ACT Research Report Series 98-1

Interpreting Differences Between Mean ACT Assessment Scores

Jeff Schiel

ACT

January 1998

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

Interpreting Differences between Mean ACT Assessment Scores

Jeff Schiel

Table of Contents

Abstract

This report examines both the substantiveness and statistical significance of differences between mean ACT Assessment scale scores. The first part of the report describes the development and results of a method for interpreting the substantiveness of mean differences in terms of test items correctly answered. For example, if Group A has a mean ACT Composite scale score of 20.9 and Group B has a mean Composite scale score of 21.0, then each student in Group B has correctly answered either one more item on the English test or one more item on the Mathematics test, relative to students in Group A. In the second part of the report, several methods for interpreting ACT Assessment mean differences, including the above method, are applied to sample data.

Interpreting Differences between Mean ACT Assessment Scores

This report examines both the statistical significance and substantiveness (i.e., significance in a practical sense) of differences between mean ACT Assessment scale scores. The first part of the report describes the development and results of a method for interpreting differences between mean ACT Assessment scores in terms of test items correctly answered. This method is intended to supplement traditional indicators of substantiveness, such as effect size. In the second part of the report, examples of methods for interpreting mean differences, including the number of items correct method, are applied to sample data.

Interpreting Differences between Mean ACT Assessment Scores in Terms of Number of Correctly Answered Items

ACT staff routinely receive inquiries from high school administrators concerning the "significance" of changes over time in mean ACT Assessment scores. Many of these inquiries pertain to the "statistical significance" of the differences between mean scores. Statistical significance refers to the probability of an observed result being reasonably thought of as due to chance. Statistical significance testing has been criticized (e.g., Carver, 1993; Thompson, 1996). One criticism concerns the number of negligibly small research results that are labeled significant simply because they are statistically significant (Carver, 1993). Moreover, it is well known that statistical significance is related to sample size; large samples are more likely to yield statistically significant results, even though the results may not be very substantive. An example of this phenomenon is evident when comparing the mean ACT Composite for the 1997 national high school graduating class (21.0; n = 959,301) to the mean for the 1996 national graduating class (20.9; n = 924,663). The difference between these two means (0.1 scale score unit) seems fairly small; however, it is highly statistically significant because the sample sizes are very large.

What about substantiveness? In the above example, the 0.1 scale score unit difference between mean ACT Composite scores may seem, to some individuals, to not have much practical meaning. After all, one might reason, a 0.1 score unit difference on a scale that ranges from 1 to 36 does not seem especially large. Other individuals may, however, have a different opinion.

One way to evaluate the substantiveness of differences between mean ACT Assessment scores is with effect sizes. Effect sizes express the difference between two means in terms of standard deviation units. For example, the standard deviation of the ACT Composite score is typically about five scale score units. An effect size of 0.5 would therefore indicate that the mean of one group is about 2.5 scale score units (½ of a standard deviation) higher than that of the other group. Cohen (1988) proposed that effect sizes of .2, .5, and .8 be considered small, medium, and large, respectively. These are just guidelines, however, and may not be appropriate for all situations. In addition, interpreting effect sizes is a somewhat subjective process. One researcher, for example, may consider an effect size of 0.5 for the ACT Assessment to be noteworthy, whereas another researcher may not.

Another, relatively simple indicator of substantiveness is based on interpreting mean scale score differences in terms of the number of test items correctly answered (raw score). ACT Assessment scale scores, which ensure comparability of scores over different forms of the test, are based on a transformation of raw scores. Scale scores range from 1 to 36 for each of the four subject-area tests (English, Mathematics, Reading, and Science Reasoning) and for the Composite score, which is calculated by averaging the subject-area test scale scores. The English test has 75 items, the Mathematics test has 60 items, and the Reading and Science Reasoning tests each have 40 items.

For many ACT Assessment users, the relationships between the familiar scale scores provided on student and institutional reports and the number of items correctly answered are unclear. It is not well known, for example, that a student who correctly answers two additional items during a second administration of the ACT English test will often earn an English scale score that is about one unit higher than his or her previous scale score. The 75 English test items, when scored as correct or incorrect, yield raw scores ranging from 1 to 75. These 75 raw score units are then mapped to 36 scale score units. The ratio of English raw scores to scale scores is therefore about two to one. This means that for approximately every two English items answered correctly, one scale score unit will be awarded. This relationship is not constant throughout the range of English raw scores, however, and may differ from form to form. At certain raw scores, for example, a student could correctly answer three additional English items and still earn the same English scale score.

Raw score and scale score relationships become more difficult for ACT Assessment users to decipher when the scores for groups of students are averaged. For example, would a one scale score unit difference between two mean ACT English scores similarly translate into two additional items correctly answered by each student in the group with the higher of the two mean scores? Even more complex interpretational difficulties arise when examining mean ACT Composite scores. How, for example, would one interpret a one-unit increase in mean ACT Composite score? Would this also represent two additional correct items and, if so, on which ACT subject-area test might they have occurred?

This study was designed to answer these kinds of questions. For example, the results presented below suggest that a 0.1 scale score unit difference between two ACT Composite means could correspond to one additional correct item on either the ACT English or Mathematics test. For example, suppose that Group A has a mean ACT Composite score of 20.9 and Group B has a mean Composite score of 21.0. This result would occur if each student in Group B correctly answered either one more item on the English test or one more item on the Mathematics test, relative to students in Group A. This type of information provides ACT Assessment users another way to determine whether differences between ACT means are substantive.

Data

The data used in this study consisted of 10% systematic random samples of ACT-tested students from the high school graduating classes of 1994, 1995, and 1996. Table 1 shows the sample sizes for each combination of testing year and ACT Assessment test form. The letters A through G designate the actual form codes. Data from 21,005 students in the 1994 10% sample file who took one particular form of the test during 1993-94 were analyzed. For the 1994-95 testing year, data from 22,560 students who had taken a different form were analyzed. These two forms were, for 1993-94 and 1994-95, the most frequently administered forms. Data from six different ACT Assessment test forms were analyzed for 1995-96; sample sizes ranged from 276 to 21,883 over forms. Note that Form F was administered in both 1994-95 and 1995-96. These data allowed a comparison of different samples within the same form.

TABLE 1

Sample Sizes, by Form and Testing Year

	Testing year		
Form	1993-94	1994-95	1995-96
A	--	--	13,355
B	--	--	13,851
C	--	--	21,883
D	--	--	16,343
E	--	--	276
F	--	22,560	551
G	21,005	--	--

Method

One way to investigate the relationships between ACT Assessment mean differences and number of items correct is with an analytic method. Perhaps an equation could be algebraically developed, for example, that would relate mean differences to number of items correct. Such a strategy is problematic, however, because the mathematical relationship that exists between ACT Assessment raw scores and scale scores is complex. For example, during the scaling of the ACT Assessment, the raw score distribution was smoothed by fitting to it a four-parameter beta compound binomial model. In the next step, the smoothed raw scores were nonlinearly transformed so that error variance was stabilized. A linear transformation was then applied to the transformed scores, yielding a specified mean and standard error of measurement. Finally, these scores were rounded and, in some cases, adjusted using judgmental procedures. Such adjustments, for example, prevent too many raw scores from converting to the same scale score. A thorough description of the scaling of the ACT Assessment is given in Kolen and Hanson (1989).

The complexity of the relationship between ACT Assessment raw scores and scale scores extends beyond the scaling process. Each form of the ACT Assessment is equated to an anchor form to ensure the comparability of scale scores over forms. The equating function used in this process differs from form to form.

An alternative to an analytic method would be to simulate the effect on mean scale scores of different numbers of items being correctly answered. For example, the mean ACT scale score for a group of students could be compared to an adjusted mean ACT scale score, where the adjusted mean represents the simulated effect of each student correctly answering a certain number of additional items. This was the approach used in this study.

Composite score analyses. Separate analyses were performed to examine ACT Composite score mean differences and subject-area score mean differences. For the Composite score analyses, several steps were involved in computing the effect of raw score unit increases on scale score mean differences:

1. For each student record, add one raw score unit to one or more subject-area raw scores. This yields, for each student, an adjusted raw score (or scores).

2. Convert adjusted raw scores to adjusted scale scores, using the standard raw score to scale score conversions for each form of the ACT Assessment. (Note that perfect raw scores were unaffected by the adjustment performed in step 1. These scores were always converted to scale scores of 36, regardless of how many raw score units were added to them.)

3. Calculate the adjusted ACT Composite scale score (i.e., sum the adjusted scale scores for the four subject-area tests, divide the result by four, then round to the nearest integer).

4. Calculate the adjusted mean ACT Composite scale score ($\bar{X}_{adj}$).

5. Calculate the unadjusted mean ACT Composite scale score ($\bar{X}$).

6. Calculate the difference between the adjusted and unadjusted mean Composite scores:

$$\delta_c = \bar{X}_{adj} - \bar{X}.$$

The difference δ_c represents the increase in mean ACT Composite score that would be expected if each student were to answer one more item correctly on different subject-area tests, individually or in combination.

Steps 1 through 6 were repeated for 15 combinations of ACT subject-area tests, as shown in Table 2 on page 10. For example, one raw score unit was added to the English raw score of each student and the above steps were followed, yielding δ_c for the English test. Then, one raw score unit was added to the Mathematics raw score, to the Reading raw score, to the Science Reasoning raw score, to the English and Mathematics raw scores in combination, and so forth. At each stage, δ_c was calculated.

All analyses were performed by test form; the results were then summarized over forms. Separate analyses were performed on the data for the two samples of students who took Form F in 1994-95 and 1995-96. The entire process was then repeated, except that two and three raw score units, rather than one, were added to students' raw scores.

Subject-area score analyses. A similar process was used to examine relationships between mean subject-area score differences and raw scores. For the English, Mathematics, Reading, and Science Reasoning tests, the raw score of each student was adjusted by adding to it one raw score unit. Adjusted raw scores were converted to adjusted scale scores, and an adjusted scale score mean was calculated. The difference between the adjusted and unadjusted scale score means was then calculated by subject area:

$$\delta_s = \bar{Y}_{adj} - \bar{Y}.$$

The subscript *s* in the above equation refers to a subject-area score, rather than to a Composite score. Analyses for the subject-area scores were done by form, and the results summarized over forms. The effect of adding two and three raw score units to each student's raw score was also investigated.

Results

Composite score. Table 2 contains values of δ_c , by test form. The first column of this table shows the ACT Assessment test(s) to which one raw score unit was added. The letter "E" represents the English test; "M," "R," and "SR" represent the Mathematics, Reading, and Science Reasoning tests, respectively. For example, the first row shows results obtained when one raw score unit was added to the English raw score of each student. The last number in that row (0.1) indicates that the median effect, over forms, of adding one raw score unit to each student's English raw score was a 0.1 scale score unit increase in mean Composite scale score. One way to interpret this result is that a 0.1 difference between the Composite scale score means of two groups would result from one more English item answered correctly by each student in the group with the higher of the two means. Alternatively, a 0.1 difference in Composite scale score means would result from one more Mathematics item answered correctly by each student in the group with the higher of the two means (see row 2 of Table 2).

TABLE 2

Increase In Mean ACT Composite Scale Score From Adding One Raw Score Unit

Subject-area test	Form A	B	C	D	E	F[1]	F[2]	G	Median scale score increase (over forms)
E	.1	.1	.1	.1	.1	.1	.1	.1	.1
M	.1	.1	.1	.1	.1	.1	.1	.1	.1
R	.2	.2	.2	.2	.2	.2	.2	.2	.2
SR	.2	.2	.2	.2	.2	.2	.2	.2	.2
E, M	.2	.2	.2	.2	.2	.2	.2	.2	.2
E, R	.3	.3	.3	.3	.3	.3	.3	.3	.3
E, SR	.3	.3	.3	.3	.3	.3	.3	.3	.3
M, R	.3	.3	.3	.3	.3	.3	.3	.3	.3
M, SR	.3	.3	.3	.3	.3	.3	.2	.3	.3
R, SR	.4	.4	.4	.4	.4	.4	.4	.4	.4
E, M, R	.4	.4	.4	.4	.4	.4	.4	.4	.4
E, M, SR	.4	.4	.4	.4	.4	.4	.3	.4	.4
M, R, SR	.4	.5	.5	.5	.5	.5	.4	.5	.5
E, R, SR	.5	.5	.5	.5	.5	.5	.5	.5	.5
E, M, R, SR	.5	.6	.6	.6	.6	.6	.6	.6	.6

Notes: 1. Administered during the 1994-95 testing year.
2. Administered during the 1995-96 testing year.

The results shown in the remaining rows may be interpreted similarly: A 0.4 difference in the mean Composite scores of two groups, for example, would result from one more item correctly answered on each of the English, Mathematics, and Reading tests by each student in the higher scoring group (a total of three more items answered correctly; see the row labeled "E, M, R").

There were some minor differences between δ_c for the 1994-95 and 1995-96 administrations of Form F. The data for the 1994-95 administration, for example, yielded a median δ_c of 0.3 scale score units when one raw score unit was added to students' Mathematics and Science Reasoning scores. In comparison, the data for the 1995-96 administration yielded a median δ_c of 0.2. These differences may reflect sampling error or differences in skill levels of the two groups of students.

The differences between adjusted and unadjusted mean ACT Composite scores in Table 3 illustrate the effect of adding two more raw score units to each student's raw score. This table can be interpreted in the same manner as Table 2. Table 3 shows that δ_c ranged from 0.2 (two raw score units added to each student's English or Mathematics score) to 1.2 (two raw score units added to each student's English, Mathematics, Reading, and Science Reasoning scores).

TABLE 3

Increase In Mean ACT Composite Scale Score From Adding Two Raw Score Units

Subject-area test	Form A	Form B	Form C	Form D	Form E	Form F[1]	Form F[2]	Form G	Median scale score increase (over forms)
E	.2	.2	.2	.2	.2	.2	.2	.2	.2
M	.2	.2	.2	.2	.2	.2	.2	.2	.2
R	.4	.4	.4	.4	.4	.4	.4	.4	.4
SR	.3	.3	.3	.3	.4	.3	.3	.3	.3
E, M	.4	.4	.4	.4	.4	.4	.4	.4	.4
E, R	.7	.6	.6	.7	.6	.6	.6	.7	.6
E, SR	.6	.5	.6	.6	.6	.6	.5	.6	.6
M, R	.7	.6	.6	.6	.6	.6	.6	.6	.6
M, SR	.5	.5	.5	.5	.6	.5	.5	.5	.5
R, SR	.8	.7	.8	.8	.7	.8	.8	.8	.8
E, M, R	.9	.8	.8	.9	.8	.8	.8	.9	.8
E, M, SR	.8	.7	.8	.8	.8	.8	.8	.8	.8
M, R, SR	1.0	.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0
E, R, SR	1.0	.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0
E, M, R, SR	1.2	1.1	1.2	1.2	1.2	1.2	1.1	1.2	1.2

Notes: 1. Administered during the 1994-95 testing year.
2. Administered during the 1995-96 testing year.

The effects on δ_c of adding three raw score units to students' ACT Assessment raw scores are shown in Table 4. Median δ_c ranged from 0.3 (three raw score units added to either English or Mathematics) to 1.8 (three raw score units added to each of the four subject-area tests).

TABLE 4

Increase In Mean ACT Composite Scale Score From Adding Three Raw Score Units

Subject-area test	Form								Median scale score increase (over forms)
	A	B	C	D	E	F[1]	F[2]	G	
E	.3	.3	.3	.3	.4	.3	.3	.4	.3
M	.3	.3	.3	.3	.3	.3	.3	.3	.3
R	.7	.6	.6	.6	.6	.6	.6	.6	.6
SR	.5	.5	.5	.5	.5	.5	.5	.5	.5
E, M	.7	.6	.7	.7	.7	.6	.6	.7	.7
E, R	1.0	.9	1.0	1.0	1.0	.9	.9	1.0	1.0
E, SR	.8	.8	.9	.9	.9	.9	.8	.9	.9
M, R	1.0	.9	.9	.9	.9	.9	.9	.9	.9
M, SR	.8	.8	.8	.8	.9	.8	.8	.8	.8
R, SR	1.2	1.1	1.1	1.1	1.1	1.1	1.1	1.2	1.1
E, M, R	1.3	1.2	1.3	1.3	1.3	1.2	1.2	1.3	1.3
E, M, SR	1.2	1.1	1.2	1.2	1.2	1.2	1.1	1.2	1.2
M, R, SR	1.5	1.4	1.4	1.5	1.4	1.4	1.4	1.5	1.4
E, R, SR	1.5	1.4	1.5	1.5	1.5	1.5	1.4	1.5	1.5
E, M, R, SR	1.8	1.7	1.8	1.8	1.8	1.8	1.7	1.8	1.8

Notes: 1. Administered during the 1994-95 testing year.
2. Administered during the 1995-96 testing year.

Values of δ_c in Tables 3 and 4 are, in many instances, multiples of δ_c in Table 2. For example, median δ_c when one raw score unit was added to students' English raw scores was 0.1. When two and three English raw scores were added, median δ_c doubled (0.2; Table 3) and tripled (0.3; Table 4), respectively. There are exceptions to this pattern, however. For example, adding one, two, and three raw score units to Mathematics and Science Reasoning scores yielded median δ_c of 0.3, 0.5, and 0.8, respectively.

Subject-area scores. Table 5 illustrates the increase in mean subject-area scores resulting from adding one raw score unit to students' raw scores on each subject-area test. Median values of δ_s ranged from 0.4 (English, Mathematics) to 0.8 (Reading). When two raw score units were added to students' raw scores, median values of δ_s ranged from 0.8 (Mathematics) to 1.6 (Reading; see Table 6). Table 7 contains δ_s representing the effect of adding three raw score units. Median values of δ_s in this table ranged from 1.2 (Mathematics) to 2.4 (Reading).

TABLE 5

Increase In Mean ACT Subject-Area Scale Score From Adding One Raw Score Unit

Subject-area test	Form A	Form B	Form C	Form D	Form E	Form F[1]	Form F[2]	Form G	Median scale score increase (over forms)
English	.4	.4	.4	.4	.4	.4	.4	.5	.4
Mathematics	.4	.4	.4	.4	.4	.4	.4	.4	.4
Reading	.9	.8	.8	.9	.8	.8	.8	.8	.8
Science Reasoning	.7	.7	.7	.7	.7	.7	.6	.7	.7

Notes: 1. Administered during the 1994-95 testing year.
2. Administered during the 1995-96 testing year.

TABLE 6

Increase In Mean ACT Subject-Area Scale Score From Adding Two Raw Score Units

Subject-area test	Form A	Form B	Form C	Form D	Form E	Form F[1]	Form F[2]	Form G	Median scale score increase (over forms)
English	.9	.8	.9	.9	.9	.9	.9	1.0	.9
Mathematics	.9	.8	.8	.8	.8	.8	.8	.8	.8
Reading	1.8	1.6	1.6	1.7	1.6	1.6	1.6	1.7	1.6
Science Reasoning	1.3	1.3	1.4	1.4	1.4	1.4	1.3	1.4	1.4

Notes: 1. Administered during the 1994-95 testing year.
2. Administered during the 1995-96 testing year.

TABLE 7

Increase In Mean ACT Subject-Area Scale Score From Adding Three Raw Score Units

Subject-area test	Form A	Form B	Form C	Form D	Form E	Form F[1]	Form F[2]	Form G	Median scale score increase (over forms)
English	1.4	1.3	1.4	1.4	1.4	1.3	1.3	1.5	1.4
Mathematics	1.3	1.2	1.2	1.3	1.3	1.2	1.2	1.2	1.2
Reading	2.7	2.3	2.4	2.5	2.4	2.4	2.4	2.5	2.4
Science Reasoning	2.0	2.0	2.1	2.1	2.1	2.1	2.0	2.1	2.1

Notes: 1. Administered during the 1994-95 testing year.
2. Administered during the 1995-96 testing year.

Tables 5-7 can be interpreted in the same manner as the Composite score tables. The results in row 1 of Table 5 suggest, for example, that a 0.4 difference between the ACT English means of two groups would result from one more English item correctly answered by each student in the group with the higher of the two means.

A pattern similar to that found in the Composite score tables is apparent in Tables 5-7. For example, when one raw score unit was added to students' Mathematics raw scores, δ_s was 0.4. Adding two raw score units doubled δ_s (0.8; Table 6); adding three raw score units resulted in a threefold increase in δ_s (1.2; Table 7).

Additional Composite score and subject-area score analyses were performed in which one and two raw score units were subtracted from, rather than added to, the raw score(s) of each student. These analyses yielded results that were nearly identical to those described previously.

Discussion

The results of this study provide supplemental information for interpreting differences between mean ACT Assessment scale scores. Such differences are interpreted in terms of the number of correctly answered test items they represent. Consider, for example, a situation in which the mean ACT Composite score for group A (20.9) is 0.1 scale score unit less than that of group B (21.0). The results of this study indicate that this Composite score difference, expressed as the number of additional test items correctly answered by each student in group B relative to those in group A, is about one English item or one Mathematics item.

Methods for interpreting a difference between means, including the method described in this study, involve some subjectivity. For example, choosing an alpha level of .05 for a test of statistical significance is subjective; .07 could instead have been chosen. Similarly, determining the substantiveness of one additional ACT Assessment item correctly answered by each student in group B, relative to those in group A, is subjective. A school administrator, eager to demonstrate the worth of a new curriculum in which he or she has invested considerable effort developing, might infer from such a result that the curriculum is performing satisfactorily. A parent and taxpayer, on the other hand, may not be nearly as impressed with an improvement of one ACT Assessment item.

One important limitation of this study is that adding one or more raw score units to the raw scores of *each* student implies that all students benefited equally from any treatment (e.g., curriculum change) they might have received. A particular treatment does not usually affect all students equally. Adding an advanced mathematics course to a curriculum may, for example, increase the ACT Mathematics score only for middle- to high-scoring students.

With a slight modification to the method used in this study, one can simulate the effect of score increases for students whose scores are in a specific part of the score distribution. As might be expected, the results of such an approach are somewhat different from those described previously. For example, if only those students with ACT Mathematics scale scores between 21 and 29 each answered one more Mathematics item correctly, then δ_c would be smaller (.05) than that reported in the results section (.1; see Table 2, Form C results). This particular result suggests that a difference of .05 between two ACT Composite score means would occur if one more Mathematics item were correctly answered by each student in the higher scoring group who had a Mathematics scale score between 21 and 29. Additional research examining the effect on δ_c and δ_s of adding one or more raw score units to raw scores in different areas of the score distribution might be useful for further interpreting ACT Assessment mean differences.

Examples

The examples presented below are intended as a refresher for school administrators who are responsible for describing to others the ACT Assessment performance of their students. Included in the examples is a practical application of the number of items correct method.

Data for XYZ School District

Table 8 contains frequency distributions of the ACT Composite scores of high school students in the XYZ School District, by graduating class. This is a small district with only one high school. Most, but not all, of the students take the ACT Assessment. Composite score means, standard deviations, and the number of ACT-tested students in each graduating class are shown at the bottom of the table.

TABLE 8

Distribution of ACT Composite Scores for XYZ School District, by High School Graduating Class

	Frequency	
ACT Composite score	Class of 1996	Class of 1997
36	0	1
35	1	0
34	0	1
33	2	1
32	3	5
31	2	5
30	5	3
29	4	4
28	10	6
27	7	12
26	5	10
25	7	10
24	7	11
23	8	10
22	8	8
21	10	13
20	14	9
19	8	13
18	6	6
17	10	9
16	6	7
15	3	2
14	5	0
Mean	22.6	23.3
Std. dev.	5.0	4.7
N	131	146

XYZ's mean ACT Composite score for 1997 (23.3) is somewhat higher than that for 1996 (22.6). Is the difference between these two means statistically significant? Is it substantive? There are several methods that can be used to address these questions.

The first two methods described below (two-sample t test, confidence interval plot) address the question of statistical significance. As previously discussed, statistical significance refers to the probability of an observed result being reasonably thought of as due to chance.

The third and fourth methods (effect size, number of items correct) address the question of substantiveness. The methods typically are most effective when used in combination. Reporting t test results together with an effect size, for example, is more informative than reporting the results of either method alone.

Two-Sample t Test

For illustrative purposes, assume that a research analyst in the XYZ district office receives an inquiry from a member of the school board who, after reviewing the 1996 and 1997 Composite score means, wants to know whether the difference between them is statistically significant. One option available to the analyst in this case is a two-sample t test.

The research analyst's goal in performing the two-sample t test is to compute a test statistic (sometimes called a t ratio) that will determine the probability of a particular result, given that the null hypothesis (there is no difference between means) is true.

The null (H_0) and alternative (H_1) hypotheses are

H_0: $\mu_1 = \mu_2$ (the population mean of Group 1 is the same as that of Group 2)

H_1: $\mu_1 \neq \mu_2$ (the population means differ).

It is assumed that the observations within each population are normally distributed and have equal variances. It is further assumed that the observations within or between each population are independent. Note that the alternative hypothesis is two-sided, or nondirectional. A directional hypothesis (e.g., H_1: $\mu_1 < \mu_2$) would be appropriate only if it were specified in advance of viewing the data in Table 8 that some treatment (e.g., a curriculum change) would be related to higher ACT scores. In the example presented here, the question about differences between means occurred after the data had already been obtained. In addition, we will assume that the research analyst has no information about whether a treatment of some type is reflected in the ACT scores of XYZ's 1997 graduates.

One formula for computing a test statistic (t^*) is

$$t^* = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} ,$$

where $\bar{x}$ and n represent the sample mean and number of students, respectively. The pooled sample standard deviation is estimated by

$$s_p = \sqrt{\frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} (n_i - 1)\, s_i^2} .$$

The subscripts denote the group (1 or 2) whose ACT scores yielded these statistics.

After substituting means, standard deviations, and sample sizes from Table 8 into Equation 1, XYZ's research analyst calculates a t^* of approximately 1.2. Assuming a level of significance (α) of .05, and referring to a table of the t distribution with $n_1 + n_2$

- 2 = 275 degrees of freedom, the analyst determines that H_0 will be rejected if $t^* < -1.96$ or $t^* > 1.96$. Because the observed value of t^* is less than 1.96, the analyst decides to accept H_0, concluding that the difference between ACT Composite score means of the two groups is not statistically significant.

Alternatively, XYZ's research analyst could use a statistical software package to perform a two-sample t test. For example, if the data in Table 8 were analyzed with the SAS System's (1990) TTEST procedure, the absolute value of t^* would again be found to be about 1.2. The corresponding probability (p value) would be .23. Given this information, the research analyst would conclude that the probability that the result (a difference between means of 0.7 or more) can reasonably be thought of as due to chance is .23.

Because the two-sample t test assumes independence of observations, it may be inappropriate when data from different years are correlated in some way. For example, if a new curriculum were implemented prior to the first year in which data were collected, then the curriculum may be a factor related to test performance both in the first year and in the second year, provided that the curriculum is continued through the second year. In this situation, statistical procedures other than a t test would likely provide relatively more accurate information about the statistical significance of differences between test score means.

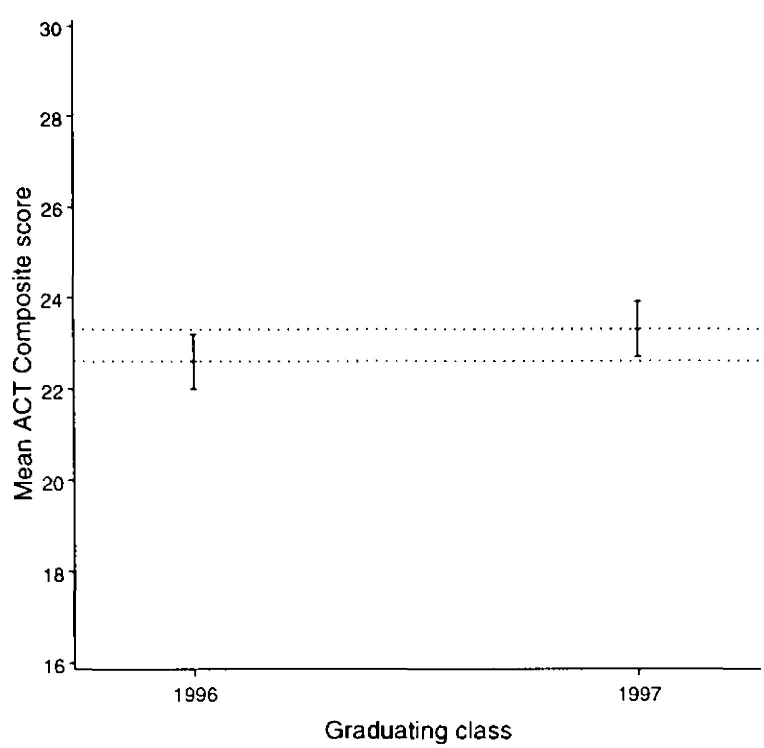
The two-sample t test is appropriate for comparing the means of two independent groups. For three or more independent group means, a method such as analysis of variance (ANOVA) is appropriate.

Confidence Interval Plot

Another way to determine whether two mean ACT scores are statistically significantly different is with a confidence interval plot. An example of this type of plot, based on the data in Table 8, is shown in Figure 1. Mean ACT Composite scores for XYZ's graduating classes of 1996 and 1997 are displayed in this figure. The two dotted, horizontal lines pass directly through the means of these two groups (22.6 and 23.3, respectively). Ninety-five percent simultaneous confidence intervals are shown around each mean.

FIGURE 1. Mean ACT Composite Scores for XYZ School District

(95% simultaneous confidence intervals)



One advantage to using a confidence interval plot is that it provides a concise graphical representation of the difference between means. In addition, it allows one to visually determine, via the confidence intervals, whether the difference between means is statistically significant.

The simultaneous confidence intervals shown in Figure 1 are based on the Tukey-Kramer multiple comparison procedure. For two groups with fairly similar, but unequal sample sizes, an approximation of a 95% simultaneous confidence interval is

$$\bar{X}_i \pm \frac{1}{2} q \frac{s_p}{\sqrt{n_i}},$$

where $\bar{X}_i$ = sample mean for group i, and

q = studentized range statistic with $n_1 + n_2 - 2$ degrees of freedom.

Tables showing percentiles of the studentized range distribution are available in many statistical textbooks. Additional information on simultaneous confidence intervals can be found in Neter, Kutner, Nachtsheim, and Wasserman (1996, chap. 17).

Because the 95% confidence intervals for the two means displayed in Figure 1 overlap, XYZ's research analyst concludes that they are not statistically significantly different from each other at an α level of .05. Had the confidence intervals not overlapped, the analyst would have concluded that the difference between the mean Composite scores was statistically significant.

It is worth noting that because there are only two means in this example, confidence intervals could have been calculated using an alternative (nonsimultaneous) procedure that does not rely on the studentized range statistic. The Tukey-Kramer procedure was illustrated here because it can be used to compare two or more means.

Effect Size of Mean Differences

By itself, a test of statistical significance will not afford much information about the substantiveness of the difference between two mean ACT Assessment scores. For this reason, it is often helpful to supplement the information provided by such a test with an estimate of an effect size.

An effect size for the difference between two means, assuming equal population variances, is estimated by

$$\Delta^{*} = \frac{\bar{X}_1 - \bar{X}_2}{s_p} .$$

In this equation, $\bar{X}_1$ is the larger of the two sample means. Additional information on effect sizes can be found in Cohen (1988), Glass (1976), Glass, McGaw, and Smith (1981), and Hedges and Olkin (1985). Effect sizes may also be calculated for three or more means. A discussion of effect sizes in ANCOVA, for example, is presented in Cohen (1988, chap. 8).

For XYZ School District's data, $\Delta^* = 0.14$. Cohen's guidelines for interpreting effect sizes (described on page 2 of this report) suggest that this effect size, representing about 0.14 ACT Composite standard deviation units, is fairly small. On the basis of this result, XYZ's research analyst concludes that the mean Composite scores for the 1996 and 1997 graduating classes are not substantively different. This finding parallels those of the two-sample t test and confidence interval plot, which indicated that the means were not statistically significantly different.

The estimated effect size can be used in combination with either the two-sample t test or the confidence interval plot. Alternatively, all three can be used together.

Number of Correctly Answered Items

The difference between XYZ School District's 1996 and 1997 mean Composite scores is 0.7 of a scale score unit. Table 3 in the first part of this report does not contain a median δ_c of 0.7, but median δ_c of 0.6 and 0.8 are shown for several test combinations. For example, a difference between means of 0.8 would result from about two more items correctly answered by each student in the 1997 graduating class on both the Reading and Science Reasoning tests (see row 10 of Table 3). If 1997 graduates had correctly answered about two more items on each of the English, Mathematics, and Reading tests, or on each of the English, Mathematics, and Science Reasoning tests, a similar result would have been obtained (see rows 11 and 12).

The ACT Assessment has a total of 215 items. Four to six additional items answered correctly by the 1997 graduates across all of the subject-area tests, relative to the 1996 graduates, is not a large difference, practically speaking. The size of this difference may, however, be interpreted differently by others. An analyst might therefore discuss the difference with colleagues before reaching a final conclusion.

The results yielded by the number of items correct method have the advantage of being practical and easily understood. The results may be used to supplement those of other methods. One could, for example, report a t ratio, its associated p value, an effect size, and the number of items represented by a particular difference between means.

Discussion

Summary. The two-sample *t* test and confidence interval plot both indicated that the difference between the mean ACT Assessment scores for the 1996 and 1997 graduating classes in XYZ School District was not statistically significant. Effect sizes suggested that the difference between means was fairly small (about 0.14 of an ACT standard deviation) and not substantive, according to common guidelines. The difference between means, when expressed as the number of additional test items answered by the graduating class with the higher of the two means, was about four to six items out of a total of 215. This did not appear to be a large difference from a practical perspective.

Correlates of ACT Assessment performance. XYZ School District staff might wish to do some further investigation to determine whether characteristics of their high school curriculum are related to the increase in mean ACT Assessment score for 1997. One important correlate of ACT Assessment performance that district staff could investigate is course-taking patterns. It has been shown, for example, that students who took and/or planned to take 3½ years of English earned higher ACT English scores than those who took and/or planned to take 2 years or less. Similar findings have been documented for the ACT Mathematics test (ACT, 1997; Harris & Kolen, 1989). Moreover, increased course taking is, in some instances, associated with higher ACT scores regardless of grades earned. For example, students' ACT Mathematics scores were found to increase, on average, by 1.3 scale score units for each additional mathematics course taken. This occurred regardless of the grades students earned in English, mathematics, or natural sciences courses (ACT, 1997, p. 41).

On the basis of previous research on course taking and ACT Assessment score relationships, XYZ district staff might decide to examine the ACT Assessment scores of students who took and/or planned to take 3½ years of English versus those who took and/or planned to take 2 years or less. Perhaps the lower scoring group (1996 graduating class in this example) contains a relatively larger proportion of students who took only two years or less of English. Perhaps relatively more students in the higher scoring group (1997 graduates) took 3½ years of English. If this were the case, then a positive relationship between course taking and ACT Assessment performance likely exists for XYZ students. XYZ district staff may therefore choose to take steps to ensure that most of their students complete at least 3½ years of English course work.

Trends in mean ACT scores. Year-to-year fluctuations in mean ACT Assessment scores are fairly common, and may be subject to overinterpretation. It is therefore advisable to examine trends over time in mean ACT Assessment scores, which provide relatively more information about students' performance. For example, consistent mean ACT score increases occurring over a five-year period would likely provide stronger evidence of a positive relationship between increased course taking and test performance than would a one-year mean ACT score increase.

Rounding error in mean ACT scores. The mean ACT Assessment scores that ACT reports to users are rounded to the nearest one-tenth of a scale score unit. The error resulting from rounding may sometimes exaggerate the size of differences between ACT Assessment means. Means of 21.43 and 21.47, for example, would round to 21.4 and 21.5, respectively. The difference between the rounded means in this example (0.1) is larger than the difference between the unrounded means (0.04). The fact that rounding error may be reflected in reported ACT Assessment means should be considered when interpreting differences between them.

References

ACT. (1997). ACT Assessment technical manual. Iowa City, IA: Author.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61 (4), 287-292.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5 (10), 3-8.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Harris, D. J., & Kolen, M. J. (1989). Examining group differences from a developmental perspective. In R. L. Brennan (Ed.), Methodology used in scaling the ACT Assessment and P-ACT+ (pp. 75-86). Iowa City, IA: American College Testing Program.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Academic Press.

Kolen, M. J., & Hanson, B. L. (1989). Scaling the ACT Assessment. In R. L. Brennan (Ed.), Methodology used in scaling the ACT Assessment and P-ACT+ (pp. 35-55). Iowa City, IA: American College Testing Program.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models (4th ed.). Chicago: Irwin.

SAS Institute, Inc. (1990). SAS/STAT user's guide, version 6, fourth addition, volume 2. Cary, NC: Author.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested performs. Educational Researcher, 25 (2), 26-30.