2. **Inadequate sample size:** There are not enough learners in each group. Studies with fewer than twenty-five learners in each group may be suspect, especially if the instructional effect is not very strong.

3. **Insensitive measure:** The dependent measure was not sensitive enough to detect differences in learning outcomes. If the test does not contain enough items or the items do not adequately test what was taught, then we really can’t tell whether the instruction was effective.

4. **Inadequate treatment implementation:** The treatment and control groups were not different enough from each other. For example, if the background music is played at a very low level, it might not be loud enough to be heard.

5. **Insensitivity to learners:** The learners were not sensitive enough to the treatment. For example, if the material was very easy for all learners, then adding the treatment feature is not really necessary.

6. **Confounding variables:** The treatment and control groups differ on another important variable. For example, the control group may have many more experienced learners than the treatment group.

Overall, in your search for good research, you need to be sure you can rule out all other explanations for saying the instructional method did not work.

### How to Interpret Research Statistics

All of these issues relate to the applicability of the research to your learning situation, that is, to the confidence you can put in the results based on the validity of the study. Throughout this book, we report the results of statistical tests of the research we summarize. Therefore, in this section we briefly summarize how to interpret those statistical tests.

Suppose you read a study comparing two groups of students—a test group and a control group. The control group received a basic multimedia lesson that explains content with graphics and audio narration. We call this the no-music group. The test group received the same lesson with background music added to the narration. We call this the music group. Suppose
the no-music group averaged 90 percent correct on a test of the material and the music group averaged 80 percent on the same test. Averages are also called means (for example, 90 percent versus 80 percent). Also suppose the scores were not very spread out, so most of the no-music students scored close to ninety and most of the music students scored close to eighty. Standard deviation tells you how spread out the scores are, or how much variation there is in the results. Powerful instructional methods should yield high averages and low standard deviations. In other words, high scores are achieved and nearly all learners score close to the average so that there is high consistency in outcomes among the learners.

As illustrated in Figure 3.5, let’s suppose the standard deviation is 10 for the no-music group and 10 for the music group. Based on these means and standard deviations, can we conclude that background music hurts learning? Generally, when the difference between the score averages is high (90 percent versus 80 percent in our example) and the standard deviations are low (10 percent in our example), the difference is real. However, to accurately decide that issue requires statistical tests. Two common statistical measures associated with research studies we present in this book are probability and effect

---

**Figure 3.5. Computing Effect Size for the Differences Between Mean Test Scores on Two Lessons.**

![Diagram showing the comparison of test scores between lessons with and without background music, illustrating the calculation of effect size.](image-url)
size. As you read research, look for results in which the probability is less than .05 (p < .05) and show an effect size of .5 or greater.

**Statistical Significance: Probability Less Than .05**

Some statistical tests yield a measure of probability such as p < .05 (which is read, “probability less than point oh five”). In the case of our background music study, this means that there is less than a 5 percent chance that the difference between 90 percent and 80 percent does NOT reflect a real difference between the two groups. In other words, if you concluded there is a difference in test performance between the groups, there is less than a 5 percent chance that you are wrong and more than a 95 percent chance that you are right. Thus we can conclude that the difference between the groups is statistically significant. In general, when the probability is less than .05, researchers conclude that the difference is real, that is, statistically significant.

**Practical Significance: Effect Size Greater Than .5**

Even if music has a statistically significant effect, we might want to know how strong the effect is in practical terms. We could just subtract one mean score from the other, yielding a difference of 10 in our music study. However, to tell whether 10 is a big difference, we can divide this number by the standard deviation of the control group (or of both groups pooled together). This tells us how many more standard deviations one group is compared with the other, and is called effect size (ES). In this case, the ES is 1, which is generally regarded as a strong effect. What this means is that an individual learner in the control group would see a 1 standard deviation increase (10 points in our example) if he or she were to study with a lesson that omitted music. If the ES had been .5 in our example, an individual learner in the control group would have a .5 standard deviation increase (5 points in our example). When the ES is less than .2, the practical impact of the experimental treatment is a bit too small to worry about; an effect size of .5 is moderate, and when it is .8 or above, you have a large effect (Cohen, 1988). In this book, we are especially interested in effect sizes greater than .5, that is, instructional methods that have been shown to boost learning scores by more than a half of a standard deviation.