How do we know what a safe learning environment is? I suspect your response will be different from mine—and my response different than my colleague's. And, what is your definition of "conducive to learning"? I suspect it's different than mine.

OK, what am I getting at? I am not convinced that we have any quantitative data that helps us with these questions. For example, what data would we have that "identifies a safe learning environment"? And what data would we have collected that would help us with identifying a "drug-free" environment? Does it mean that we have no evidence of marijuana use but some students smoke cigarettes? Or does it mean that there are no incidents of smoking on campus but we realize that smoking at home occurs?

My point is that I am not sure we have those data available—and that we would need to utilize qualitative methods (or at least mixed methods) to get at the information we need to be able to address the NCLB goal of "students will be educated in learning environments that are safe, drug free, and conducive to learning."

So, develop a strategy for collecting qualitative data that gives you a base or benchmark for pursuing the NCLB goal. How will you determine if an environment is "drug free"? How will you determine if a learning environment is "safe"? And how will you determine if an environment is "conducive to learning"?

CHAPTER THIRTEEN

Putting It All Together

An Evidence-Based Practice Field

In earlier chapters, I emphasized that educators rarely examine the data existing in schools to assess in a systematic way the quality of teaching and learning in our schools. Much of the fault lies with our universities and their failure to adequately prepare teachers and administrators to deal effectively with data (and evidence). I agree totally with Holcomb's (2005) contention that existing programs at the university level (and the course delivery of statistics and tests and measurement classes) have created a structure that emphasizes esoteric experimental designs that cannot be replicated in a normal school setting. I also echo Bracey's (1997) belief that in all too many instances we teach statistics and other data-related courses in a theoretical manner based on hard-to-understand formulas and far too many examples unrelated to the daily life of educational practitioners.

Yes, I agree that much progress has been made since the first edition of *Schools and Data* was published in 2001. But a new problem has surfaced! In our response to the recent accountability movement (for example, No Child Left Behind [NCLB]), we need to be careful that we do not "just collect data for the sake of collecting data." And let's focus our attention beyond the data to look closely at the *evidence*. I have related my thoughts on *data* versus *evidence* in earlier chapters. And for the second edition of *Schools and Data*, I have changed the subtitle and focus of this chapter the coefficient if it meets the more stringent criteria of .001. Not sure what this may mean at this point—but you need to think about why this might be the case.

Second, we see another significant correlation—between science and writing ($r = .252$, $p = .011$). Again, this correlation is positive, so it means that as students tend to score (up or down) in science, they have a similar tendency in writing. Not quite as strong as the language and science, but still significant at the .05 significance level.

We will not go further here—but you see there are many other observations we can make with this particular analysis that will help us with more *evidence-based decision making.* For example, we would assume that students who score well in language would also have a tendency to score well in writing. With our 100 students, this is not the case, as there is not noticeable or significant correlation in that direction. Really, the lack of a correlation means that there is no pattern between these two subject areas. Hmmm. We might want to investigate this situation further. We might want to discuss collaborative cross-curricular activities with our two faculty teaching writing and language—or if the same instructor teaches both, are both subjects being included in the curriculum? Just an idea!

Conclusion

Argyris and Schoen, in their book *Organizational Learning* (1978), stated that people function with a gap between their espoused theories (what they believe to be the right course of action) and their theories in use (what they choose to do given the surrounding circumstances). Sound like our life in the schools? I think so. The authors continue by saying that failure to recognize and close those gaps impedes organizational learning.

The evidence-based practice field can help close some of those gaps. A crucial component, however, is the opportunity to take what has been learned in the practice field and apply it to the real life of the school and the community. So creating practice fields is important but is in itself inadequate (Kim, 1995). There is danger in allowing practice fields to become training grounds as an end in themselves. The learning experiences are no better than the old

traditional ways of doing things if they are not moved and adapted to the workplace.

I continually state that the real culprits in this dilemma are the university teacher and administrator preparation programs. Though there are a few "bright spots" in some preparation programs, for the most part there is not a substantive attempt to increase teachers' and administrators' understanding of data analysis or the use of analysis to improve real teaching and learning.

Gravetter and Wallnau introduced their excellent book *Statistics for the Behavioral Sciences* (2000) by asking their readers to read the following paragraph, which was adapted from a psychology experiment by Bransford and Johnson (1972). I would like to use the same exercise as a conclusion to this book:

> The procedure is actually quite simple. First you arrange things into different groups depending on their makeup. Of course, one pile may be sufficient, depending on how much there is to do. If you have to go somewhere else due to the lack of facilities, that is the next step; otherwise you are pretty well set. It is important not to overdo any particular endeavor. That is, it is better to do too few things at once than too many. In the short run this may not seem important, but complications from doing too many easily arise. A mistake can be extensive as well. The manipulations of the appropriate mechanisms should be self-explanatory, and we need not dwell on them here. At first, the whole procedure will seem complicated. Soon, however, it will just become another facet of life. It is difficult to foresee any end to the necessity of this task in the immediate future, but then one never can tell. (Gravetter & Wallnau, 2000, p. 4)

Perhaps the quoted paragraph sounds like some complicated statistical procedure. But it actually describes the everyday task of doing laundry. Knowing this, now go back and read the passage again.

The authors' purpose, like mine here, was to point out the importance of context. When things are out of context, even the simplest procedures can seem complex and difficult to understand. I suggest that this has been one of the problems with the instructional delivery of data-related courses in teacher and administrator preparation programs. They lack context and applicability.

from "A Data-Driven Practice Field" to "An Evidence-Based Practice Field."

Why the Change to Evidence-Based Decision Making?

Data-based decision making is the current buzz term in education circles today. But since the publication of the first edition of *Schools and Data*, I have come to be concerned that not only do we disagree on the meaning of data-based decision making, but that our confidence in data analysis in our schools is mistaken. Let me explain my position further.

I use as an example elsewhere (Creighton, 2005) the comparison of attendance rates with absence rates. You are familiar with the term *average daily attendance* (ADA). We collect these data for several reasons, but primarily they are used to determine the number of dollars we receive in state and federal monies. Simply put, if students are present, schools receive additional funding to help educate them. In addition, schools are now given accountability ratings (NCLB Act) based upon ADA. So principals and superintendents give heavy emphasis on implementing strategies to keep attendance rates high.

Many of the schools receiving the highest accountability rating ("Exemplary") report ADA in a range of 92% to 98%. A 92% sounds pretty good and even has a connotation of a grade of A, right? So we, as school administrators, report that on average, we have a 92% attendance rate.

We have accurately reported the data—*but*, the evidence reveals something completely different and troublesome. If the average daily attendance rate is 92%, what is the average daily absence rate? Well, that's easy—8%. Investigating a little further, we discover that if, on average, 8% of our students are absent, and based on a year of 180 days of instruction, students on average are missing approximately 2 weeks of school per year (8% of 180 equals 14.4 days). Basing a school's Exemplary rating on 92% attendance is one thing, but students on average missing 2 weeks of school each year is quite another matter. You may argue that focusing on attendance is no different than focusing on absence. Well, it is all a matter of perspective, and reflects the difference

between *data-driven decision making* and *evidence-based decision making*. Reporting on attendance is a matter of reporting the data. The absence rate is evidence that is not reported or considered to be existing data. I also argue that by looking at the evidence rather than the data, we are more likely to touch more of our *marginal* students—those placed at risk of educational failure. Often, these students are not represented in the data. For example, when we report attendance rates or graduation rates, missing are students not in attendance and students not graduating.

This is my concern now—and the reason for my new approach in this second edition. We have been quick to report the data—and that's good. But in our haste to look at the data, we have been missing the more important evidence. I will not belabor this point further, but do reference a more detailed treatment of my position in *Leading From Below the Surface* (2005), also published by Corwin Press. You can access this book by visiting the *Schools and Data* Web site at: http://www.schoolsanddata.org.

Practice and Performance Fields

For some time, I have studied the work of Daniel Kim (1995), an organizational consultant and public speaker who is committed to helping problem-solving organizations transform into learning organizations. Kim is a colleague of Peter Senge (*The Fifth Discipline*, 1990) and cofounder of the MIT Organizational Learning Center, where he is currently the director of the Learning Lab Research Project.

Kim (1995) argued that the accelerated pace of change has overwhelmed our ability to keep up with and understand how these changes affect our organizations. We lack a place to practice decision making where we can make mistakes and step out of the system temporarily so that we can work "on the system" and not just "in the system." He calls this kind of a place a "managerial practice field."

Think for a minute about our jobs in education. Except for a brief experience with some sort of internship or student teaching (both of which continue to suffer from a lack of quality and relevance), where and when do we get an opportunity to leave the day-to-day pressures temporarily and enter a different kind of

Figure 13.5 Two-Way ANOVA for Ethnicity, Gender, and Math Scores

Descriptive Statistics

Dependent Variable: math

male=1, female=2	*white=1, black=2, hispanic=3*	*Mean*	*Std. Deviation*	N
male	White	242.57	8.218	37
	Hispanic	245.60	10.319	10
	Black	249.67	8.262	6
	Total	243.94	8.798	53
female	White	245.58	9.483	19
	Hispanic	248.22	5.847	18
	Black	249.20	9.578	10
	Total	247.36	8.253	47
Total	White	243.59	8.701	56
	Hispanic	247.29	7.659	28
	Black	249.38	8.823	16
	Total	245.55	8.674	100

Tests of Between-Subjects Effects

Dependent Variable: math

Source	*Type III Sum of Squares*	*df*	*Mean Square*	F	*Sig.*
gender	53.194	1	53.194	.740	.392
ethnic	379.338	2	189.669	2.639	.077
gender * ethnic	35.707	2	17.854	.248	.781

a. $r^2 = .093$ (Adjusted $r^2 = .045$)

schools can be addressed or answered with correlations. Though there are several analyses we can conduct, let's take a look at whether or not there might be a correlation among our four subject areas (math, science, language, and writing). The specific questions might be:

1. Do our students seem to perform similarly in the four subject areas across subjects? In other words, do students who score high in math also have a tendency to score high in science (and any other combination of comparisons)? This answer to this type of question would be revealed by a positive correlation coefficient (for example, $r = .67$).

2. Do our students seem to perform dissimilarly in the four subject areas across subjects? In this case, we are interested to see if, for example, students who score high in math might have a tendency to score lower in language or writing. The answer to this kind of question might result from a negative correlation coefficient (for example, $r = -.65$). If one set of scores tends to be high, while another set of scores goes down, there is a negative correlation.

Figure 13.6 displays the Pearson correlations for our students and their four subject area scores. Remember, the cells that show 1 are simply a correlation of the same variable (that is, math correlated with math). We disregard these cells. Also, be aware that each cell is repeated—because of the order of the variables being correlated (for example, math with science is no different than science with math).

Figure 13.6 Math, Science, Language, and Writing Correlation

Correlations

		math	language	science	writing
math	Pearson Correlation	1	−.004	.004	.004
	Sig. (2-tailed)		.969	.971	.967
	N	100	100	100	100
language	Pearson Correlation	−.004	1	.334(**)	.183
	Sig. (2-tailed)	.969		.001	.068
	N	100	100	100	100
science	Pearson Correlation	.004	.334(**)	1	.252(*)
	Sig. (2-tailed)	.971	.001		.011
	N	100	100	100	100
writing	Pearson Correlation	.004	.183	.252(*)	1
	Sig. (2-tailed)	.967	.068	.011	
	N	100	100	100	100

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

We see several things of interest. First, the largest and statistically significant relationship exists between science and language ($r = .334$, $p = .001$). This means that as students tend to score (up or down) in science, they have a strong tendency to score in the same direction in math. You notice SPSS places two stars next to and Blacks show .046. OK, significant, but really pretty close to .05—this finding is not as powerful, so we must be careful in our interpretation. But suffice it to say that we have a significant difference—reflecting back to earlier chapters, we likely have a *practical significance* here rather than a strong significant difference.

Though there are differences between the mean scores of Whites and Hispanics and between Hispanics and Blacks, they are not significant (.147 and .710), being much greater than the selected criteria of .05.

The last printout displayed in Figure 13.4 is one of homogeneous subsets. This analysis places like sets together (little difference) and unlike (difference) sets apart. So we see here that Subset 1 consists of Whites and Hispanics and Subset 2 consists of Hispanics and Blacks. The Hispanics' mean is close to the Whites', but also shares closeness with the Blacks'. What this really means is that the greatest difference is between the Whites and Blacks—*but*, nonetheless, there is also less difference among the Hispanics and Whites and Blacks. Sounds confusing, perhaps, but we see a very serious pattern of concern here. There are disparities among all three groups, though one more serious than the others. We must investigate further to see what we might be able to do in regard to curriculum and instruction to reduce or eliminate these disparities among ethnic groups and math achievement. More *evidence-based decision making!*

Activity 7. Two-Way Analysis of Variance

Analysis of variance looks for significant differences between groups by comparing the means of those groups with some selected variable. In the one-way ANOVA described in Activity #6, we were interested to see if the three ethnic groups differed from each other on their performance on standardized math assessment scores. The one-way part of the analysis indicates that there is only one independent variable (the ethnic groups) and only one dependent variable (math scores).

The two-way ANOVA allows us to use two independent variables. We are now interested to discover if there is a relationship (as measured by math scores) between ethnic group and gender (two independent variables). The two-way ANOVA will allow us to determine if gender or ethnic group or an interaction between gender and ethnic group has an effect on math achievement.

Questions

1. Do females and males differ significantly in math performance as measured by standardized test scores? This question addresses the *main effect* for gender.
2. Do students in the three ethnic groups differ significantly in math performance as measured by standardized test scores? This question addresses the *main effect* for ethnic group.
3. Is there an interaction between gender and ethnic group? This question addresses the possibility, as an example, that Hispanic females score higher in math but White females score lower.

Discussion

Figure 13.5 displays descriptive statistics and the ANOVA tests of between-subjects effects.

Let's use the information in Figure 13.5 to answer the three questions posted at the beginning of Activity #7:

1. There is no statistically significant main effect for gender. Females (mean = 247.36) did not score significantly higher than males (mean = 243.94), $F = .74$, $p = .405$.
2. There is no significant main effect for ethnicity. Though the test did not meet the criteria of .05 significance level, .077 is not too distant. Again here, we may want to emphasize *practical significance.* Blacks had higher math scores in both the male and female categories. However, the difference is not considered significant because the difference was not great enough to meet our selected .05 criteria.
3. There is no statistically significant *interaction* between gender and ethnicity, $F = .242$, $p = .781$. Actually, this is good news. Math performance seems to be consistently similar across gender and ethnic groups. If this were not the case, we would have cause for greater alarm.

Activity 8. Pearson Correlation

I saved the last activity in the evidence-based practice field for last, since I want to make the point that much of what we do in space in which we can practice and learn? Think another moment about your limited training and preparation in the use of data and evidence to improve decision making. What has been missing? The opportunity or place does not exist where aspiring teachers and administrators can improve their skills in problem analysis, program and student evaluation, data-based and evidence-based decision making, and report preparation. The central idea here is that a *leadership practice field* provides an environment in which prospective school leaders can experiment with alternative strategies and policies, test assumptions, and practice working through the complex issues of school administration in a constructive and productive manner.

I can think of no other profession that does not value or provide opportunities for new professionals to practice in a different kind of space, where one can simply practice and learn. When practicing a symphony, the orchestra has the ability to *slow down* the tempo in order to practice certain sections. A medical student in residence has the opportunity to *slow down* and practice certain medical diagnoses and procedures. The New York Knicks spend most of their time in a practice field, slowing down the tempo, and practicing certain moves, strategies, and assumptions. All of these practice fields exist in an environment with opportunities for making mistakes, in a "safe-failing space to enhance learning" (Kim, 1995).

Let me return to the work of Daniel Kim for a moment.

> Imagine you are walking across a tightrope stretched between two large buildings. The wind is blowing and the rope is shaking as you inch your way forward. One of your teammates sits in the wheelbarrow you are balancing in front of you, while another perches on your shoulders. There are no safety nets, no harnesses. You think to yourself, "One false move and the three of us will be taking an express elevator straight down to the street." Suddenly, your trainer yells from the other side, "Try a new move! Experiment! Take some risks! Remember, you are a learning team!" (Kim, 1995, p. 353)

Kim argued that although this may sound ludicrous, it is precisely what many companies expect their management teams to do: experiment and learn in an environment that is risky, turbulent, and unpredictable. Unlike a high-wire act, a sports team, musician, or physician, however, management teams do not have a practice field in which to learn; they are nearly always on the performance field.

Replace the words *management teams* with *principals* or *school leaders* in the preceding paragraph. Will you not agree that our jobs in education resemble the *risky, turbulent,* and *unpredictable* analogy described by Kim?

An Evidence-Based Practice Field

It is my desire in this final chapter to provide you with a "practice field" designed around the actual work we do in schools. This practice field will also give you the opportunity to practice each of the data analysis strategies discussed in earlier chapters. In addition, you will have the opportunity to experiment with alternative strategies, test some of your assumptions, and practice working with the data found in your schools.

One of the goals of this chapter is to provide a *real enough* practice field so that the activities are meaningful to you as an educational leader but also *safe enough* to encourage experimentation and learning. My hope is that you will be able to step out of the day-to-day pressures faced in your workplace (or your graduate research class) and spend time in the practice field. Musicians practice. World-class athletes practice. But educators, for the most part, are constantly performing.

Again, part of the problem relates to teacher and administrator preparation programs at the university level. Because I am more familiar with and involved in administrative preparation, allow me to make a few comments. Murphy and Forsyth (1999) reported that although supervised practice could be the most critical phase of the administrator's preparation, the component is notoriously weak. Along with other educational leaders (Griffiths, 1999; Milstein, 1990), Murphy claims that field-based practices do not involve an adequate number of experiences and are arranged on the basis of convenience. I am not happy to report that since the first edition of *Schools and Data,* not much has changed. Yes, there are bright spots on the horizon; yes, there are new and innovative approaches to the preemployment experiences of the school leader. But, the bright spots are few—and are the exception rather than the rule. So, here we are again—five years later, and facing the same dilemma of not providing the appropriate preparation for principals and other school leaders in the important work of evidence-based decision making.

I also believe that many of our individual courses in research and statistics should have a practice field where educators can apply what they learn away from the pressures of day-to-day school business. What follows is an attempt to place you in such a practice field. *Try a new move! Experiment! Take some risks! Remember, you are a learning team!*

Horizon High School: A Practice Field

Using the data found in Resource E, create a file titled "Horizon High School Data." The data represent 100 students in Grades 9 through 12. Though the data are fictitious, the variables and entries represent the same kind of data found in your schools. Table 13.1 displays the variables and coding used.

Table 13.1 Coding of Variables for Horizon High School

Gender	Gender of the student: 1 = male, 2 = female
Ethnicity	Ethnicity of the student: 1 = White, 2 = Hispanic, 3 = Black
Freeredu	Qualification for free/reduced lunch: 1 = yes, 2 = no
Math	Standardized test scores in math
Language	Standardized test scores in language
Science	Standardized test scores in science
Writing	District-created writing assessment scores
Grade	Present grade level of student
GPA	Current grade point average of student

Activity 1. Frequency Distributions and Cross-Tabulation

With the use of frequency distributions and cross-tabulations in SPSS or Excel, we would like to take a quick glimpse of the Horizon High School student body. For example, we want to first know:

1. The number of males and females
2. The ethnic makeup of the school
3. The number of students qualifying for free or reduced lunch
4. The number of students in each grade

Though you may not consider the number of males and females or the ethnic makeup of the school to be terribly important, the information regarding ethnicity and free or reduced lunches is used for most of your funding formulas, grant applications, and state and federal reimbursements.

You might want to obtain a graphic representation of your frequency distributions for Horizon High School. Simply select **Chart**, and then select the desired form, such as **Bar Chart**. *Experiment* with several different kinds.

Questions

1. Is there anything unusual about the number of boys qualifying for free or reduced lunch compared to the number of girls? For what reasons might this information be relevant or helpful?
2. Does the gender representation seem normal for an average high school? Is there any weighted effect between gender and any of the other variables?

Discussion

Tables 13.2, 13.3, and 13.4 display some of the information that we need to address these questions.

Table 13.2 Gender of Horizon High School Students

Gender	*Number*	*%*	*Cumulative %*
Male	53	53	53
Female	47	47	100
Total	100	100	

Table 13.3 Gender of Horizon High School Students by Free or Reduced Lunch Qualification

	Free or Reduced Lunch?		
Gender	*Yes*	*No*	*Total*
Male	13	40	53
Female	19	28	47
Total	32	68	100

Table 13.4 Ethnicity of Horizon High School Students

Ethnicity	*Number*	*%*	*Cumulative %*
White	56	56	56
Hispanic	28	28	84
Black	16	16	100
Total	100	100	

Examining data by the use of frequency distributions and tabulation tables allows us to determine very quickly and easily what our population or sample looks like. Such information as gender, race, age, and socioeconomic status is helpful *before* we proceed with further investigation.

Activity 2. Measures of Central Tendency and Variability

You serve as the chairperson of the Horizon High School Curriculum Committee. The director of secondary education in your district has just notified you that she wants a recommendation from you and the committee for this year's textbook adoption. As is the practice at Horizon High School, the board of education only allows one subject area to have new textbooks annually. You and your committee must decide which subject area would benefit most from receiving this year's new textbooks.

Looking at the standardized test scores and the district writing assessment, can you use the measures of central tendency and variability to help you and your committee make a wise decision (based on the evidence) and recommendation to the director of secondary education?

Questions

1. Based on the standardized test data you have for Horizon High School, what subject area might you and your committee recommend for new textbooks this year? And what reasons would you give for your decision?
2. The mean and standard deviation of which subject area might concern you? Approximately 68% of your students scored between what two points in that same subject area? Does it perhaps make sense that the standard deviation on this subject area might be higher than in other subjects? Why or why not?

Discussion

Table 13.5 displays some of the information to address the questions above.

Table 13.5 Horizon Students' Descriptive Statistics of Test Score Data

Subject	N	*Minimum*	*Maximum*	*Mean*	*SD*
Math	100	234	263	246.86	7.69
Language	100	232	260	248.26	6.38
Science	100	220	276	245.41	8.51
Writing	100	175	234	205.76	11.88

We do notice that the mean scores for the writing assessments are much lower (on average) than those for other subject areas. In addition, the standard deviation is higher than the others, meaning that the writing scores are a bit more spread out from the mean. Specifically, we say that approximately 68% of the high school students scored between 194 and 218 on their writing assessment scores. The standard deviation is 11.88 (rounded to 12), and the mean is 205.76 (rounded to 206). The 12 points represent one standard deviation unit, and we know that approximately 68% of the distribution (students) falls between one standard deviation unit below the mean and one standard deviation unit above the mean.

The scores of the other subject areas are noticeably higher and the standard deviations are a bit smaller, indicating that those scores are centered a little more closely around the mean. Though we need to take many more factors into consideration (evidence), the writing scores give us reason to perhaps suggest new instructional textbooks for that subject area.

Activity 3. A One-Sample *t* Test

The national testing service responsible for publishing your standardized tests states that the average score across the nation on the mathematics assessment is 240. Your principal has asked you to prepare a report to the Horizon Board of Education and wants you to specifically address the progress of your high school students in mathematics. Can you use a one-sample *t* test to help you with your report?

Remember, a one-sample *t* test is a procedure used to determine if the mean of a distribution (your students) differs significantly from a preset value. In other words, does the mean of the Horizon High School math scores differ significantly from the preset value of 240, which represents the national average? As an appropriate test for significance, the one-sample *t* test compares a sample mean with a single fixed value.

Questions

1. Are the Horizon High School students' scores significantly different (higher or lower) from the national average as measured by the standardized math assessment you currently use?
2. Using the same national average score of 240, what can you say about your high school students in other subject areas?

Discussion

Figure 13.1 displays information helpful in addressing the Horizon School Board regarding the math achievement of your

high school students. This table is a result of an SPSS one-sample *t* test. Notice that you first get a printout of the descriptive statistics (*N*, mean, and standard deviation). The second printout has the important information we need. Good news for you, your students, and the school board. This one-sample *t* test indicates that the mean score for your high school students in math is 245.55, and is significantly higher than the national average of 240. Remember, on first glance, you might say, "well certainly they are higher." But the real issue and question is, Is the difference large enough to be considered statistically significantly different. This is really based on the sample size (100, *df* 99) and the alpha level we choose as criteria ($p = .05$).

Figure 13.1 One-Sample *t* Test for Math Scores

One-Sample Statistics

	N	*Mean*	*Std. Deviation*	*Std. Error Mean*
math	100	245.55	8.674	.867

One-Sample Test

	Test Value = 240					
					95% Confidence Interval of the Difference	
	t	*df*	*Sig. (2-tailed)*	*Mean Difference*	*Lower*	*Upper*
math	6.398	99	.000	5.550	3.83	7.27

With an alpha level of .05 and 99 degrees of freedom ($N-1$), the *t* value required for significance is approximately 1.9, as shown in Resource B. You notice from your SPSS test report that the *t* value is actually 6.4 (rounded), with a significance much lower than our targeted .05. Though SPSS reports .000, we recall that this figure is rounded to three decimals, so we can report $p < .001$. This level is well into the critical region required to reject the null hypothesis of no significant difference between your high school students and the national average.

Figure 13.2 SPSS One-Sample *t* Test for Writing Scores

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
writing	100	205.85	11.892	1.189

One-Sample Test

	Test Value = 240					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
writing	–28.717	99	.000	–34.150	–36.51	–31.79

Using a one-sample *t* test to compare the national average of 240 with the other subject areas reveals that both science and language means are also significantly higher than the national average. However, you will notice that the one-sample *t* test for writing reveals that the mean score is significantly *lower* than the national average, as shown in Figure 13.2.

This draws attention to why we want to conduct a two-tailed test. We are interested in *any difference*—higher or lower. In this case, the significance is equally important to know. The board of education will no doubt feel pretty good about what is happening in math, language, and science. But, the evidence reveals a problem with writing—and you have some evidence-based decisions to make before discussing this finding with them. Maybe you have a high population of limited-English speakers; maybe the difference relates to the kind of performance students are asked to complete on their writing exam; or maybe it is gender specific (females with high scores and males with very low scores), skewing the mean downward? Do some more investigating to see if you can find further *evidence* explaining this particularly troublesome finding.

Activity 4. An Independent-Samples *t* Test

You have taught high school algebra, geometry, and trigonometry for more than 10 years at Horizon High School and currently

serve as department chair. Though your superintendent has tried to recruit a woman or two for the math department, the department continues to be dominated by male instructors. With the increased attention on Title IX, social justice, and gender equity, educators must be sensitive to and aware of the issues of gender equity in both academic and extracurricular activities.

A local parent activist group has approached your school board and made the claim that because of the dominance of male math instructors, males are receiving better instruction and more attention in math courses than are the female students. Specifically, these parents claim that boys are outperforming girls in math achievement.

Your superintendent has asked you to report your assessment of the situation at the Friday afternoon faculty meeting. Use an independent-samples *t* test with a significance level of .05 to help prepare for the Friday meeting.

Questions

1. Do the males and females differ significantly from each other on standardized math test performance?
2. What unusual finding presents itself, and how might you explain this finding to the faculty and superintendent on Friday?
3. How might you use evidence-based decision making to address the general issue of gender equity at Horizon High School?

Discussion

As displayed in Figure 13.3, we find there is a 3.41 mean difference between the means (males = 243.94; females = 247.36).

Hmmm . . . Actually, we discover that the girls are outperforming the boys in standardized math scores. Careful, let's check the power of this finding. Again, the real question is, Is the difference large enough to be statistically significant? Yes, the *t* value is –1.996, with a significance level of .049. By-the-book significant, but look how close we are to .05. This finding is too close to call significant. But, good news—there does not seem to be enough *evidence* to conclude that there is any difference academically in math between girls and boys. So, I would feel good about the fact that math achievement is pretty balanced in terms of gender performance.

Figure 13.3 Independent-Samples *t* Test: Gender and Math Scores

Group Statistics

	male=1,female=2	N	Mean	Std. Deviation	Std. Error Mean
math	male	53	243.94	8.798	1.208
	female	47	247.36	8.253	1.204

Independent-Samples Test

		t	df	Sig. (2-tailed)	Mean Difference
math	Equal variances assumed	–1.996	98	.049	–3.418

What kinds of further investigations or decisions would you want to consider with your faculty? And even with the gender finding, is this *proof* that gender and math achievement are not problem areas? And what about this perception that your parent group had about male math instructors? Are there perhaps other variables or factors at play here? Evidence-based decision making!

Activity 5. A GPA Analysis

As a guidance counselor at Horizon High School, you spend considerable time advising seniors about attending college. Your local university just announced that it is raising the grade point average (GPA) entrance requirement from 2.8 to 3.0.

Questions

1. Can you produce from the Horizon High School data file an analysis of your seniors' GPA in relation to the newly announced GPA requirement by the university?
2. If you analyzed the GPA scores across all grade levels, what concerns might you have?

Discussion

I threw an easy one at you for a change of pace. This issue can be addressed with a single cross-tabulation, as shown in Table 13.6.

Table 13.6 Horizon High School Grade Level by GPA

	GPA								
Grade	*2.5*	*2.8*	*2.9*	*3.0*	*3.1*	*3.2*	*3.3*	*3.4*	*Total*
9	1	8	3	7	9		5		33
10	3	4	6	3	8		2		26
11	7	10		3					20
12			4	7	5	2	1	2	21
Total	11	22	13	20	22	2	8	2	100

Things look pretty good with our seniors' GPAs. All but four seniors have GPAs of 3.0 or higher, and those four have GPAs that are very close (2.9). Perhaps we might want to provide some intervention to the four who are "on the bubble." Or perhaps they are students planning on attending the local community college for 2 years before transferring to the university, in which case the issue is not so serious.

Wow! The number of students in Grades 9, 10, and 11 with noticeably low GPAs is somewhat alarming. Slightly more than 30% of the ninth graders, 50% of tenth graders, and 85% of eleventh graders have GPAs below 3.0. Of special concern is the large group of eleventh graders, for they only have 1 year to improve significantly. Again, lots of opportunities for *evidence-based decision making.*

Activity 6. One-Way Analysis of Variance

Having had a close call with the local activist parent group over gender equity issues, you, as the Horizon High School principal, decide that you want to look at any significant differences in math scores among the three ethnic groups at your school. In other words, you are interested to see whether any of your three ethnic groups (White, Hispanic, and Black) differ significantly from each other in the subject area of math.

A one-way analysis of variance (ANOVA) will help address this concern. The ANOVA will reveal any significant differences within any of the comparisons of the three ethnic groups and their respective math scores. In this case, the dependent variable is the math test and the independent variables are the three ethnic groups. Recall that the ANOVA indicates only *whether* there is any difference among the three groups—it does not identify the exact location of any difference. We will need to select a post hoc test (Tukey HSD) to help answer that question.

The requirement for a *t* test is that we can compare *only* two means—it is the ANOVA that allows us to compare many means at one time. The means for the math scores for each of the ethnic groups will be compared with each other: Whites with Hispanics, Whites with Blacks, Hispanics with Blacks. The one-way ANOVA will generate a significance value indicating whether there are significant differences within the comparisons being made. The significance value does not indicate where the difference is or what the differences are, but the post hoc test will help identify pairwise differences.

Questions

1. After running a one-way ANOVA on math scores, run another for each of the other subject areas.
2. With the help of the Tukey HSD test, where does the problem seem to lie?

Discussion

I will not display all of the tables from this analysis. Figure 13.4 displays some important descriptive statistics, an analysis of ethnic groups and math, and the Tukey HSD test.

First of all, we realize that the group sizes are unequal in this analysis. In other words, it would be better if all three ethnic groups consisted of similar numbers of students. Admitting a weakness of our study, let's look at the results. The ANOVA reveals a significant difference (or differences) within comparisons of math scores and the three ethnic groups, based on $F(2, 97) = 3.743$, $p = .027$. You notice that the significance level is .027, much lower than the selected criteria of .05, making the finding somewhat more powerful.

Now let's see if we can specifically identify *where* the significant difference is. Looking at the Tukey HSD test, and focusing on the "Multiple Comparisons" printout in Figure 13.4, we see significant difference *only* in two places: comparing Whites with Blacks and

Figure 13.4 One-Way ANOVA and Tukey HSD for Ethnicity and Math

ANOVA

math

	Sum of Squares	*df*	*Mean Square*	F	*Sig.*
Between Groups	533.732	2	266.866	3.743	.027
Within Groups	6915.018	97	71.289		
Total	7448.750	99			

Multiple Comparisons

Dependent Variable: math
Tukey HSD

(I) white=1, black=2, hispanic=3	*(J) white=1, black=2, hispanic=3*	*Mean Difference (I–J)*	*Std. Error*	*Sig.*	*95% Confidence Interval*	
					Lower Bound	*Upper Bound*
White	Hispanic	–3.696	1.954	.147	–8.35	.96
	Black	–5.786(*)	2.393	.046	–11.48	–.09
Hispanic	White	3.696	1.954	.147	–.96	8.35
	Black	–2.089	2.646	.710	–8.39	4.21
Black	White	5.786(*)	2.393	.046	.09	11.48
	Hispanic	2.089	2.646	.710	–4.21	8.39

* The mean difference is significant at the .05 level.

Homogeneous Subsets

Tukey HSD

white=1, black=2, hispanic=3	N	*Subset for alpha = .05*	
		1	*2*
White	56	243.59	
Hispanic	28	247.29	247.29
Black	16		249.38
Sig.		.262	.648

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 25.846.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

comparing Blacks with Whites (really the same comparison, but in reverse, causing one value to be positive and the other to be negative). Notice the significance value next to the comparison: Whites