremental progress. For example, it is not reasonable to set a goal of moving from 300 suspensions a year to zero.

The process for determining *practical significance* from the *Unweighted Means*, *Standard Deviations*, and *Medians* are summarized in Table 3.6:

Table 3.6
Criteria for Determining Practical Significance

Criteria	Statistics
How did the students in the experimental group actually do?	Unweighted Mean Standard deviation Unweighted Median
How similar is the sample to your setting(s)	Leader judgment
Does the performance of the experimental group represent a BIG improvement over the current performance of your students?	Leader judgment of yes/no

Sorry, but there is no magical quantitative formula for making the decision in this case of what constitutes *practical significance*. This determination is left to the insight and initiative of leaders.

While this may disappoint those seeking a clearly objective quantitative way to decide if the *practical significance* of a study warrants the decision to adopt an intervention, that is an unrealistic expectation. Indeed, the need for human judgment at this critical point in the decision-making process is not unique to education—it is universal. Consider the use of quantitative data to help a woman make the difficult decision as to whether to undergo regular mammography tests. This is a case of dueling probabilities. There is the probability that if a woman does not undergo regular testing that she will incur an undiagnosed breast cancer that may kill her. On the other hand, there is the probability that the test will produce a false positive resulting in overtreatment, i.e., undergo a Mastectomy that is not needed. The role of quantitative research is to produce more accurate estimates of what these probabilities are. (At this point there is a good bit of debate as to the values of each probability.) The role of science is to develop better, more accurate tests, and to simultaneously develop better treatments to cure a wider variety of cancers. This changes the probabilities at both ends of the decision equation. So each women must make a very personal decision based on which risk she fears the most; i.e., the probability of dying from an undiagnosed cancer or receiving overtreatment using the best current estimate of each probability.

While the results of quantitative research are critical to the decision-making process, ultimately the making of the actual decision in all fields is ultimately based on personal judgment. Similarly, once an education leader knows that a research finding has *practical significance*, it becomes a judgment call as to whether there is sufficient *practical significance* to make a BIG difference for his/her school(s).

iii. How to determine practical significance if only Weighted Means are provided?

If only *Weighted Means* are provided be **very suspicious of the results**. There is probably an ulterior motive as to why the researcher did not provide the *Unweighted Means*.

iv. How to determine practical significance if only relative (ES) and correlational data are provided with no Means?

Suppose the study does not provide the *Means*, either weighted or unweighted, of how the experimental group did. What then? How to you determine practical significance if the only results provided are the relative *ES* differences between the groups, or data on the relationships between variables expressed as correlations or regressions? The simple answer is that under those circumstances you cannot determine whether the results have practical significance. All you can do is look for *potential practical significance.*

Determining the Potential Practical Significance of Research Evidence

…Under development…

Disconnect Between Evidence Provided by Researchers And That Needed for Leadership Decision-Making

Table 3.1 showed how Academicians/researchers ask different types of questions than educational leaders. Given the discussions of practical significance and potential practical significance, it is now clear that the evidence sought and/or needed by these different types of individuals is also different. This is illustrated in Table 3.7.

Table 3.7
Evidence Provided By Researchers and Evidence Needed for Practical Significance

	Questions Asked	Evidence Typically Provided	Evidence Needed for *Practical Significance* for Leadership Decision-Making
Academicians/Researchers	Did the experimental group in the research context do better than the comparison group, and was that difference *statistically significant*? What are the implications of the findings for theory?	*Statistical significance* of *Effect Size (ES)* between groups	
Leaders	If I adopt the approach used by the experimental group in the study is it likely to produce a BIG difference in my school(s)?		Was the sample similar in some ways to the population of my school(s), or at least to a significant subpopulation? If 'yes', how did the experimental group do in an absolute sense, i.e. the group's unweighted, Mean/Median results, and are those results substantially better than how my students—school(s) are doing?

C'mon, does it really make a real world difference if you rely on tests of *statistical significance*, or see *practical* or even *potential practical* significance?

Consider the following example. The What Works Clearinghouse (WWC) is the federal agency in the U.S. Department of Education (ED) charged with making educational research understandable to practitioners. (The 'wonderful' quality of this agency's work was critiqued earlier in this chapter.) The WWC recently released a report that summarized the research on the comparison of the effectiveness of traditional public schools and charter schools. Powers & Glass (2014) summarized WWC's findings as follows based on *statistical significance*:

WWC table 1

Statistically Significant Differences Between Charter Schools and Traditional Public Schools Across Five Studies

Study	Reading Gains for Charter School Students vs. Traditional Schools	Math Gains for Charter School Students vs. Traditional Schools
16 States	-	-
Indiana	+	+
National	+	-
New Jersey	+	+
New York	+	+

Based on statistically significant differences, it is 'clear' that charter schools did better in both reading and math. However, Powers & Glass (2014) then went a step further and used the data within the results to produce the table below that presents the evidence based on the *potential practical significance* of the studies.

WWC Table 2 Compiled by Powers & Glass (2014)

Effect Size Differences Between Charter Schools and Traditional Public Schools Across Five Studies

Study	Reading Gains for Charter School Students vs. Traditional Schools	Math Gains for Charter School Students vs. Traditional Schools
16 States	Charter – the equivalent of moving the median student from the 50th to slightly higher than the 49th percentile	Charter – the equivalent of moving the median student from the 50th to the 49th percentile
Indiana[ii]	Charter + 0.05 standard deviations higher. equivalent to moving the median student from the 50th to the 52rd percentile	Charter + 0.07 standard deviations higher, equivalent to moving the median student from the 50th to the 53rd percentile
National	Charter + 0.01 standard deviations higher	None
New Jersey	Charter + equivalent to moving the median student from the 50th to the 52rd percentile	Charter + equivalent to moving the median student from the 50th to the 53rd percentile
New York	Charter + 0.06 standard deviations higher, equivalent to moving the	Charter +

median student from the 50th to the 52rd percentile.	0.12 standard deviations higher, equivalent to moving the median student from the 50th to the 55th percentile

In WWC Table 2, he differences between the traditional public schools vs. charters are presented in terms of *Standard Deviation Units* and/or hypothesized, not actual, differences in percentiles.[6] None of these results meets the evidentiary criterion of *potential practical significance*. The only one that comes close is the math gains in NY. In addition, the tiny differences in the hypothesized percentile differences are so small as to indicate that:

- There are really no *potential practical significance* between the performance of charters vs. standard public schools,
- The small hypothesized percentile differences support the data from the *Standard Deviation Units* that there are no real substantive differences,
- The criterion of *potential practical significance* is a superior standard of evidence than *statistical significance*. The latter is misleading since small differences show up as significant if there is a large enough sample.

So under *statistical significance* charter schools are better. Using the more valid evidentiary criteria of *potential statistical significance* there is no practical difference between the two types of schools. So by presenting a dumbed down summary of incorrect criteria for evidence, the WWC is misleading the field, and doing an injustice to those working in traditional public schools. Clearly, the standard of evidence makes a huge real world difference.

Powers & Glass (2014) conclude that:

> *If summaries generated from research studies are intended to be useful guides to practitioners, they must provide a consistent and careful accounting of findings that allows them to assess their* ***practical importance****. Summaries of research that are hard to understand and misleading run the risk of eroding practitioners' trust in research...*[bold added]

The bottom line is, as stated earlier, is that it is critical for leaders to not rely on summaries of research, regardless of who puts them out. Instead, it is critical for leaders to conduct their own review.

Leadership Issues in Utilizing Research That Has Practical Significance

Suppose you find research about an intervention that has strong evidence of effectiveness in terms of *practical significance* or *potential practical significance*. Should you adopt it?

...to be continued…

[6] The results also indicate what difference that would make I terms of percentile scores. However, keep in mind, these are not the actual percentile scores. These are extrapolated scores if the results were normally distributed around the *Mean* and equal *Standard Deviations* which is highly unlikely.

BIBLIOGRAHY

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum. Hillsdale, NJ.

Joan I. Heller, J.I., Hanson, T., & Barnett-Clarke, C. (2010). The Impact of Math Pathways & Pitfalls on Students' Mathematics Achievement and Mathematical Language Development: A Study Conducted in Schools with High Concentrations of Latino/a Students and English Learners. A report prepared for the U.S. Department of Education. Institute of Education Sciences, Washington, D.C.

Muschkin, C.G., Glennie, E. & Beck, A.N. (2014). *Teachers College Record* Volume 116 Number 4, http://www.tcrecord.org ID Number: 17405, Date Accessed: 3/10/2014 5:55:58 PM

Pogrow, S. (1998). What is an Exemplary Program and Why Should Anyone Care? A Reaction to Slavin and Klein. *Educational Researcher*, October 1998, pp. 22-29.

Pogrow, S. (2000). The Unsubstantiated 'Success' of Success for All. Implications for Policy, Practice, and the Soul of the Profession. *Phi Delta Kappan*, April, pp. 596-600.

Powers, J.M. and Glass, G.V. (2014). When Statistical Significance Hides More Than it Reveals, *Teachers College Record*. ID Number: 17591, Downloaded: 7/10/2014 from http://www.tcrecord.org.

Thompson, B. (2005). Foundations of Behavioral Statistics: An Insight Based Approach. The Guilford Press, New York.

Venezky, R. L. (1998). An alternative perspective on Success for All. Advances in educational policy, Wong, K. (ed.) Greenwich, CN: JAI Press. 4, 145-165.