Empowering Educational Consumers to Analyze Educational Assessment Data: The Educational Impact Calculator (EIC)

Technical Report 2016-2

Jean Stockard, Ph.D., Director of Research and Evaluation

March 24, 2016

Table of Contents

List of Tables

List of Figures

Executive Summary

NIFDI's Educational Impact Calculator (EIC) is designed to help educational consumers analyze publicly available data on student achievement, such as information that is often disseminated by state departments of education and school officials. The web-based calculator uses aggregate level data to answer five general questions about a school or district:

- Are students at one school doing better (or worse) than those in another school?
- Are they doing better or worse than those in the district, the state, or the nation?
- Are they doing better this year than last year?
- Are changes in one school different than changes in other schools?
- Are changes in one school different than changes in the district, the state, or the nation?

The same questions can be asked about aggregate achievement of a classroom or a school district.

Input data for the EIC may be the percentage of students reaching a given benchmark, group averages and standard deviations, or the percentile rank of the average student. Output statistics include effect size, improvement index, and, if users know the number of students tested, the probability that results would occur by chance. These statistics can help educational consumers determine if trends in achievement patterns in a school or district meet the criteria that researchers typically use to denote educationally important and statistically significant results.

The body of this Report provides background to help users of the EIC. The first two sections discuss terminology and describe the structure of the EIC. The following five sections give examples of the use of the EIC to answer the queries listed above. Examples use each of the possible types of input data and different types of groups (e.g. classrooms, schools, and districts). A final section discusses ways in which the EIC could potentially help students and schools and provides cautions regarding its use. An extensive appendix explains the underlying research designs and gives the equations used in the statistical analyses. The techniques are identical to those covered in introductory college level statistics courses. While they are not complex, they are fully sufficient for answering the questions that are generally of most concern to educational consumers.

Empowering Educational Consumers to Analyze Educational Assessment Data: The Educational Impact Calculator (EIC)

The No Child Left Behind (NCLB) Act and its requirements for routine student assessment have resulted in an educational system that teems with data. Each year, educational consumers – teachers, school administrators, policy makers, and parents – receive reports of the progress of their students on state assessments and, often, other tests. There are many questions they may want to answer, such as:

- Are students at our school doing better (or worse) than those in other schools?
- Are they doing better or worse than those in the district, the state, or the nation?
- Are they doing better this year than last year?
- Are changes in my school different than changes in other schools?
- Are changes in my school different than changes in the district, the state, or the nation?

Similar questions could be asked about achievement patterns in a classroom, comparing results to other classrooms, the school or district. They could also be asked about achievement in a district, comparing results to other districts or to the state or nation.

Educational consumers can easily see general patterns that address these questions. Yet, wise educational consumers want to know how strong the differences are? Would they be considered "educationally significant?" Could they have occurred by chance? The answer to these questions are important for helping to decide if teachers, schools, or entire districts should consider changing programs or procedures or if they should continue with their present activities.

The research community has well developed methods and criteria to answer these questions, but most educational consumers are far from comfortable with the underlying calculations and statistical procedures. They may be left wondering what the assessment results mean and feel that they are at the mercy of researchers and other "experts" to interpret the data.

NIFDI's Educational Impact Calculator (EIC) is designed to help break the barrier between the "experts" and the consumers, empowering those who are closest to the data to answer questions regarding assessment results. Specifically, the EIC provides a simple way to answer questions about assessment results using data that are routinely given to school officials, posted on state department of education websites, and sent to parents. It helps educational consumers use these publicly available data to determine if changing achievement patterns in

their school would meet the usual criteria that researchers use to denote educationally important and statistically significant results.

This Technical Report provides background to help users of the EIC. The first section provides a brief discussion of the terminology used, and the second provides an overview of the structure of the EIC. The following five sections give examples of the use of the EIC to answer the questions noted above. Three types of data can be used: 1) the percentage of students reaching a given benchmark or standard, 2) means and standard deviations, or 3) percentile ranks that correspond to scores of the average student. A final section returns to the issue of empowering educational consumers, discussing the ways in which these procedures could potentially help students and schools. It also includes cautions regarding their use.

While the discussion in the body of this report is at a relatively general level, an appendix provides explanations of the underlying research designs and computations and more detailed explanations of the various terms and concepts involved. Readers will find that the techniques are identical to those covered in introductory college level statistics courses. In other words, the logic and statistics involved are not complex. However, they are fully sufficient for answering the questions that are generally of most concern to educational consumers.

Input and Output Data and Terminology

The EIC requires users to input data on achievement and then calculates three types of output statistics: effect sizes, the probability that a result would occur by chance, and an "improvement index." The analyses of changes over time also involve the concept of a cohort.

Input Data

Reports on assessment data that appear in the media and are sent to school administrators typically include information about a given school or district, a state and even, for some assessments, the nation.

One type of information often given to consumers is the percentage of students who score at and above (or below) certain proficiency levels or benchmarks. This type of data is often included with curriculum based measures such as DIBELS or AIMSweb or with annual tests administered by state departments of education. Occasionally consumers also might be interested in the percentage of students scoring at or above a given percentile. Examples with the use of benchmark data are given in Examples 1, 4, 7, 10 and 13 below.

Sometimes consumers are given information on the mean, or average, score obtained by a group of students, and results with this type of data are in Examples 2, 5, 8, 11, and 14. For

some analyses using average scores users also need to know the standard deviation. For instance, when comparing results for a school to a larger entity, such as a national norming sample, the standard deviation for the school is not needed (see Examples 6 and 14). The standard deviation for the norming sample is sufficient for completing the calculations. When comparing average scores between two groups of the same type (e.g. two schools or two districts, as in Examples 2 and 11), the user needs to know the standard deviation of both groups.

The third type of data used by the EIC is the percentile rank of the average student (see Examples 3, 6, 9, 12, and 15). The wording "percentile rank associated with the average score" is important because, technically, percentiles should not be averaged. However, raw scores (or their equivalents) can be averaged and translated to percentiles. In addition, percentiles can be translated to Normal Curve Equivalent (NCE) scores, which can be used in analyses. As explained in the appendix to this report, the EIC incorporates the appropriate translations needed for calculations. The reports supplied to consumers by testing companies typically report the "percentage associated with the average score" and do not average the percentiles. Thus, these results can be used in the spreadsheets. Users are advised, however, to use the two other types of input data rather than percentiles if these other data are available.

To use the EIC it is not necessary to know the sample size. However, if this information is inputted the EIC will report the probability that a result occurs by chance.

Statistical Output

The first type of output given by the EIC is an effect size. Researchers often use effect sizes to describe the magnitude of a result. Technically, an effect size describes the magnitude of a difference between two groups in standard deviation terms. A value of zero indicates no difference, while larger absolute values (i.e., either positive or negative) indicate more of a difference. A value of positive one (+1.00) indicates that a target group had scores that were one standard deviation larger than the comparison group; while a value of negative one (-1.0) indicates that a target group's scores were one standard deviation lower than the comparison. An effect size of .50 indicates a difference of one-half of a standard deviation, etc. Within the field of education, effect sizes of .25 or larger have traditionally been considered educationally significant (Tallmadge, 1977). It should be noted, however, that the effect sizes associated with a strong curriculum are generally substantially larger. The average effect size associated with implementations of Direct Instruction is estimated to be to be well over twice the .25 level.[1] In

[1] Hattie analyzed the results of four meta-analyses that included Direct Instruction (DI), incorporating 304 studies, 597 effects and over 42,000 students and found an average effect size of .59. Stockard (2013) used methods like those described in this paper to examine assessment data from 18 different sites using the DI curriculum and

addition, the criterion of .25 should be seen as a touchstone or helpful guide to interpreting results. There is no magic associated with this particular number. It just provides a useful signpost, and examples of interpretations of results are included in the sections below.

The second output statistic is the "improvement index," which translates the effect size into percentiles. The number tells the difference between the percentile rank of an average student in a user's group and the percentile rank of an average student in the comparison group. Like the effect size, the improvement index can be positive or negative, depending upon whether the target group had better results or worse results than the comparison group.

The final output statistic is the probability that the results would have occurred by chance. The measure of statistical significance should always be interpreted cautiously, primarily because calculations of statistical significance are highly influenced by the number of students in a comparison. With large samples relatively small differences will be statistically significant; with small samples relatively large differences will not be significant. In contrast, the effect size and improvement index are not affected by the number of students in a comparison. They remain the same no matter how many students are included. As a result they are easier to compare from one situation to another and are often more useful for educational consumers. Traditionally probability values of .05 or less have been considered statistically significant. However, when looking at results in real life settings and, especially with very large or very small samples, this value should only be considered a general point of reference. Users who are not familiar with the notion of probability should use this output statistic cautiously.

Comparing Cohorts

One additional bit of terminology can be useful – the notion of cohorts. As groups of students move through the grades they are called cohorts. As described more fully in the appendix, because they generally move through school together these cohorts are, in statistical jargon, "independent" of each other. This simply means that the cohorts are discrete entities, with very little movement from one group to another. In addition, in most schools the composition of student cohorts is quite similar from one year to the next. The proportion of students at risk, often measured by the percentage receiving free or reduced lunch and/or students' entry level skills, generally varies only slightly across time. More important, when there are variations, teachers and administrators almost always know about it and can alert users to these differences. In statistical terms, the high similarity of one cohort to another is called "equivalence." Because cohorts are independent and equivalent they can be compared across

found an average effect size of .56, slightly smaller than the value reported by Hattie, but still more than twice the level used by Tallmadge.

time in ways that are statistically valid. When analyzing changes over time with the EIC users compare the achievement of different cohorts (Queries 3, 4, and 5, examples 7 to 15).

Choosing an Approach

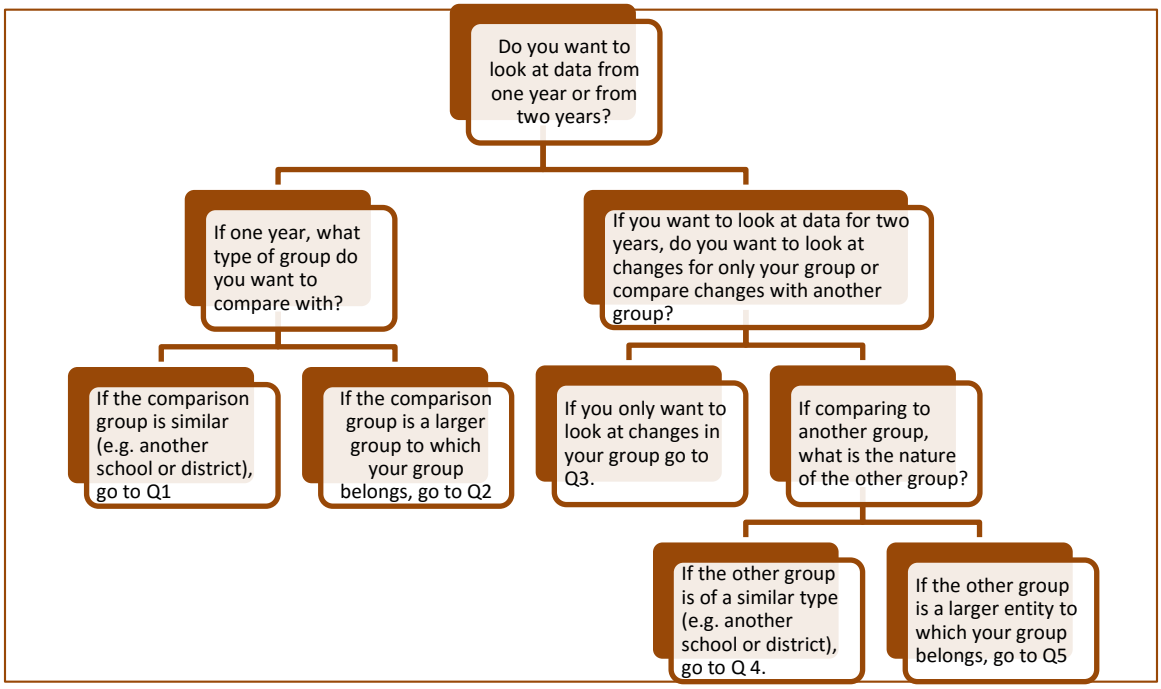
The next five sections provides example of using the EIC to answer the five general queries outlined at the start of this document:

1. Are students in my classroom (school or district) doing better (or worse) than those in other classrooms (schools or districts)? (Examples 1 to 3)
2. Are students in my group doing better or worse than those in a larger group to which they below, such as the district, the state, or the nation? (Examples 4 to 6)
3. Are my students doing better this year than in a previous year? (Examples 7 to 9)
4. Are changes in my classroom (school or district) different than the changes in another classroom (school or district)? (Examples 10 to 12)
5. Are changes in my classroom (school or district) different than the changes in a larger group (school, district, state or nation)? (Examples 13 to 15)

The first two queries involve comparisons at just one time point, while the next three involve comparisons over time (between cohorts). Each section includes examples with the three possible types of input data: the percentage of students scoring at benchmark, average (mean) scores, and the percentile rank associated with the score of the average student.

Figure One is a decision tree that illustrates the difference in the five possible questions and when they would be appropriate. Users who only wish to examine data from one year would go to the left-hand side of the diagram. Those who wish to compare results to those in a similar group, such as another school or district, would use the part of the calculator associated with Query One. Those who wish to compare to a larger group in which theirs is embedded (e.g. their school to the district or their district to the state) would go to Query Two. Those who want to compare data from two different years (two cohorts) would go to the right side of the decision tree. If they only wanted to look at changes within their own group they would go to Query Three. If they wanted to compare changes in their group with those in another they would go to Query Four or Query Five. The former would be appropriate if the comparison group were another school or district. The latter would be appropriate if the comparison group were a larger entity to which their group belonged.

Figure One
Decision Tree: Choosing the Appropriate Part of the EIC for A Comparison



Within each query users then select the part of the EIC that matches their available data: 1) the percentage at benchmark, 2) means and standard deviations, or 3) the percentile rank corresponding to the score of the average student. Separate calculation sheets are available for those who know the number of students who were tested and for those who do not have that information. The following sections give examples of the use of each portion of the EIC.

Query One: Comparing Results in One Group with Those in Another

One of the first questions a consumer may ask is, "How does the performance of students at my school compare with the performance of students in a similar group?" For instance, one might want to compare performance of students in one school with those in a nearby school that serves students with very similar background characteristics but uses a different curriculum. Or one might want to compare scores of students in one district with those in a nearby district. A wise consumer would want to know if any differences were large enough to be educationally important. Examples are given below comparing two classrooms, two schools, and two districts. Examples use each of the three types of data.

Example One: Comparing the Percentage of Students at Benchmark in Two Schools

Principal Mary Brown wanted to compare the achievement of students in her school with the achievement of students in a nearby school. Fifty percent of the students in fifth grade at her school were rated as proficient on the state assessment, while 65 percent of the fifth graders in the nearby school were rated as proficient. One hundred students were tested at Principal Brown's school and 120 students were tested at the nearby school. Clearly the students at Principal Brown's school did not do as well as those at the other school, but was this difference large enough to be considered educationally significant? Would it be considered statistically significant?

The data were entered into the EIC, as shown in the first part of Table One. The results produced by the EIC are in the second part of the table. The effect size of -.31 is beyond the level of .25 generally seen as educationally significant. The improvement index, which translates the effect size into percentile terms, indicates that the average student in Principal Brown's school scored 12 percentile ranks lower than the average student in the comparison school. The value of .02 in the final line of results indicates that a difference as large as that between Principal Brown's school and the other school would occur by chance only 2 times out of 100. In other words, it is quite unlikely that the results were a fluke. Taken together, these results suggest that Principal Brown would be wise to be concerned about her students' achievement and consider corrective action.

Table One	
Example One: Comparing Two Schools, Percent at Benchmark	
Data Entered for Example One	
Data for Your Group	
Percentage for your group	50
Number of students tested for your group	100
Data for the comparison group	
Percentage for the comparison group	65
Number of students tested for the comparison group	120
Results from Example One	
Effect Size	-0.31
Improvement Index	-12.1
Probability this effect would occur by chance.	0.02

Example Two: Comparing Means and Standard Deviations of Scores in Two Districts

Superintendent Paul Johnson had data on student achievement on a standardized achievement test for students in his district and in a nearby district with similar demographic characteristics. The average (mean) score of his students was 110 with a standard deviation of 15. The average in the other district was 107, with a standard deviation of 14. One hundred fifty students had been tested in both districts. Clearly Superintendent Johnson's students scored higher than those in the other district. But was this difference large enough to be considered educationally significant? Could it have just appeared by chance?

To answer this question Superintendent Johnson could enter the data into the EIC, as shown in the top panel of Table 2. Note that for this comparison Superintendent Johnson needed to know the mean and standard deviation for each group. The number of students is not necessary for calculating the effect size and information index. It is only needed if one wants to know the probability the result would occur by chance.

The results are shown in the bottom panel of Table 2. The effect size of .21 is close to the level typically deemed educationally significant and corresponds to a difference of 8 percentile ranks between average students in the two districts. The probability that differences this large would occur by chance is only 7 out of 100. Many would suggest that, based on these results, Superintendent Johnson was entitled to be quite proud of the accomplishments of his students relative to those in the nearby district.

Table Two	
Example Two: Comparing Two Districts: Average Scores	
Data Entered for Example Two	
Data for Your Group	
Mean (Average) score	110
Standard deviation	15
Number of students tested (if available)	150
Data for the comparison group	
Mean (Average) score	107
Standard deviation	14
Number of students tested (if available)	150
Results from Example Two	
Effect Size	0.21
Improvement Index	8.2
Probability this effect would occur by chance	0.07

Example Three: Comparing Percentile Ranks of the Average Student in Two Classrooms

Principal Margaret White was interested in differences in scores of students in two third grade classrooms in her school. The students had been randomly assigned to teachers at the beginning of the school year, and they had very similar skills at that point. Each classroom had 25 students. Yet at the end of the school year the scores in Classroom A, where the percentile rank of the average student was 66, seemed markedly lower than the scores in Classroom B, where the percentile rank of the average student was 78. Principal White wondered if this difference was large enough to be considered educationally significant or if it could have just occurred by chance. To answer that question data were entered into the EIC as shown in the top part of Table 3. The results obtained are shown in the bottom part of Table 3.[2]

The results from the EIC confirm Principal White's concerns. The effect size of -.36 would be seen as indicating that the gap between classroom A and classroom B is educationally significant. The probability level of .20 is above the .05 cut-off that is often used, but that no doubt reflects the relatively small number of students in each group. Principal White would probably want to consider corrective actions.

Table Three	
Example Three: Comparing Two Classrooms, Percentile of the Average Student	
Data Entered for Example Three	
Data for Your Group	
Percentile of average score	66
Number of students tested (if available)	25
Data for the Comparison Group	
Percentile of average score	78
Number of students tested (if available)	25
Results from Example Three	
Effect Size	-0.36
Improvement Index	-14
Probability this effect would occur by chance	0.20

[2] Careful readers will note that the improvement index is not equivalent to difference of the two percentiles used as input data. That occurs because of the differences between percentile scores and the NCE scores used in calculations. One could argue that the Improvement Index calculated with NCE scores provides a more accurate estimate of the effects than simple comparisons of the percentiles. (See appendix for more details on calculations.)

Query Two: Comparing Results in One Group to Those in a Larger Group

In addition to comparing results from one group to another similar group, as was shown in Examples 1, 2, and 3, consumers might want to compare results regarding students in one group to a larger group to which they belong. For instance, one might want to compare results of a school to the district, of a district to the state, or a classroom to the school. Often the reports provided by testing agencies contain such comparative information. This section includes examples of these comparisons with the three types of input data. Note that to obtain tests of statistical significance with these analyses users only need to know the number of students tested within their own group, not in the larger comparison group.

Example Four: Comparing the Percentage of Students at Benchmark in a School to the Percentage at Benchmark in the District

The Chair of Central School District's School Board had information for each school in the district on the percentage of fourth grade students who scored at benchmark on the state assessment and the percentage for the district as a whole. She was especially concerned about the scores for Elm Elementary. At that school 55 percent of the students scored at benchmark in contrast to 60 percent of the fourth graders in the district as a whole. Fifty students were tested at Elm Elementary.

Table Four	
Example Four: Comparing Percent at Benchmark in One School to the Percent at Benchmark in the District	
Data Entered for Example Four	
Data for Your Group	
Percentage of students meeting benchmark	55
Number of students tested (if available)	50
Data for the Larger Group	
Percentage of students meeting benchmark	60
Results from Example Four	
Effect Size	-0.10
Improvement Index	-4.1
Probability this effect would occur by chance	0.47

The data the Chair entered into the EIC are shown in the top panel of Table 4 and the results are in the bottom panel. The effect size of -.10 does not reach the .25 criterion of educationally important and there is almost a 50 percent chance that the difference between Elm Elementary and the district could have appeared by chance. Thus, unless these results were part of a

recurring pattern appearing with other grades and years, the Chair would probably decide to wait before pursuing further action.

Example Five: Comparing Means and Standard Deviations of Scores in One School to National Norms

Principal Evans, of Central High School, knew that the tenth graders in his school had scored above the national average on a nationally normed achievement test. But, he wondered, was the difference large enough to be considered educationally significant? The average for the Central students was 105, while the national average was 100. The standard deviation for the nation was 15. Fifty Central students had taken the test.

Principal Evans could enter this information into the EIC, as shown in the top part of Table 5. (Note that he only needed to know the standard deviation for the larger group, not his school.) The results are shown in the bottom panel of the table and reinforce Principal Evan's pride in his students. The effect size of .33 is well beyond the level traditionally used to indicate educational significance, and the probability that the result would occur by chance is only two out of 100. The improvement index of 13 indicates that that the average student at Central High had a score that was 13 percentile ranks higher than the average student in the nation.

Table Five	
Example Five: Comparing Average Scores in One School to the National Average	
Data Entered for Example Five	
Data for Your Group	
Average score of students in your group	105
Number of students tested (if available)	50
Data for the Larger Group	
Average score of students in the larger group	100
Standard deviation of scores in the larger group	15
Results from Example Five	
Effect Size	0.33
Improvement Index	13.1
Probability this effect would occur by chance	0.02

Example Six: Comparing the Percentile Rank of the Average Student in a District to the Percentile Rank of the Average Student in the State

Superintendent Jensen had recently accepted a position in Seacoast district. She knew that the district had a history of achievement problems, but wanted to understand the issue in greater detail. She was especially interested in knowing if the achievement of Seacoast's students was significantly lower than the achievement of other students in the state. In the previous year the average Seacoast third grader scored at the 35th percentile on the state assessment, while the average student in the state scored (by definition) at the 50th percentile. One hundred fifty Seacoast third graders had been tested.

Superintendent Jensen could enter this information into the EIC, as shown in the top panel of Table 6. The results are shown in the bottom panel. The effect size of -.39 would be considered educationally significant. The probability value of 0.000 reflects less than one out of 1000 (<.001) and indicates that an effect size of this magnitude would occur very rarely. The improvement index of -.15 shows that the average Seacoast third grader scored 15 percentile points below the average student in the state. Superintendent Jensen would no doubt conclude that there were indeed serious achievement problems in Seacoast District.

Table Six	
Example Six: Comparing Percentile Rank of the Average Student in a District to the Percentile Rank of the Average Student in the State	
Data Entered for Example Six	
Data for Your Group	
Percentile of average score	35
Number of students tested (if available)	150
Data for the Larger Group	
Percentile of average score	50
Results for Example Six	
Effect Size	-0.39
Improvement Index	-15
Probability this effect would occur by chance	0.000

Query Three: Comparing Results from One Year (Cohort) to Another

Consumers often want to know about changes in assessment scores over time. Given the requirements of legislation such as NCLB they want to know if students are doing better now than in previous years. They also often want to know the impact of a new curriculum. What difference does an implementation have on students' achievement? Would changes be

considered educationally important? To answer this question, consumers want to compare the achievement of cohorts with different educational experiences. Three examples are given below, each with a different type of data.

Example Seven: Comparing the Percentage of Students at Benchmark in a School in One Year to the Percentage Four Years Later

Oak Elementary began using *Reading Mastery* in grades K-3 in the fall of 2012. Teachers felt that their students were doing better after exposure to the program, but wanted to find out if changes in state assessment scores were educationally significant. In Spring 2016, four years after beginning the new curriculum, 65 percent of Oak Elementary third graders scored at the proficient or advanced level on the state assessment. In Spring 2012, before starting the new curriculum, 40 percent of Oak Elementary third graders scored at that level. Eighty-five students were tested in 2016 and 70 students were tested in 2012.

Table Seven	
Example Seven: Comparing Percent at Benchmark in One Year to the Percent at Benchmark in Another Year	
Data Entered for Example Seven	
Data for the more recent year	
Percentage of students at benchmark	65
Number of students	85
Data for the comparison year	
Percentage of students at benchmark	40
Number of students	70
Results for Example Seven	
Effect Size	0.52
Improvement Index	19.8
Probability this effect would occur by chance	0.001

The information Oak Elementary teachers entered into the EIC is shown in the top panel of Table Seven, and the results are shown in the bottom panel. Note that for this query the EIC specifies the lines in which data for each year should be entered. The data for the more recent year are entered first, followed by the data for the later year. The results clearly support the teachers' impression of higher achievement. The effect size of .52 is far beyond the level considered educationally significant (and, in fact, just slightly less than the average effect size associated with the use of *Reading Mastery* that is reported in the research literature). The probability of .001 indicates that increases of this magnitude would occur by chance only once out of 1,000 times. The Improvement Index is also impressive, indicating that the average third

grader at Oak Elementary in 2016 scored almost 20 percentile ranks higher than the average third grader in 2012.

Example Eight: Comparing the Average Scores of Students in One Cohort to the Average in a Later Cohort

When a new superintendent came to Mountain View School District he mandated the use of "balanced literacy" and "whole language" programs throughout the elementary schools. After two years of this new approach, some school board members became concerned about rumors of lowered achievement. They asked the Superintendent for data on reading achievement of first graders over the last five years, and he provided information on the average and standard deviation of first graders on a curriculum-based measurement (e.g. DIBELS or AIMSWeb). As the school board members suspected, scores had declined after the change in reading programs. The Superintendent insisted that the change was simply due to chance and wasn't significant. The school board members used the EIC to test that assertion.

In the current year, two years after the curriculum change, the average reading composite score of district first graders was 58, with a standard deviation of 16. In contrast, two years earlier and before the change the average score was 65, with a standard deviation of 12. Three hundred students were tested in each year. The top part of Table Eight shows the data that were entered into the EIC, and the bottom part shows the results.

Table Eight	
Example Eight: Comparing Average Scores in One Year to Average Scores in Another Year	
Data Entered for Example Eight	
Enter the data for the more recent year	
a) Average (mean) score	58
b) Standard deviation	16
c) Number of students	300
Enter the data for the comparison year	
a) Average (mean) score	65
b) Standard deviation	12
c) Number of students	300
Results for Example Eight	
Effect Size	-0.50
Improvement Index	-19.1
Probability this effect would occur by chance	0.000

The results indicate that the school board members had good reason to be concerned. The effect size of -.50 indicates that the decline in first graders' reading skills since the curriculum was instituted was educationally significant. The probability level shows that this result was very unlikely to have occurred by chance (a probability of less than one in 1,000). The average first grader in the current year had scores that were 19 percentile ranks lower than the average first grader two years earlier using the previous curriculum.

Example Nine: Comparing the Percentile Rank of the Average Student in One Cohort to the Percentile Rank of the Average Student in a Later Cohort

Parents of students in Valley View Elementary were concerned that the achievement of students had declined over the last few years. From looking at past communications they saw that in 2012 the average fifth grader scored at the 58th percentile rank on a nationally normed test. But in 2014 the average fifth grader scored at the 54th percentile rank. Was this decline large enough to be considered educationally significant? There were 85 students in the fifth grade in each year.

The parents entered this information into the EIC, as shown in the top panel of Table Nine. The results are shown in the bottom panel. The effect size is -.10, below the level typically seen as educationally significant, and the probability level of .51 indicates that the result was likely to have occurred by chance. Based on these results one could suggest that the parents need not be overly worried at this point about declining achievement, although they certainly would be advised to continue to monitor the students' achievement.

Table 9

Changes Over Time: Comparing Scores in One Cohort with Scores of Another Cohort - Percentile Rank of the Average Score, Sample Size Known

Enter the data for the more recent year	
a) Percentile of the average score	54
b) Number of students tested for the more recent year	85
Enter the data for the comparison year	
a) Percentile of the average score	58
b) Number of students in the comparison year	85
Results	
Effect Size	-0.10
Improvement Index	-4.0
Probability this effect would occur by chance	0.51

Query Four: Comparing Changes from One Year to Another in One Group to the Changes in a Group of the Same Type

The comparisons between years shown in Examples 7, 8 and 9 are certainly informative. But wise consumers might have another question: Are the changes seen over time in my school (or district) greater than those that may have occurred in another school (or district)? Are the changes at my school (or district) large enough, when compared to the changes in the other group, to be considered educationally important? As with Examples 7 to 9, these comparisons involve cohorts, two groups that are independent of each other but passing through the same organization, such as fourth graders in one year and fourth graders in another year. The examples given in this section build on the data given in the comparisons of data between two groups described in the examples associated with query one.

Example Ten: Comparing Changes in the Percentage of Students at Benchmark in Two Schools

Example One described Principal Brown's comparison of scores of her fifth graders on the state proficiency test to scores of the fifth graders in a nearby school. After discovering that her students scored significantly lower than others she and her staff worked diligently to change the situation. Three years later 70 percent of the fifth graders at her school scored at the proficient level, compared to only 50 percent at the earlier testing. At the same time the scores also rose at the nearby comparison school – from 65 percent to 75 percent proficient. Principal Brown and her staff knew that their students were doing better, but was the change enough greater than the change in the nearby school to be considered educationally or statistically significant?

Table 10 shows the data that Principal Brown entered into the EIC and the results that were obtained. Note that data are entered first for the two cohorts for the user's group and then the data are entered for the comparison group. Data for the more recent year are entered first. Principal Brown and her staff would, no doubt, be gratified by the results. The positive effect size of .20 shows that fifth graders in her school had improved one-fifth (20%) of a standard deviation more than those in the comparison school. This is equivalent to a change, for the average student, of almost eight percentile ranks and would occur by chance only 4 percent of the time.

Table Ten	
Comparing Changes in the Percentage of Students at Benchmark in One Group with Changes in a Larger Group	
Data Entered for Example Ten	
Enter the data for your group	
a) Percentage for the more recent year	70
b) Number of students tested in the more recent year (if available)	105
c) Percentage for the comparison year	50
d) Number of students tested in the comparison year (if available)	100
Now enter the data for the Comparison Group	
a) Percentage for the more recent year	75
b) Number of students tested in the more recent year (if available)	125
c) Percentage for the comparison year	65
d) Number of students tested in the comparison year (if available)	120
Results for Example Ten	
Effect Size	0.20
Improvement Index	7.7
Probability this effect would occur by chance	0.04

Example Eleven: Comparing Changes in Average Scores of Students in Two Districts

Example two describes how Superintendent Johnson compared standardized achievement test scores for students in his district to those in another district. A few years later Superintendent Johnson found that the scores in his district had fallen, from an average of 110 to 105. Scores in the comparison district had also fallen, but only by one point (from an average of 107 to 106). (One hundred fifty students were tested in each district in each year.) Was this difference large enough to be seen as educationally significant or statistically significant?

Table 11 shows the data that Superintendent Johnson could enter in the EIC to answer this question. Data for the user's group are entered first, followed by data for the comparison group, and for each group data for the more recent year are entered first. The results indicate that Superintendent Johnson would be wise to worry about the results. Relative to the other district, his students' achievement had declined by .28 of a standard deviation, a decline that is equivalent, for the average student, to 11 percentile ranks. The probability value indicates that these results would be very unlikely to have occurred by chance (only once out of 1,000).

Table Eleven	
Comparing Changes in the Average Scores of Students in One Group with Changes in Another Group	
Data Entered for Example Eleven	
Enter the data for your group	
a) Mean (Average) for the more recent year	105
b) Standard deviation for the more recent year	16
c) Number of students tested for the more recent year	150
d) Mean (Average) for the comparison year	110
e) Standard deviation for the comparison year	15
f) Number of students tested for the comparison year	150
Enter the data for the other group	
a) Mean (Average) for the more recent year	106
b) Standard deviation for the more recent year	13
c) Number of students tested for the more recent year	150
d) Mean (Average) for the comparison year	107
e) Standard deviation for the comparison year	14
f) Number of students tested for the comparison year	150
Results for Example Eleven	
Effect Size	-0.28
Improvement Index	-10.9
Probability this effect would occur by chance	0

Example Twelve: Comparing Changes over Time in Percentile Rank of the Average Student in Two Classrooms

Example Three described Principal White's comparison of scores of students in two classrooms. Students had been randomly assigned at the beginning of the year, but at the end of the year those in Classroom A had markedly lower scores, a difference that was large enough to be seen as educationally significant (an effect size of -.36, see Table 3). Given those results Principal White worked with the teacher in Classroom A to help her improve her skills. At the end of the following year, Principal White was gratified to find that the average student in Classroom A was now doing much better than in the previous year, with a percentile rank of 75. The average student in Classroom B scored just slightly higher than the average student in the previous cohort, with a percentile rank of 79.

Principal White could use the EIC to examine the gains made in Classroom A relative to the gains made in Classroom B. Table 12 shows the data that she would enter and the results. Both Principal White and the teacher in Classroom A would no doubt feel gratified by the findings.

The effect size of .23 shows that the improvement in Classroom A from the previous year was almost a quarter of a standard deviation greater than the change in Classroom B. This difference corresponds to a difference of nine percentile ranks. Thus, while the students in Classroom A were still not scoring at equivalent levels to those in Classroom B, differences for the current cohort were much smaller than for the previous cohort. (The results of the probability line of the EIC no doubt reflect the relatively small samples involved.)

Table Twelve

Comparing Changes in the Percentile Rank of the Average Student in One Group to Changes in the Percentile Rank of the Average Student in Another Group	
Data Entered for Example Twelve	
Enter the data for your group	
Enter the data for the more recent year	
a) Percentile of the average score	75
b) Number of students tested for the more recent year	24
Enter the data for the comparison year	
c) Percentile of the average score	66
d) Number of students in the comparison year	25
Enter the data for the other group	
a) Percentile of the average score	79
b) Number of students tested for the more recent year	26
Enter the data for the comparison year	
c) Percentile of the average score	78
d) Number of students in the comparison year	25
Results for Example Twelve	
Effect Size	0.23
Improvement Index	9.0
Probability this effect would occur by chance	0.26

Query Five: Comparing Changes from One Year to Another in One Group to the Changes from One Year to Another in a Larger Group

Examples 7 to 9 looked at changes over time in just one group, and Examples 10 to 12 compared these changes to another group of similar size. But wise consumers might have yet another question: Are the changes seen over time in my school (or district) greater than those that may have occurred in the larger group to which my school (or district) belongs? Are the changes at my school (or district) large enough, when compared to the changes in the total group, to be considered educationally significant? This question is especially important to consider when both the target group and the larger group have been the focus of improvement efforts. As with Examples 7 to 12, these comparisons involve cohorts, two groups that are independent of each other but passing through the same organization, such as fourth graders in one year and fourth graders in another year.

Example Thirteen: Comparing Changes in the Percentage of Students at Benchmark in One Group with Changes in a Larger Group

As described in Example Seven above, Oak Elementary began using *Reading Mastery* in grades K-3 in the fall of 2012. In Spring 2012, before starting the new curriculum 40 percent of Oak Elementary third graders scored at the proficient or advanced level on the state assessment. In Spring 2016 65 percent of Oak Elementary third graders scored at that level. Seventy students were tested in 2012 and 85 students were tested in 2016. But, over that time period, there was also a change in the percentage of students in the state as a whole who scored at the proficient level – from 55 percent in 2012 to 60 percent in 2016. Was the change at Oak Elementary educationally significant when compared to the changes in the state as a whole? Could it have just appeared by chance?

To answer these questions, data could be entered into the EIC as shown in the top part of Table 13. Note that, as with examples 7 to 12, the EIC specifies the lines in which data for each year should be entered. The data for the more recent year are entered first, followed by the data for the later year. Also note that, to obtain tests of significance, the user only needs to know the number of cases at the organization of interest (in this case Oak Elementary), not the larger group.

The results are shown in the bottom part of Table 13. The effect size of .40 is beyond the level generally used to denote educational significance, the improvement index is large, and there is only a remote possibility the results could appear by chance. Thus, the Oak Elementary teachers could conclude that the improved achievement of their students was greater than that of students in the state as a whole.

Table Thirteen	
Comparing Changes in the Percentage of Students at Benchmark in One Group with Changes in a Larger Group	
Data Entered for Example Thirteen	
Enter the data for your group	
Percentage for the more recent year	65
Number of students tested in the more recent year (if available)	85
Percentage for the comparison year	40
Number of students tested in the comparison year (if available)	70
Now enter the data for the larger comparison group	
Percentage for the more recent year	60
Percentage for the comparison year	55
Results for Example Thirteen	
Effect Size	0.40
Improvement Index	15.7
Probability this effect would occur by chance	0.012

Example Fourteen: Comparing Changes in the Average Scores of Students in One Group with Changes in a Larger Group

Principal Greene, of East High school, was concerned that seniors in recent years seemed to have lower Scholastic Achievement Test scores than those in previous years. He wanted to compare the change from 2010 to 2015 for East seniors to those for the nation as a whole. In 2010 the average score of the 150 students tested was 685. In 2015 170 students were tested, and their average score was 650, reflecting a decline of 35 points in the average. One hundred fifty East students were tested in each year. For the nation, the average in 2015 was 585 and the average in 2010 was 590, a decline of 5 points. In both years the standard deviation for the national data was 150. Principal Greene could use the EIC to examine the extent to which the decline in East students' scores was greater than the decline among all students in the nation. Was this difference educationally significant?

The data entered are in the top part of Table 14. Note that, to obtain information regarding the probability of the results occurring by chance, Principal Greene did not need to know how many students were in the national testing. He only needed to know the number of students in his school in each year. Also note that it is only necessary to know the standard deviation for the larger group.

The results computed by the EIC are in the bottom part of Table 14. The effect size of -.20 is close to the .25 threshold of educational importance, and the associated probability level of .08 indicates that an effect of this size would occur less than eight times out of a hundred. From 2010 to 2015 decline in the SAT scores of the average East student, relative to the decline among students in the nation as a whole, was almost eight percentile ranks. Given this information it would seem logical for Principal Greene to be concerned about his students' performance on the SAT.

Table Fourteen	
Comparing Changes in the Average Scores of Students in One Group with Changes in a Larger Group	
Data Entered for Example Fourteen	
Enter the data for your group	
Mean for the more recent year cohort 1 - your group	650
Number of students tested in the more recent year (if available)	170
Mean for the comparison year - your group	685
Number of students in the comparison year (if available)	150
Now enter the data for the larger comparison group	
Mean for the more recent year	585
Standard deviation for the more recent year	150
Mean for the comparison year	590
Standard deviation for the comparison year	150
Results for Example Fourteen	
Effect Size	-0.20
Improvement Index	-7.9
Probability this effect would occur by chance	0.075

Example Fifteen: Comparing Changes in One Group with Changes in a Larger Group – Percentile Ranks of the Average Student

The teachers at Hillside school were proud of the change in their students' achievement over the past few years. For instance, in the most recent year the average student scored at the 60th percentile on the district's standard test of achievement. Three years earlier the average student was at the 40th percentile. (One hundred students were tested each year.) At the same time, scores in the district, which had always been higher than those at Hillside, were also rising. In the most recent year the average district student scored at the 78th percentile, while three years previous the average district student had scored at the 76th percentile. The Hillside

teachers wanted to know the extent to which the change in achievement among their sixth graders was greater than that of sixth graders in the district.

To answer this question they entered data in the EIC, as shown in the top panel of Table 15. The results are in the bottom panel. The effect size of .44 is well over the .25 threshold and the probability level of .002 indicates that the results would be very unlikely to have occurred by chance. The Improvement Index of 17 indicates that the improvement of the average student in Hillside was 17 percentile ranks greater than the improvement of the average student in the district. In short, the results of the EIC indicate that the Hillside teachers could be justifiably proud of their students' accomplishments.

Table 15

Table Fifteen	
Comparing Changes in the Percentile Rank of the Average Student in One Group to Changes in the Percentile Rank of the Average Student in a Larger Group	
Data Entered for Example Fifteen	
Enter the data for your group	
Percentile of the average student in the more recent year	60
Number of students in the more recent year	100
Percentile of the average student in the comparison year	40
Number of students in the comparison year	100
Enter the data for the comparison group	
Percentile of the average student in the more recent year	78
Percentile of the average student in the comparison year	76
Results for Example Fifteen	
Effect Size	0.44
Improvement Index	17
Probability this effect would occur by chance	0.002

Summary and Discussion

The No Child Left Behind Act requires schools to publish information on their students' performance on standardized assessments. State departments of education make this information publicly available on the web and often disseminate it through the media. These data allow educational consumers – parents, administrators, policy makers, and teachers – to compare scores of their students with those in their district or state or with students in other schools. They also allow consumers to look at changes over time. The EIC lets educational consumers use these data in ways typically employed by educational researchers and compute

effect sizes that can indicate the extent to which comparisons would be deemed educationally significant.

Educational research has, in recent decades, become its own form of big business, with consultants paid enormous sums of money to analyze data and various agencies devoting hundreds of hours and millions of dollars to the endeavors. Yet, much of this work is relatively inaccessible, and the analyses are often cloaked in a manner that implies that only certain people are capable of examining or understanding the information. In reality, as shown in the appendix to this report, the calculations needed to examine trends in educational achievement are relatively simple, well within the skill level of those who have completed an introductory statistics class. The EIC makes it even easier, by embedding the calculations in a spreadsheet available on the web.

Thus, the Educational Impact Calculator is designed to empower educational consumers and counter the impression that only "specialists" can understand achievement data. The hope is that it will give parents and policy makers the tools to independently assess the extent to which achievement in their schools differs from that in others and the extent to which changes in achievement in their schools over time would be deemed educationally important and statistically significant. The procedures described are simple and use publicly available data. Because they do not require high priced educational consultants, they are also inexpensive. Most important, they provide accurate and valid results to the questions that educational consumers typically ask and provide data for the most effective advocacy for their students.

While the tools described in this paper can be useful to educational consumers, users need to remember their limitations. Most important, they should realize that the findings that are produced are only as good as the data that are available. The assessment information that is typically released to the public is, at best, a snapshot in time. Thus, it provides only a partial view of the learning that may be occurring in a school. Wise users will want to consult as many sources of information and data as possible, looking at a variety of assessments, data from a range of time periods, and information for different grade levels. In addition, users need to remember all the many factors that can influence student and learning. Changes in administration, staffing levels, supports for teachers, curricular materials, time for instruction, and behavioral climate within a school can all be factors that influence learning. Wise consumers will want to ensure that they have considered as many relevant factors in their analyses as possible. In short, while the results certainly provide useful information, wise consumers continue to gather and assess data – knowing that more information is always better than less information.

Appendix

This appendix describes the methodological and statistical details that underlie the Educational Impact Calculator (EIC). It includes descriptions of the underlying research designs, statistical formulas, and the calculations embedded in the EIC. The underlying statistics are all relatively simple and should be familiar to those who have completed an elementary college level course in quantitative methods. The first section provides a brief overview of key elements of the EIC, including basic definitions of the input data needed and the output statistics. The next sections describe the designs and calculations used to answer each of the five queries outlined in the body of this report. The examples used in those sections are based on those detailed in the body of the paper.

Elements of the EIC

This section describes the input data used in the EIC and the three output statistics. Explanations of these data and relevant definitions and equations are given below.

Input Data

The EIC accepts three types of data regarding student achievement. All of these statistics are measures of central tendency and aggregate in nature: 1) the percentage of students who have reached a given benchmark or proficiency level, 2) average values on a scale or continuous measure of achievement, and 3) the percentile of the average student.

Percentages. When users input data as percentages, the EIC transforms these values to proportions by simply dividing the percentage by 100. The standard deviation of binominal distribution, the distribution used with proportions, is calculated as follows:

$$SD_p = \sqrt{pq}, \quad (1)$$

where p is the proportion and q is its complement (1-p). For instance, with a percentage of 60, p = .60, q = .40, and $SD_p \sqrt{.60 * .40} = \sqrt{.24}$**,** = .490.

Means and Standard Deviations. The second measure of central tendency used by the EIC is the arithmetic average or mean, the sum of all scores divided by the number of cases. When inputting average values the standard deviation (SD) is also needed. When comparing two schools or two cohorts with each other (queries 3 and 4), users need to know the SD in both groups. When comparing one group to a larger entity to which the group belongs (queries 2 and 5), the user only needs to know the standard deviation for the larger group.

Percentile Ranks. Percentiles require special consideration and treatment in statistical analyses and computations such as those involved in the EIC. Percentile ranks are transformations of the frequency distribution of scores on a given test, reporting the percentage of test takers that score at or below a given point. For instance, a school in which the average student scored at the 75th percentile indicates that this average student had scores that were equal to or higher than 75 percent of the other students taking the test. If graphed, a percentile distribution would be a simple rectangular, or even, distribution. For instance, ten percent of the test takers would be evenly distributed across each decile; 25 percent of test takers would be evenly distributed in each quartile. Yet, most tests actually have a normal, or bell-shaped, curve as the underlying distribution. Relatively more students have results bunched around the middle, or mean, of the distribution, and higher (and lower) scores are more spaced out. As a result, percentile scores misrepresent the extent to which students' scores are relatively similar or dissimilar; and computing averages with percentiles can be misleading.

To counteract this problem researchers typically transform percentiles to Normal Curve Equivalent (NCE) scores and use the transformed scores in the calculations. This transformation alters the distribution of scores to a normal, or bell-shaped, curve with a mean of 50 and a standard deviation of 21.06. A student with a percentile score of 50 would have an NCE score of 50, reflecting the symmetric nature of the normal curve where the mean and median (50th percentile) are equal. A percentile rank of 99 corresponds to an NCE score of 99, and a percentile of 1 corresponds to an NCE of 1. Differences occur in midranges. For instance a percentile of 40 corresponds to an NCE score of 45, and a percentile of 30 corresponds to an NCE score of 39, reflecting the way in which the normal curve has a larger percentage of area in the midpoint of the distribution. The EIC transforms percentile scores to NCE scores using the following EXCEL formula:

NCE = 21.06*NORMSINV(PR/100)+50, (2)
where PR is the percentile rank associated with the average score

Two additional points regarding the use of percentiles should be made. First, some have suggested that effect sizes calculated with NCE scores may provide conservative (i.e. smaller) estimates of effects (McLean, O'Neal, & Barnette, 2000). Second, as explained in more detail below, the EIC calculations with NCE scores use the population value for the SD (21.06), which may vary from the sample values. Thus, users are cautioned to use the other two available modes of input data (percent at benchmark or means and SDs) whenever possible.

Sample Size. If users know the number of students the EIC will give results of tests of significance. For comparisons of one school or cohort with another (queries 1, 3, and 4), the user needs to know the number of students in both groups. For comparisons with a larger entity (queries 2 and 5), the user only needs to know the number of cases in the smaller group, not within the larger population.

Output Statistics

The EIC produces three output statistics: effect size, improvement index, and probability level.

Effect Size. Effect sizes describe the magnitude of a difference or a statistical result, translating differences into a standard format. The effect size calculated by the EIC is Cohen's d, which uses the common standard deviation as the denominator.[3] Cohen's d is defined as the difference between two means divided by the common standard deviation.

$$ES = (M_1 - M_2)/SD_c, \quad (3)$$

where M_i is the mean, i denotes the two groups involved, and SD_c is the common standard deviation of the two groups.

An effect size of zero indicates no effect. Larger values indicate more of an effect, either positive or negative. An effect size of 1.0 indicates that the scores differ by one standard deviation, an effect size of 0.5 indicates that they differ by one-half of a SD, etc. Somewhat arbitrarily, Cohen deemed effect sizes of .8 of an SD and greater as large, those of .5 as medium, and those of .2 to .3 as small (Cohen, 1988). Within the field of education an effect size of .25 has generally been considered educationally significant (Tallmadge, 1977).[4]

Improvement Index. The improvement index translates the effect size into percentile form and is designed to help users interpret results. It equals the difference between the percentile rank of the average student in the group of interest and the percentile rank of the average student in the comparison group. The following set of formulas is used to calculate the improvement index in Excel:

Improvement in NCE Scores (INCE) = (21.06*d) + 50, where d is the effect size; (4)

[3] The other most commonly used effect size is Hedge's g. It differs from Cohen's d in the denominator by using degrees of freedom, rather than sample size. Not surprisingly the two values are almost perfectly correlated and differ only very slightly in magnitude.

[4] Recently, Lipsey, Puio, Yun, et al (2012) examined variation in effect sizes across 181 studies in education published after 1995 and using a randomized control group design. They reported a mean value of .28 and a median of .18. As noted in the body of this report, the effect sizes associated with efficacy studies of the Direct Instruction curriculum are, on average, at least twice the typical benchmark of .25.

Improvement in %ile scores (I%ILE) = 100*NORMSDIST((INCE-50)/21.06); and (5)

Improvement Index (II) = I%ILE – 50. (6)

Equation 4 translates the effect size, d, into NCE units. Equation 5 then transforms the NCE score produced in equation 4 to a percentile rank. And equation 6 compares this percentile to the percentile of the average student, which is, by definition, 50. For instance, an effect size of .50 would, using equation 4, result in INCE = (21.06*.50) + 50 = 10.53 + 50 = an NCE score of 60.53. Using equation 5, one would find that this NCE score corresponds to a percentile rank of 55. Then one would use equation 6 to find that this corresponds to an improvement index of 5 percentile ranks (=55-50).

Careful readers will note that the improvement index is sometimes not equivalent to the value one would receive through simple subtraction of percentiles. That occurs because of the differences between NCE scores and percentiles described above. One could argue that the Improvement Index calculated with NCE scores provides a more accurate estimate of the effects than simple comparisons of the percentile ranks.

Tests of Significance. Researchers use tests of significance to decide if differences are "random noise," that is, whether they could simply have occurred by chance. The EIC computes t-ratios and z-scores as the basis for tests of significance. In the EIC the 2-tail probability level is calculated with the following excel formula:

p = 2*((1-NORMSDIST(z))) (7)

where z is the absolute value of the test statistic (z or t).[5] The EIC reports only two-tail (non-directional) probabilities.

Although again admittedly arbitrary, researchers have, for many years, used a probability figure of .05 as indicating statistical significance. When comparing two groups, a probability of .05 or less would indicate that such differences would, by chance, only occur five times out of 100. Yet, tests of significance are highly influenced by the number of students in a comparison. It is easier to find significant results when many students are examined and harder to do so when few students are included.

[5] The program used to post the spreadsheet underlying the EIC to the web does not support the use of a t-distribution. Fortunately, the t distribution differs from the z (normal) distribution only when samples are relatively small and, even with such small samples, the differences in probabilities are minimal.

Effect sizes remain the same no matter how many students are included and are easier to compare from one situation to another. Thus, they are often more useful for educational consumers. The inverse association of effect size and level of significance, as well as the ways in which levels of significance are influenced by sample size, often make interpretation more difficult for those with less experience working with statistical results. Thus the body of this paper recommends that only those who are comfortable with the concept of probability employ results of tests of significance in their analyses.

Query One: Comparing Scores in One Group with Scores in Another Group

Education consumers often want to compare achievement results of one group with those of another group. For instance, they might want to compare scores in one classroom with those in another classroom, scores in one school with those in another school, or scores in one district with those in another district. Using the language of experimental design developed by Campbell, Stanley, Shadish and Cook (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002), these comparisons involve a post-test only control group design. This design is illustrated in Figure A-1, using the conventions of the Campbell et al tradition. The first line in the figure represents the user's group and the second represents the comparison group. The X in the first line is commonly used to indicate an experimental treatment or the way in which the two groups differ. The Oi values at the end of each line indicate that each group had an assessment (observation) that is compared. Note that an important element of this design is the independence of the two groups. That is, the students in one group are not in the other.[6]

Figure A-1
Posttest Only Control Group Design

Group One (e.g. user's school)	X	Assessment (O_1)
Group Two (e.g. a comparison school)		Assessment (O_2)

Effect Size and Improvement Index

Comparing scores of the two groups in this design is a classic use of effect size methodology, using formula (3) defined above. The difference of two central tendencies is divided by the

[6] A dependent design would be a "repeated measures," pretest-posttest, or panel design in which the same students would be tested at each time point. At present, the EIC does not handle that design.

common standard deviation. For calculations involving proportions (example one in the body of this paper), the appropriate formula is

$d = (p_1 – p_2)/s_{pc}$, (8)

where p_1 is the proportion in group 1 (the target group), p_2 is the proportion in group 2 (the comparison group), and s_{pc} is the common standard deviation.

For calculations involving means (examples two and three), the appropriate formula is

$d = (M_1 – M_2)/s_c$, (9)

where M_1 and M_2 are the average values in the 2 groups and s_c is the common standard deviation.

The common standard deviation can be computed as the weighted average of the standard deviations of the two groups. For proportions,

$s_{pc} = (n_1sp_1 + n_2sp_2)/(n_1+n_2)$, (10)

where s_{pi} is defined as in equation 1 and n_i refers to the size of each group i (i = 1 and 2).

For continuous data and means and standard deviations,

$s_c = ((n_1*s_1) + (n_2*s_2))/(n_1+n_2)$, (11)

where n_i refers to the size of each group and s_i refers to standard deviation of each group.

If the number of cases is not available the EIC assumes the groups are of equal size and computes the simple average.

The improvement index is calculated using equations 4, 5, and 6.

Test of Significance

The appropriate test of significance for Query One is a simple t-test for independent groups. For proportions, the null hypothesis is

Ho: $P_1 = P_2$, or alternatively $P_1 – P_2 = 0$,

For means and standard deviations, the null hypothesis is

Ho: $\mu_1 = \mu_2$, or alternatively $\mu_1 – \mu_2 = 0$,[7]

[7] Technically, μ_1 and μ_2 refer to the means of the hypothetical populations that M_1 and M_2 represent; and P_1 and P_2 refer to the proportion in the hypothetical populations.

The formula for the t-statistic differs from the effect size in only the denominator. For effect sizes the denominator is the common standard deviation, but for the t-ratio and tests of significance the denominator is the standard error. The standard error is directly related to the standard deviation, getting larger as the standard deviation increases. However, it is inversely related to sample size. With larger samples the standard error is smaller, but with smaller samples it is larger – even if the standard deviation is the same. In other words, as noted above, effect sizes are not influenced by the number of students included in the assessments, but tests of significance are affected. It is easier to have significant results when samples are larger.

The formula for the standard error of the difference of two measures of central tendency can be calculated from the standard deviations. The formula for the standard error for the difference of proportions is

$$б_{s.e.p1-p2} = [(s_{p1})^2/n_1] + [(s_{p2})^2/n_2], \quad (12)$$

where s_{pi} is the standard deviation calculated as in equation one and n_i refers to the sample size.

The formula for the standard error for the difference of means is

$$б_{s.e.M1-M2} = √[(б^2_1/n_1) + (б^2_2/n_2)], \quad (13)$$

where $б^2_i$ is the variance, or the square of the standard deviation.

The t-ratio is defined as the difference between the measures of central tendency divided by the standard error of the difference between the means. For the difference of proportions the t-ratio is calculated as

$$t = (p_1 – p_2)/б_{s.e.p1-p2} . \quad (14)$$

where p_1 is the proportion in group 1, p_2 is the proportion in group 2 and $б_{s.e.p1-p2}$ is the standard error for the difference of proportions defined in equation 12.

For the difference of means the t-ratio is calculated as

$$t = M_1 – M_2)/б_{s.e.M1-M2,} \quad (15)$$

where M_1 is the mean of group 1, M_2 is the mean of group 2 and $б_{s.e.M1-M2}$ is the standard error of the difference of the means, defined in equation 13.

Examples One, Two, and Three

Table A-1 gives the results of the calculations for examples 1, 2, and 3 in the body of this paper. All of the results were computed in Excel using the equations given above. Example 1 involved data in the form of percentages, example 2 involved means and standard deviations, and example 3 involved percentile ranks associated with the average score. The first two lines of

data give the central tendencies for each example and the third gives the difference of these values. The data for example 1 were entered as percentages, but all calculations were done after these percentages were transformed to proportions. The data for example 3 were entered as percentile ranks, but the calculations involve NCE scores computed using equation 2 above. Thus, the differences given in the third line of data represent, for example 1, the difference in proportions, and, for example 3 the difference in the NCE scores that correspond to the percentiles (58.7 – 66.3 = -7.6). Readers can use these data to reproduce the results using the equations given above.

Table A-1

Statistics Calculated for Examples One to Three - Comparing One Group With Another

	Example 1	Example 2	Example 3
Central Tendency User's Group	50%	110	66
Central Tendency Other Group	65%	107	78
Difference of Central Tendencies	-0.15	3	-7.6
Common SD	0.49	14.5	21.06
Standard Error of Difference	0.07	1.68	5.96
effect size	-0.31	0.21	-0.36
Improvement Index	-12.1	8.2	-14
t-ratio	-2.26	1.79	-1.27
prob.	0.02	0.07	0.20

Note: Example 1 used percentages as input, example 2 used means and standard deviations, and example 3 used percentile ranks corresponding to scores of the average student.

Query Two: Comparing Performance of One Group with a Larger Entity

The second question that can be answered by the EIC involves the extent to which average scores of a group differ from those of a larger group to which it belongs. As shown in Examples 4, 5, and 6, this can involve comparisons of a classroom, school or district with the school, district, state, or nation in which it is embedded.

The logic involved in these comparisons is that of a norm comparison design, with the larger entity analogous to a norm group. This design is illustrated in Figure A-2. As in Figure A-1 the top line refers to the user's group, the second line refers to the comparison group, X refers to the way in which the user's group differs from the larger group, and O*i* refers to the assessment. In contrast to the design used with query one, the comparison group in this design is a larger unit to which the user's group belongs.

Figure A-2
Norm Comparison Design

Group One (user's group)	X	Assessment (O_1)
Group Two (the larger group)		Assessment (O_2)

Effect Size and Improvement Index

The logic involved in statistical analyses associated with this question involves comparing results with a sample (e.g., the user's school or district) to those of a larger population or universe (e.g., the district or the state). This is one of the first topics studied in elementary statistics classes. The effect size is simply the standard score, or z-score.

In statistical jargon, the values for the schools are the sample statistics, and those for the district or state are the parameters of the population or universe. For proportions values associated with a sample are typically denoted by the subscript "s," while those for the population or universe are denoted by the subscript "u" for "universe." For instance, the proportion associated with the user's group (the sample school or district) is termed p_s, and its complement as q_s $(1\text{-}p_s)$. The proportions associated with the larger group are the population parameters, and often denoted P_u and Q_u, for the proportion and its complement. For means and standard deviations, sample values are denoted by the Latin alphabet and population values by Greek letters. For instance, the means and standard deviations for the sample are denoted by M and s, while the values for the population parameters are denoted by μ (the population mean) and б (the population standard deviation).

The effect size is the difference between the sample statistic and the population parameter divided by the standard deviation of the population. For proportions the formula is

$$d = z = (P_s - P_u)/ б_p, \quad (16)$$

where P_s is the proportion in the sample, P_u is the proportion in the population and $б_p$ is the standard deviation of the population calculated using formula (1).

For continuous data (means and standard deviations), the formula is

$$d = z = (M - \mu) / б, \quad (17)$$

where M is the sample mean, μ is the population mean, and б is the population standard deviation. Note that the standard deviation is that of the population, not the sample.

The improvement index is calculated from the effect size using equations 4 to 6.

Test of Significance

The test of statistical significance is based on the null hypothesis that the sample statistic equals the population parameter:
Ho: $p_s = P_u$, or, alternatively, Ho: $p_s - P_u = 0$ for proportions; and
Ho: $M = \mu$, or, alternatively, Ho: $M - \mu = 0$ when means and standard deviations are given.

Again, the difference in calculating the effect size and the test of significance lies in the denominator. The test statistic, unfortunately also typically called z, is calculated by simply dividing the difference by the standard error. (To help minimize confusion this is called the z-ratio in this paper.).

For tests with proportions, the standard error is defined as

$$б_{s.e.p.} = б_p/\sqrt{(1/n)}, \quad (18)$$

where $б_p$ is the standard deviation for the population and n is the number of students in the user's group.

$$\text{The z-ratio} = (P_s - P_u)/б_{s.e.p}, \quad (19)$$

where P_s is the proportion in the user's group (the sample), P_u is the proportion in the larger group, and $б_{s.e.p}$ is the standard error for proportions as defined in equation 18.

For tests with continuous data (means and standard deviations), the standard error is defined as

$$б_{s.e.} = б/\sqrt{(n)}, \quad (20)$$

where б is the population standard deviation and n is the number of students in the user's group.

$$\text{The z-ratio} = (M - \mu)/б_{s.e.} \quad (21)$$

where M is the sample mean, μ is the population mean, and $б_{s.e}$ is the standard error of the mean defined in equation 20.

Examples 4, 5, and 6

Table A-2 gives results of the calculations for examples 4, 5, and 6, all based on the equations given above. Example 4 used data in the form of percentages, example 5 involved means and standard deviations, and example 6 involved percentiles associated with the average score. As in the calculations for examples 1, 2, and 3 the percentages were transformed to proportions and the percentiles to NCE scores. Again, these data can be used to calculate the results.

Table A-2

Statistics Calculated for Examples Four to Six - Comparing One Group with a Larger Group to Which it Belongs

	Example 4	Example 5	Example 6
Central Tendency User's Group	55%	105	35
Central Tendency Other Group	60%	100	50
Difference of Central Tendencies	-0.05	5	-8.1
Common SD	0.49	15	21.06
Standard Error of Difference	0.07	2.12	0.29
Effect size	-0.10	0.33	-0.39
Improvement Index	-4.1	13.1	-15.0
z-ratio	-0.72	2.36	-8.4
prob.	0.47	0.02	<.001

Note: Example 4 used percentages as input data, example 5 used means and standard deviations, and example 6 used percentiles of the average student's score.

Query Three: Changes in a Group Over Time

Educational consumers are often interested in changes that occur in schools, classrooms, or districts over time, especially as new programs or procedures are implemented. Examples 7 to 9 showed how the EIC can address this question. The techniques used to answer these questions are an example of the cohort control group or recurrent institutional cycle design in which the achievement of one cohort is compared to that of another. This design is described in the classic experimental design literature as a useful alternative to randomized control trials in organizational settings. As Shadish, Cook, and Campbell (2002), put it,

> Many institutions experience regular turnover as one group "graduates" to another level and their place is taken by another group. Schools are an obvious example of this, as most children are promoted from one grade to the next each year….The term cohort designates the successive groups that go through processes such as these. Cohorts are particularly useful as control groups *if* (1) one cohort experiences a given treatment and earlier or later cohorts do not; (2) cohorts differ in only minor ways from their contiguous cohorts; (3) organizations insist that a treatment be given to everybody, thus precluding simultaneous controls and making possible only historical controls; and (4) an organization's archival records can be used for constructing and then comparing cohorts. (Shadish, et al., 2002, pp. 148-149, emphasis in original,; see also Cook & Campbell, 1979, pp. 126-127 and Campbell & Stanley, 1963, pp. 56-61; Stockard, 2013).

Figure A-3 illustrates the logic of this design. As with the designs discussed above comparisons are made with two groups. But, in contrast to the previously discussed designs, the comparisons are between two cohorts or groups of students within an organization, such as third graders in the current year and third graders in a previous year. T1 refers to time one, or the period in which the earlier cohort was assessed; T2 refers to time two, or the period in which the more recent cohort was assessed; and X refers to the way in which the more recent cohort differs from the earlier one. Just as in the analyses associated with query one (and as described in the quote from Shadish, et al., above) it is assumed that the two cohorts are independent of each other. That is, the students in the more recent cohort were not in the earlier cohort.

Figure A-3
Cohort Control Group Design

	T1		T2
Group One – Earlier Cohort	O_1		
Group One – Recent Cohort		X	O_2

Effect Size and Improvement Index

Comparing scores of two cohorts is based on the techniques and equations used to compare scores from two similar groups described in the discussion of query one above. Because the two cohorts are independent of each other (e.g. second graders in one year are a different group than second graders in the next year), the classic effect size methodology and a simple independent t-test can be used. Thus, as described in the previous section, the effect size is simply the difference between the measures of central tendency divided by the common standard deviation (equations 8 to 11) and the test of statistical significance is the classic t-test of the difference between the means (equations 12 to 15). The only difference from the discussion in the previous section is that the groups involved in the comparisons are cohorts, rather than organizations.

Examples 7, 8 and 9

Table A-3 gives results of the calculations for examples 7, 8 and 9 in the body of this paper. All of the results are based on the equations given above. Example 7 used data in the form of percentages, example 8 involved means and standard deviations, and example 9 involved percentiles associated with the average score. As in previous examples the percentages were transformed to proportions and the percentiles to NCE scores.

Table A-3

Statistics Calculated for Examples Seven to Nine - Comparing One Cohort with Another Cohort

	Example 7	Example 8	Example 9
Central Tendency User's Group	65%	58	58
Central Tendency Other Group	40%	65	54
Difference of Central Tendencies	0.25	-17	2.1
Common SD	0.48	14	21.06
Standard Error of Difference	0.08	1.4	3.23
Effect size	0.52	-0.50	0.10
Improvement Index	19.8	-19.1	4.0
t-ratio	-3.20	-4.99	0.66
prob.	0.001	<.001	0.51

Note: Example 7 used percentages as input data, example 8 used means and standard deviations, and example 9 used percentiles of the average student's score.

Query Four: Changes over Time in One Group Compared to Changes in Another Group

The fourth query addressed by the EIC involves comparing changes over time in one group to changes in a similar group. For instance, users might compare changes in their school to those in another school or changes in their district to those in another district. This type of analysis involves a pretest-posttest cohort control group design. It combines the post-test only control group design used in Query One and the cohort control group design used in Query Three. When a user wants to examine the possible impact of an intervention, such as a new curriculum, the pretest-posttest cohort control group design is arguably more powerful than either of the other designs. This is because the design controls to at least some extent for factors, such as historical events or current events that might have affected change in both groups.

The logic of the pretest-posttest cohort control group design is illustrated in Figure A-4. It involves four cohorts – two in the user's group and two in the comparison group. Within each group data are obtained for a cohort at an earlier time point (T1) and for a cohort at a more recent time point (T2). Note that the differences between the two groups (using the logic of query one) are obtained by comparing assessments at T1 (O_1 and O_3) and at T2 (O_2 and O_4). If

the intervention in the user's group has made a difference, it would be expected that the difference at T2 would vary from that at T1. Similarly, one can compare the achievement of the two cohorts within each group (comparing O_1 and O_2 for group one and comparing O_3 and O_4 for group 2). If the intervention has made a difference, the effect size for the user's group would be larger than the effect size for the comparison group.

Figure A-4
Pretest-Posttest Cohort Control Group Design

	T1		T2
Group One – Earlier Cohort	O_1		
Group One – Recent Cohort		X	O_2
Group Two – Earlier Cohort	O_3		
Group Two – Recent Cohort			O_4

Effect Size and Improvement Index

The technique involved in calculating the effect size for the pretest-posttest cohort control group design is a simple extension of the development related to query one regarding the comparison of one group to another similar group and builds on the fact that the effect size d is in standard deviation units. Thus, the effect size for the change in one group relative to the change in another entity is simply equal to the difference of the effect size for the two years in the comparison. That is,

$$d_{2\text{-}1|2\text{-}1} = d_2 - d_1, \quad (22)$$

where d_2 is the effect size from the second year in the comparison, d_1 is the effect size in the earlier year in the comparison, and $d_{2\text{-}1|2\text{-}1}$ refers to the effect size related to the change in the user's group over time relative to the change in the other group. In other words, it tells, in standard deviation units, the extent to which change in the user's group differs from that in the other group. Equivalent results are obtained by comparing the effect sizes associated with the cohort comparisons in each group, using the logic embedded in query one. Again, the improvement index is calculated with equations 4 to 6.

Test of Significance

The null hypothesis examined for comparisons in the pretest-posttest cohort control group design is simply H_o: $d_{2\text{-}1|2\text{-}1} = 0$, that the effect size equals zero. Recalling that the standard deviation for effect sizes is, by definition, 1.0, the formula for the standard error is

$\sigma_{s.e.d2\text{-}1|2\text{-}1} = \sqrt{[1/(n_{11}+n_{12}-2)] + [1/(n_{21}+n_{22}-2)]}$, (23)

where n_{ij} is the sample size for school j in year i.[8]

The associated t-ratio is then

$t\text{-ratio} = d_{2\text{-}1|2\text{-}1} / \sigma_{s.e.d2\text{-}1|2\text{-}1}$ (24)

where $d_{2\text{-}1|2\text{-}1}$ is defined as in equation 22 and $\sigma_{s.e.d2\text{-}1|2\text{-}1}$ is defined as in equation 23.

Examples

Table A-4 gives results of the calculations for examples 10, 11, and 12 in the body of this paper. All of the results are based on the equations given above. Example 10 used data in the form of percentages, example 11 involved means and standard deviations, and example 12 involved percentiles associated with the average score. As in previous examples the percentages were transformed to proportions and the percentiles to NCE scores. Again the table includes sufficient details to allow readers who are interested in the intricacies of the calculations to verify the results obtained by substituting values into equations 24 to 26 above.

Table A-4

Statistics Calculated for Examples Ten to Twelve - Comparing Changes from One Cohort to Another in one Group to Changes in Another Group of a Similar Nature

	Example 10	Example 11	Example 12
Central Tendency User's Group, T2	70%	105	75
Central Tendency User's Group, T1	50%	110	66
Central Tendency, Other Group T2	75%	106	79
Central Tendency, Other Group T1	65%	107	78
Difference of Central Tendencies, T2	-0.05	-1.0	-2.78
Difference of Central Tendencies, T1	-0.15	3.0	-7.58
Effect Size of Difference T2	-0.11	-0.069	-0.13
Effect Size of Difference T1	-0.31	0.207	-0.04
Effect Size of Difference of Differences	0.2	-0.28	0.23
Standard Error of Difference of Differences	0.10	0.08	0.20
Improvement Index	7.7	-10.9	9.00
t-ratio	206.00	-3.37	1.12
prob.	0.04	0.001	0.26

Note: Example 10 used percentages as input data, example 11 used means and standard deviations, and example 12 used percentiles of the average student's score. T2 refers to the more recent time period as entered into the EIC.

[8] This formula for the standard error is easily derived from the formula s.e. = $\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$, by recalling that the standard deviation of z scores is, by definition, equal to one.

Query Five: Changes in One Group over Time Compared to Changes in a Larger Group

The fifth query addressed by the EIC involves comparing changes over time in one group to changes in a larger group to which it belongs. For instance, users might compare changes in their school to those in the district as a whole or changes in their district to those in the state. This type of analysis combines the norm comparison design used in Query Two and the cohort control group design used in Query Three. It is called a cohort control group historical comparison design and is arguably more powerful than either of the other designs.

The logic of the cohort control group historical comparison design is illustrated in Figure A-5. It involves four cohorts – two in the user's group and two in the larger comparison group. Within each group data are obtained for a cohort at an earlier time point (T1) and for a cohort at a more recent time point (T2). Note that the differences between the two groups (using the logic of query two) are obtained by comparing assessments at T1 (O_1 and O_3) and at T2 (O_2 and O_4). If the intervention in the user's group has made a difference, it would be expected that the difference at T2 would vary from that at T1. Similarly, one can compare the achievement of the two cohorts within each group (comparing O_1 and O_2 for group one and comparing O_3 and O_4 for the larger group). If the intervention has made a difference, the effect size for the user's group would be larger than the effect size for the larger comparison group.

Figure A-5
Cohort Control Group Historical Comparison Design

	T1		T2
User's Group – Earlier Cohort	O_1		
User's Group – Recent Cohort		X	O_2
Larger Comparison Group – Earlier Cohort	O_3		
Larger Comparison Group – Recent Cohort			O_4

Effect Size and Improvement Index

The effect size for the change in a school or district relative to the change in a larger entity is calculated in the manner described above for query four regarding comparisons to changes in a similar type of organization. The effect size is simply equal to the difference of the effect size for the two years in the comparison. That is,

$$d_{2\text{-}1|2\text{-}1} = d_2 - d_1, \quad (25)$$

where d_2 is the effect size from the more recent year in the comparison, d_1 is the effect size in the earlier year in the comparison, and $d_{2\text{-}1|2\text{-}1}$ refers to the effect size related to the change over time relative to the change in the other group. Equivalent results would be obtained by comparing the effect sizes associated with the cohort comparisons in each group. Again, the improvement index is calculated with equations 4 to 6.

Test of Significance

The null hypothesis examined for comparisons in the pretest-posttest cohort control group design is simply H_o: $d_{2\text{-}1|2\text{-}1} = 0$, that the effect size equals zero. Again, the tests of significance build upon the effect size calculations, but use the standard error, rather than the standard deviation, in the denominator. The standard error is a function of sample size in the user's group in the comparison years. (As with query four the logic derives from the fact that the d values are standardized scores where, by definition, s.d. = 1.00.) Thus, the standard error,

$$\text{s.e.}_{d2\text{-}1|2\text{-}1} = \sqrt{[(1/n_1) + (1/n_2)]}, \quad (26)$$

where n_1 = the sample size in year 1 and n_2 = the sample size in year 2.

The t-ratio to test the null hypothesis that the effect size equals zero is a simple function of the effect size and the standard error.

$$\text{t-ratio} = d_{2\text{-}1|2\text{-}1} / \text{s.e.}_{d2\text{-}1|2\text{-}1}, \quad (27)$$

where $d_{2\text{-}1|2\text{-}1}$ is defined as in equation 25 and $\text{s.e.}_{d2\text{-}1|2\text{-}1}$ is defined as in equation 26.

Examples 13, 14, and 15

Table A-5 gives results of the calculations for examples 13, 14, and 15 in the body of this paper. All of the results are based on the equations given above and involve data in the form of percentages, means and standard deviations, and the percentile rank associated with the average score. As in previous examples the percentages were transformed to proportions and the percentile ranks to NCE scores before calculations were completed. Again the table includes sufficient details to allow readers who are interested in the intricacies of the calculations to verify the results obtained by substituting values into equations 27 to 29 above.

Table A-5

Statistics Calculated for Examples Thirteen to Fifteen - Comparing Changes from One Cohort to Another in One Group to Those in a Larger Group to Which it Belongs

	Example 13	Example 14	Example 15
Central Tendency User's Group, T2	65%	650	60
Central Tendency User's Group, T1	40%	685	40
Central Tendency, Larger Group T2	60%	585	78
Central Tendency, Larger Group T1	55%	590	76
Difference of Central Tendencies, T2	0.05	65.0	-10.9
Difference of Central Tendencies, T1	-0.15	95.0	-20.2
Effect Size of Difference T2	0.10	0.43	-0.52
Effect Size of Difference T1	-0.30	0.63	-0.96
Effect Size of Difference of Differences	0.4	-0.20	0.44
Standard Error of Difference of Differences	0.16	0.11	0.14
Improvement Index	15.7	-7.9	17.00
t-ratio	2.50	-1.78	3.12
prob.	0.012	0.075	0.002

Note: Example 13 used percentages as input data, example 14 used means and standard deviations, and example 15 used percentiles of the average student's score. T2 refers to the more recent time period as entered into the EIC.

References

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Available at http://ies.ed.gov/ncser/.

McLean, J. E., O'Neal, M. R., & Barnette, J. J. (2000, November). *Are all effect sizes created equal?* Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Bowling Green, KY. (ERIC No. ED448188)

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Stockard, J. (2013). "Merging the Accountability and Scientific Research Requirements of the No Child Left Behind Act: Using Cohort Control Groups" *Quality and Quantity: International Journal of Methodology*, 47, 2225-2257.

Tallmadge, G. (1977). *The Joint Dissemination Review Panel Idea Book*. Washington, D.C.: NIE, U.S. Government Printing Office.