twelve

Designing Ex Post Facto and Experimental Studies

Debbie Rohwer

Fuente: Debbie Rowhwer (2013) En Hildegard C. Froehlich,Carol Frierson—Campbell (eds) Inquiry in Music Education: Concepts and Methods for the Beginning Researcher. London: Routledge.

This chapter highlights the following concepts:

- Types of difference studies;
- Internal and external validity:
    - design options;
- Analysis of Variance:
    - the F statistic;
    - statistical issues with the ANOVA;
- Inputting data into a statistical program;
- Difference studies in the classroom.

Ex post facto (sometimes called causal-comparative) and experimental research investigates differences between two or more groups. This chapter describes and distinguishes between these two types of research design.

Introduction

> As a high school music teacher, *Juan* noted a trend in his band classes over the last two years. Of the current students, male percussionists seemed to play at a higher level than female percussionists. He wondered if this was just an issue with this group of students or whether the issue was more widespread. To answer this question, he arranged for all high school percussionists in his district to take a performance test. He found that the difference between males and females was statistically significant, favoring males (M = 87.6, SD = .65) over females (M = 73.1, SD = .80). He pondered the clarity of this result. Maybe it wasn't gender in isolation, but something else that caused the result. Perhaps the male percussionists began instruction earlier or took more private lessons. Or some other variable was causing this finding altogether. It was challenging to know for sure.

Juan's dilemma—trying to understand differences between groups—is familiar to music education researchers. Research that investigates differences uses *ex post facto* (sometimes called causal-comparative), and/or *experimental* designs. The distinction between ex post facto and experimental research is determined by the *independent variable,* which is a nominal level grouping variable.

The independent variable in Juan's study is gender (with the distinct levels of male and female). In this case, the independent variable (gender) is pre-existing; the design of the study is therefore called *ex post facto* (Latin for "after the fact"). Studies in which the researcher is in control of the independent variable are called *experimental* (as, for example, in an experiment where you can determine who gets what treatment). The dependent variable in both ex post facto and experimental studies is a test. In Juan's case, the dependent variable was an interval level test of performance ability. The following section details considerations for both types of studies.

Types of Difference Studies

In ex post facto studies the independent variable is pre-existing (as in Juan's study). Such variables can also be called self-selected or non-manipulated. Examples relevant to music teachers may include gender (male, female), instrument family (woodwinds, brass, percussion), voice part (soprano, alto, tenor, bass), aptitude groupings (high, medium, low scorers), or grade level (middle school, high school).

As an example, take the following article titled "The effect of grade level on motivation scores." Here, the independent variable (the nominal level group variable) is grade level (middle school and high school), and the dependent variable (the interval level test) is motivation scores. Because the nominal-level independent variable is pre-existing—that is, the participants are already in middle school or high school—the study is ex post facto.

In an experimental design, the independent variable can be manipulated (or controlled) by the researcher. Examples might include treatment groups (block chord accompaniment versus arpeggiated accompaniment) or conditions (sight singing a song with Curwen hand signs, and also without hand signs). A researcher can also design a study that combines the two: one independent variable being manipulated (in the treatment groups) and one independent variable being pre-existing (e.g., gender).

As an elementary music teacher, *Keisha* was interested in measuring the difference between those students who receive large movement training and small movement training. She had learned that large movements may help flow and feel related to rhythm, while small movements may be more transferrable to the Orff instrument techniques she used. She wondered if small movements would also help rhythmic feel. She designed an experiment where she tested the rhythmic steadiness of all four of her classes of third graders. Then she took two of the classes and taught them rhythms using small motor movements specific to Orff mallet technique. The other two classes learned the same rhythms but used large motor movements. She then measured all students at the end and found no significant differences in steadiness across the two treatment groups (large motor: M = 36.1, SD = .20; small motor: M = 35.8, SD = .39).

In contrast to Juan, Keisha designed a study with two different treatment groups (large motor training in one group and small motor training in the other group). Her study was experimental because as the researcher, she was in control of the independent variable.

Notice that in both of these examples, the independent variable is a *grouping* or *categorical variable* (at the nominal level of measurement, as discussed in Chapter 10), which places subjects into families or subgroups of people. The independent variable is able to differentiate better between scores on the test (the dependent variable) if the groupings maximize variance. This concept is the first component of Kerlinger's (1986) MAXMINCON principle, wherein studies ideally MAXimize variance, MINimize error, and CONtrol extraneous variables.

Maximizing variance (MAX) can be done by choosing independent variable groupings that allow for differences to be apparent. For instance, if you were trying to give clear examples to a young child of what the words "short" and "tall" meant, you probably would not pick as the examples one person who was 5 ft. 3 inches and another who measured 5 ft. 4 inches. You would most likely choose people who could maximize the difference, thereby clarifying the terms: Someone 4 ft. tall might represent "short" and someone 7 ft. tall would represent "tall." In the case of an experiment, then, you would not choose test score groupings (high and low groups, for instance) as an independent variable and then choose groupings that vary only by one point on the test (low: scoring 99, and high: scoring 100). Instead, it would be important to have a spread of representative scores that would maximize the groupings, such as lows having scores between 1 and 50, and highs having scores between 51 and 100.

Minimizing error variance (MIN) can be done by doing a validity check to make sure: (1) the testing instrument is appropriate for the age level of the test takers; (2) the items represent the concept being measured; and (3) the test does not have superfluous content or items that are missing but are necessary to represent the content appropriately. After a validity check and a field test to check for clarity, a pilot test can be done to assess the reliability of the instrument (see Chapter 10), with the goal of this three-step process being to lessen the impact of error variance.

The CON (CONtrol extraneous variables) in Kerlinger's principle can be accomplished by designing and implementing a study that allows confidence that factors other than the independent variable were not the cause of the results. In order to control for extraneous variables as possible causes, researchers must analyze, redesign, and discuss any research-related issues that could question the claim of the independent variable having been the most likely cause of the results of a study. Integral to such confidence is investigating the internal validity of research designs and procedures.

Internal and External Validity

Campbell and Stanley (1963) published a seminal list of what they called *threats to internal validity.* Such threats, detailed further below, are factors that could weaken a study by causing readers to question whether the independent variable or some other extraneous variable caused the results to happen the way they did. Since 1963, there have been further expansions of the initial list of threats to internal validity (Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2001), but the issue is the same: Studies should be designed and conducted in such a way that the likelihood of any extraneous variables being the cause of the findings can be ruled out to the greatest extent possible.

Design Options

Threats to internal validity are partly based on the choices you make when planning a study as well as the choices made in the midst of carrying the study out. For instance, when designing an experimental study, you may choose: (1) whether to have a comparison group or not; (2) whether to use a pretest or not; and (3) whether to use random assignment or not. *Pre-experimental* (or less controlled) designs use fewer control options than true experimental designs.

A common middle-ground design, called a *quasi-experimental design*, uses a realistic combination of the three issues mentioned; it has a treatment and comparison group, each of which was intact (already formed) before the study began. Individuals in each group are given a pretest at the beginning of the study to see if the groups score equally on the test. Then each group receives "a level of the independent variable," technical language for saying that, for instance, one class receives guitar instruction on accompanying in block chords whereas the second class is instructed in arpeggiated chord accompaniment. After some period of training, members of both groups take a posttest. This design is called the *non-equivalent control group design* due to the use of non-randomly assigned, intact groups. Keisha's example earlier in this chapter was such a non-equivalent control group design because she used intact classes that were pretested, then "treated," and then posttested. (For a detailed discussion of design issues, see Frankel and Wallen's (2009) *How to Design and Evaluate Research in Education.*)

Threats to Internal Validity

Designs that do not have a pretest or that do not use randomization have a potential *selection* threat. This means that the groups could be different from the start without the researcher knowing it. Consider the challenge faced in the following scenario: A researcher has two groups, one receiving "hand/sign and solfege" training and the other receiving "solfege alone" training. At the end of the treatment, all students take a sight singing performance exam. The "hand/sign and solfege" group had a mean score of 47.3 on the posttest and the "solfege alone" group had a mean score of 71.4 on the posttest. No pretest was given before the treatment. The results are shown in Table 12.1.

The posttest information alone may lead you to conclude that the training caused the "solfege alone" group to score higher than the "hand/sign and solfege" group. We don't know, however, whether the groups were equal at the beginning. It could have been that the "solfege alone" group actually might have *decreased* from an initial pre-treatment mean of 88.7 and the "hand/sign and solfege" group actually might have *increased* from an initial pre-treatment mean of 21.8.

Now think about the finding reported in Table 12.2.

Based on this example, you can see that a pretest is pivotal to understanding whether groups differ at the start of a study. If they do differ, you would then need to undertake further statistical solutions to analyzing the data, or change the design of the study. A statistical solution would be to use an advanced statistic to control for initial differences between groups. A design-based solution would be to take the initial groups and reorganize them by placing them into equal groups based on the pretest scores. Either way, interpretation of results without pretest information may be incorrect. Therefore, researchers need to weigh the pretest variable carefully in order to avoid a selection threat.

The selection threat is only one of a number of possible challenges that researchers face when trying to determine cause and effect. Other examples are:

TABLE 12.1 Scores on a Posttest-Only Design

Pretest Score	Treatment	Posttest Score
?	hand/sign and solfege	47.3
?	solfege alone	71.4

TABLE 12.2 Pre- and Posttest Scores on Sight Singing With and Without Solfege

Pretest Score	Treatment	Posttest Score
21.8	hand/sign and solfege	47.3
88.7	solfege alone	71.4

- *Implementation threat:* The researcher who gives positive feedback to one treatment group and less positive feedback to the other group, when feedback is not the measured independent variable (i.e., the treatments are not being implemented consistently across the groups).
- *Instrumentation threat*: The individuals in the performance judging panel who tire as the day goes on and become more lenient in their scoring (i.e., the scoring is not being conducted consistently across the participants).
- *Location threat*: The one treatment group is in a well-lit room and the other that is in the broom closet (i.e., the environment/setting is not consistent across groups).
- *History threat*: The students in one of the treatment groups get together outside of the treatment sessions to do extra study/work, while the other treatment group participants do not meet (i.e., students' study/work activities outside of the treatment are not consistent across groups).

Especially in ex post facto studies, almost anything in addition to or in place of the independent variable may have caused the results of the study.

In *Juan*'s study of male and female percussionists, the independent variable was out of his hands, and anything may have worked in conjunction with the independent variable to confound the clarity of the findings. As he wondered aloud, "Maybe the male percussionists are beginning instruction earlier or taking more private lessons, or some other variable is causing this finding altogether." Juan clearly had good reason to ponder additional causes for the result he found.

Threats to External Validity

While possible threats to internal validity make you consider the clarity of the independent variable as a cause for the study's results, *threats to external validity* lead to questions about whether a given study's design or methodological choices may be generalized to other settings. Whether an expert provides instruction for the treatment groups in a study may not be problematic for determining cause and effect (i.e., the internal validity), but it could make the results ungeneralizable to settings where the instructor is not an expert on the treatments. Or, a study that compares two treatments with a group of students who have participated in many different experiments may not call into question the internal validity of the study, but due to the students' experimental savvy, the results may not be generalizable to other, less savvy student samples. In fact, simply having students know that they are part of an investigation may make its results less generalizable to settings where the treatment is being implemented in an ordinary, non-research classroom setting. Both internal and external validity issues need to be weighed so that the best choice for determination of cause, issues of generalizability, and considerations of feasibility can be considered.

Analysis of Variance

The most common statistical method for analyzing data in difference studies (i.e., ex post facto studies or experimental studies) is the *analysis of variance* or *ANOVA*. In this chapter, the principles of the ANOVA will be explained as the main statistic by which you compare means of subgroups to determine the extent of difference between them.

In an ANOVA, the number of independent variables determines the naming of the ANOVA: A study with one independent variable is called a one-way ANOVA; a study with two independent variables, such as "treatment" and "gender," is called a two-way ANOVA.

Both Juan's and Keisha's research efforts (described earlier in this chapter) were one-way ANOVAs with two levels or subgroups to the independent variable. Gender was the independent variable in Juan's case, with male and female as the levels. In Keisha's study, treatment was the independent variable, with large motor and small motor treatment groups as the levels.

A two-way ANOVA with two levels to each independent variables (such as having both of the variables "treatment and gender" in the same study) can also be called a 2 × 2 ANOVA ("2 by 2"). This term describes the number of levels or subgroups in each independent variable. In the case of a 2 × 2 ANOVA, then, the variable "treatment groups" has two levels (such as moveable do training and fixed do training) and the variable "gender" also has two levels (male and female). This terminology changes according to the number of levels of each independent variable. For instance, you would call the statistic a 3 × 2 ANOVA for the following scenario: An ANOVA had two independent variables, but the first independent variable (treatment group) had three levels (such as Eastman counting system training, Gordon counting system training, and Kodaly counting system training) while the second independent variable had two levels (high and low). This design is illustrated in Figure 12.1.

The F Statistic

The result of an ANOVA is documented as an *F statistic.* The larger the F, the greater the calculated difference between groups. Also, the larger the F, the smaller the calculated significance value will be, documenting the smaller likelihood that differences were due to error or chance. Results of an ANOVA are listed together with a test of significance in which a p value (significance) below .05 indicates that the sets of means for the subgroups are significantly different from each other, with less than a 5% risk of the result being due to error or chance.

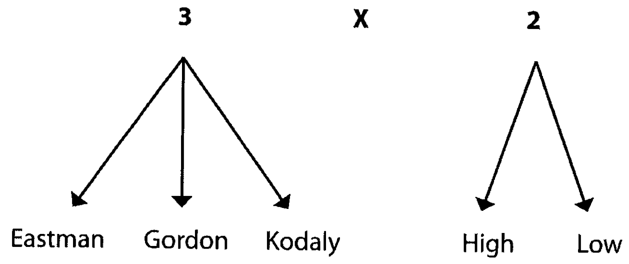

FIGURE 12.1 Illustration of a 3 × 2 ANOVA Design

For example, if a group of females and males was assessed for differences on a theory test, the ANOVA would be used to assess if the theory scores for the males (as seen in the first graph of Figure 12.2) differed from the theory scores for the females (as seen in the second graph of Figure 12.2). The height of the bars on the graphs in Figure 12.2 displays the number of people with that score; for instance, there were 12 females who scored 80 on the second graph, which makes up the tallest bar on the graph. Just by looking at the middle of each graph (a mean of 40 for the males and a mean of 80 for the females) you can envision that the means are probably going to be different enough not to be attributable to error or chance.

When the ANOVA is calculated for this data set, it documents a significant difference for the variable gender, *F* (1, 62) = 220.44, *p*< .001, favoring females (*M* = 80.00, *SD* = 1.91) over males (*M* = 40.00, *SD* = 1.91). The p value being smaller than .001 in the result sentence tells us that we are at least 99.9% sure that the groups differ, and not by error or chance. The large F value aligns with the small p value. The numbers in the parenthesis after the F are part of the calculation of the F; you would see these numbers in the ANOVA table if you looked at the degree of freedom column.

Whether you are looking at an ANOVA table or reading the results embedded in an article's results section, such as *F* (1, 62) = 220.44, *p*< .001, it is important to know what the numbers mean so that you can be an educated reader. The numbers in the parentheses are called the *degrees of freedom* (abbreviated as *df*). They serve as part of the calculation of the F statistic, but also tell you important information about the choices the researcher made. In the case of the gender study listed above, the parenthetical content reads "(1, 62)."

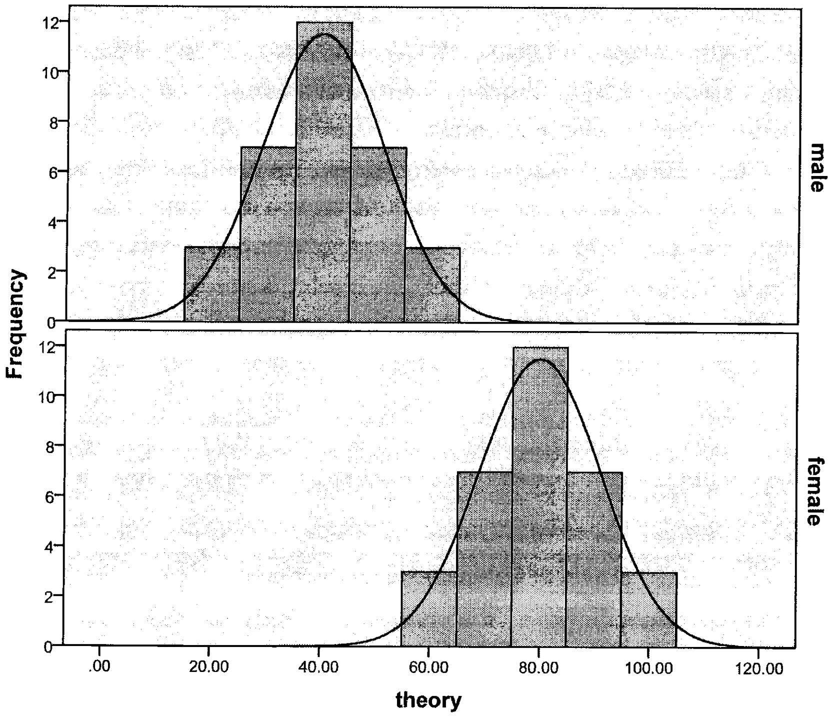

FIGURE 12.2 Music Theory Score Distributions for a Sample of Male and Female Students

The first number is the df for the independent variable, gender. If you add 1 to this number, it tells you the number of levels of the independent variable (note that this only works for independent variables). In this example, then, there are two levels to the variable gender (male and female). The second number in the parenthesis is the degree of freedom for error, and if you add this number (62) to the degree of freedom for the independent variable (1+62) and add 1 more, you will get the sample size for this one-way ANOVA. This study, therefore, had 64 people in it and the result that would be stated in the article would be that there was a significant main effect for gender, favoring females. You know that females outperformed males by looking at the original means for both subgroups. The term "main effect" is commonly used when discussing the results of independent variable findings.

Both Juan and Keisha could provide degree of freedom information for the results of their studies. Such information would remind the reader of the number of levels of the independent variable in each study as well as the sample size. In the case of Juan, comparing male and female percussionists on a performance test resulted in the following: $F(1, 92) = 27.3, p = .02$. He found a statistically significant difference between male and female percussionists, favoring males.

To determine sample size as well as number of subgroups (or levels) in this study, look at the first number in the parenthesis, the degree of freedom for the independent variable (1). The 1 (plus 1) states that there are two levels to the variable gender (males and females). Add the degree of freedom for gender number (1) to 92 and add 1 additional and you get the sample size of 94 students.

In Keisha's experiment with two treatment groups, no significant differences across the two groups were found: $F(1, 78) = .03, p = .92$. From the degrees of freedom, you can determine that she had two treatment levels (you add 1 to the 1 in the parenthesis). Her sample size can be obtained by adding together the two numbers in the parenthesis and then adding an additional 1, which equals 80.

The ANOVA table listed in many research articles provides the reader with the same information you can glean from the parenthetical way of reporting statistical findings. The table for a one-way ANOVA with the ex post facto variable "voice part" being the independent variable, for example, may look like Table 12.3.

The finding shown in Table 12.3 documents that for the variable voice part, there is a significant main effect or difference ($p = .003$) that is below the normal significance cutoff of .05. In the example shown in Figure 12.2, there were only two levels to the independent variable "gender" (male and female), and just by looking at the original subgroup means (80 and 40) you could say that females outscored males. In this example, however, there are four levels to the independent variable "voice part." You can discern this information by looking at the degree of freedom for the variable "voice part" (3) and then adding 1, which equals 4. You can also tell from this table that the degree of freedom for error is 76 and if you add 76 to 3 and add 1 more you will get the sample size for this study, which is 80. (For a discussion of how degree of freedom information works in all forms of ANOVAs and other statistics, see Huck's (2012) *Reading Statistics and Research.*)

TABLE 12.3 Summary Table of a One-Way ANOVA

Tests of Between-Subjects Effects

Dependent Variable: score

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Voice part	1403.750	3	467.917	4.984	.003
Error	7135	76	93.882		

Because the example deals with four levels to the variable "voice part" (soprano, alto, tenor, and bass), the basic ANOVA cannot describe the specifics of "where" the difference(s) lie; it only documents that there is a difference. After obtaining a significant ANOVA finding for a study that uses a variable with more than two levels, you could calculate a post hoc (Latin for "after this") test to determine information on the specific subgroup differences. Another option would be to calculate planned comparisons based on suggestions in the literature. Conducting further statistical tests like post hocs can sometimes be like data mining. The debate of data mining versus specified comparisons is a topic that needs to be purposefully and philosophically weighed.

Returning to the data set represented in Table 12.3: Descriptively, the sopranos had the lowest overall mean (49.00), with the altos (51.00), tenors (58.00), and basses (58.50) having higher mean scores. If a post hoc statistic were calculated, you could document that the lowest mean from the sopranos (49.00) was not significantly different from that of the altos (51.00), but the sopranos (49.00) were significantly different from both the tenors and the basses (58.00 and 58.50). The alto, tenor, and bass means did not significantly differ from each other (51.00, 58.00, and 58.50).

If a study has more than one independent variable, then it will document results for each independent variable, and will also document an *interaction* result, which highlights whether the findings need to be qualified or not. If there is a significant interaction between the independent variables, then the results are in some way hazy and need to be clarified. For instance, a researcher might find a significant main effect for gender, favoring females, but females in one of the treatment groups actually scored lower than the males in both of the treatment groups. It would then be misleading and incomplete to say that there was a significant main effect for gender, favoring females. The significant *interaction* helps clear up that misinterpretation by qualifying the main effect result. An interaction qualification statement may read:

> While there is a significant main effect for gender favoring females, the main effect was caused by females in treatment group one who outscored all other subgroups. Females in treatment group two, however, scored lower than both male treatment subgroups, so it may be wise to consider carefully the interaction of gender and treatment groups when making instructional decisions.

Table 12.4 displays findings from a study with two independent variables (treatment and age) and one interaction result. The table has degree of freedom (df) information, the F statistic, and the p values (Sig.). The other table information (i.e., Sum of Squares and Mean Square) is used as part of the calculation of the F statistic.

TABLE 12.4 Summary Table of a Two-Way ANOVA

Tests of Between-Subjects Effects

Dependent Variable: score

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Treatment	4.000	1	4.000	.038	.846
Age	23104.000	1	23104.000	217.791	.000
Treatment*age	23104.000	1	23104.000	217.791	.000
Error	10184.000	96	106.083		

Once Table 12.4 has been analyzed to determine significance or non-significance, the original descriptive means of each subgroup could be looked up, a graph could be made, and then the following description could be written as results of the study:

1. There is no significant main effect for the variable treatment groups [F (1, 96) = .038, p = .85], with the visual group scoring: M= 50.0, SD = 11.43, and the auditory group scoring: M = 50.0, SD = 21.55.
2. There is a significant main effect for age [F (1, 96) = 217.79, p<.001], favoring younger students (M = 65.0, SD = 18.65) over older students (M = 35.0, SD = 17.34).
3. There is a significant interaction between treatment and age [F (1, 96) = 217.79, p<.001], showing the need to qualify the significant main effect for age. While younger students in the auditory group scored higher than all other subgroups (M = 80), younger students in the visual group achieved lower scores (M = 50), with those scores being equal to the scores of the older students in the visual group (M = 50). Also, older students in the auditory group scored the lowest of all subgroups (M = 20). This interaction can be seen in visual form in Figure 12.3.

Any time a significant interaction exists, you need to examine the subgroup means to determine where the non-parallel lines are occurring so that an explanation can be made of the qualifying information.

Statistical Issues with the ANOVA

The Analysis of Variance (ANOVA) is a *parametric* statistic (see Chapter 9) that has certain rules, or assumptions, governing its use. Most importantly, you need to be concerned with issues of *normality* (normal distribution of all scores) and *homogeneity of variance* (equal spread of scores). If these assumptions are not met, then non-parametric statistics (see Appendix K) or some other choice (such as transforming the data) must be used.

In order to calculate the results of a study by using an ANOVA, you first must establish that the distribution of the scores resembles the bell-shaped curve (see Figure 12.4). Distributions that are asymmetrically skewed (as shown in Figure 12.5) or too peaked (Figure 12.6), do not meet the assumption of normality and would not be appropriate for an ANOVA.

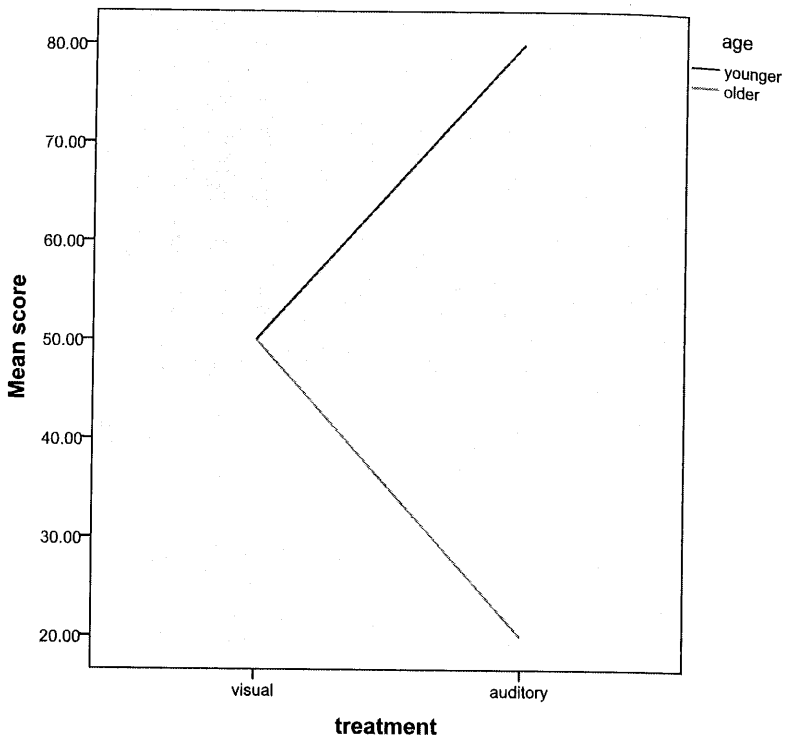

FIGURE 12.3 Graph of Interaction of Two Independent Variables

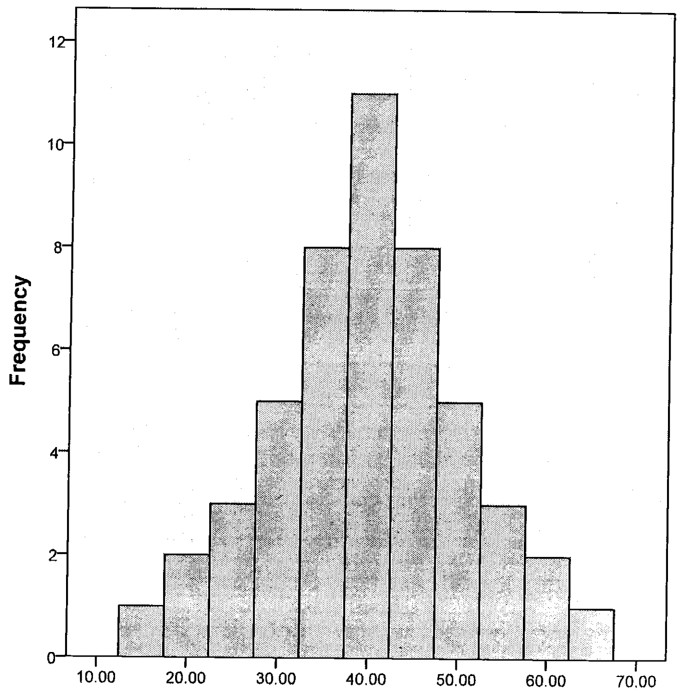

FIGURE 12.4 Normal Curve Histogram Based on Frequency of Scores in a Range from 0 to 70

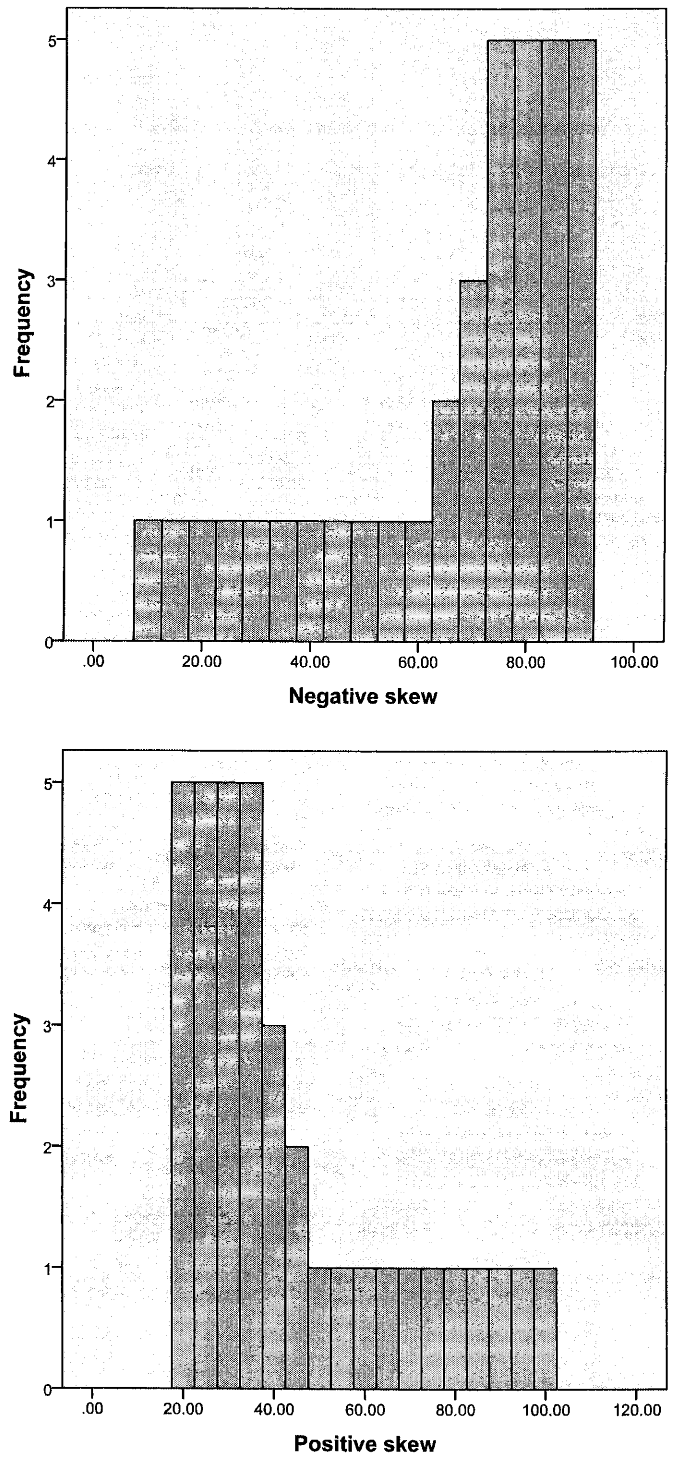

FIGURE 12.5 Asymmetrically Skewed Score Distributions

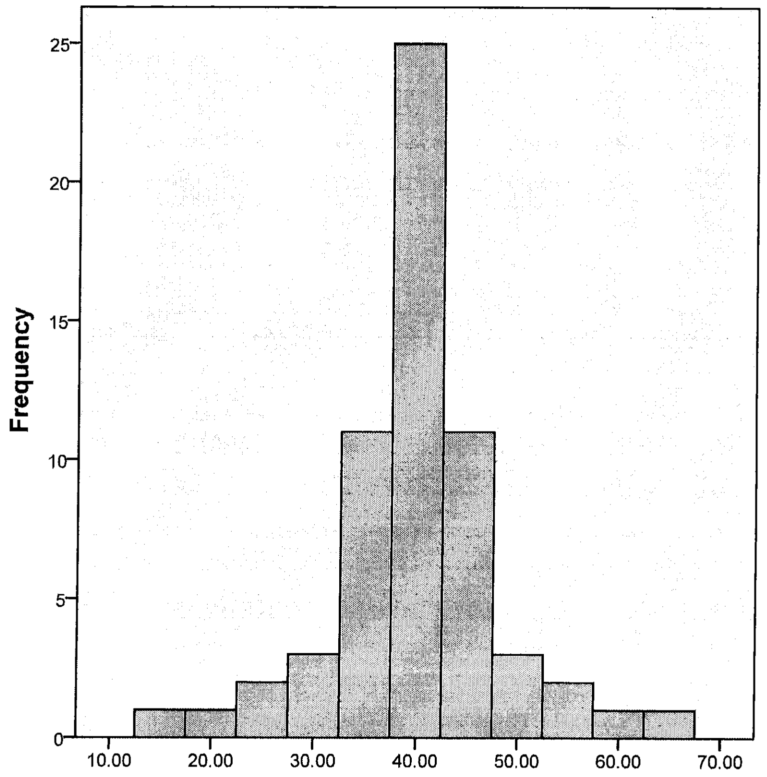

FIGURE 12.6 A Highly Peaked Score Distribution

It is also important that the spread of scores across each subgroup is similar, documenting "*homogeneity of variance.*" Figure 12.7 displays a pair of *box and whiskers* plots wherein the medians (represented by the horizontal line in the center of each box) are equal at 80. The equal picture sizes show an equality of variability.

The medians represented by the pictures shown in Figure 12.8 are similar, but the size of the box and whiskers shows a difference of spread (or variability). This discrepancy documents a problem with homogeneity of variance that would need to be solved in some way before an ANOVA could be calculated.

In general, it helps to have group sizes that are large in order to achieve a normal curve and to have enough *power* (see Cohen, 1988) so that the statistic can detect differences if they exist. A basic rule of thumb of 30 people in a group is an appropriate place to start for the purpose of meeting assumptions. Also important, however, is looking at past research to analyze sample size and significance results to determine how large your sample size might need to be to find differences if they exist.

As in other types of research, doing one or more pilot studies can help you estimate the power of a test to detect differences as well as the magnitude of the differences (called *effect size*). Since sample size can have an impact on statistical significance, adding effect size estimates can help the reader of an article gauge whether the stated statistical

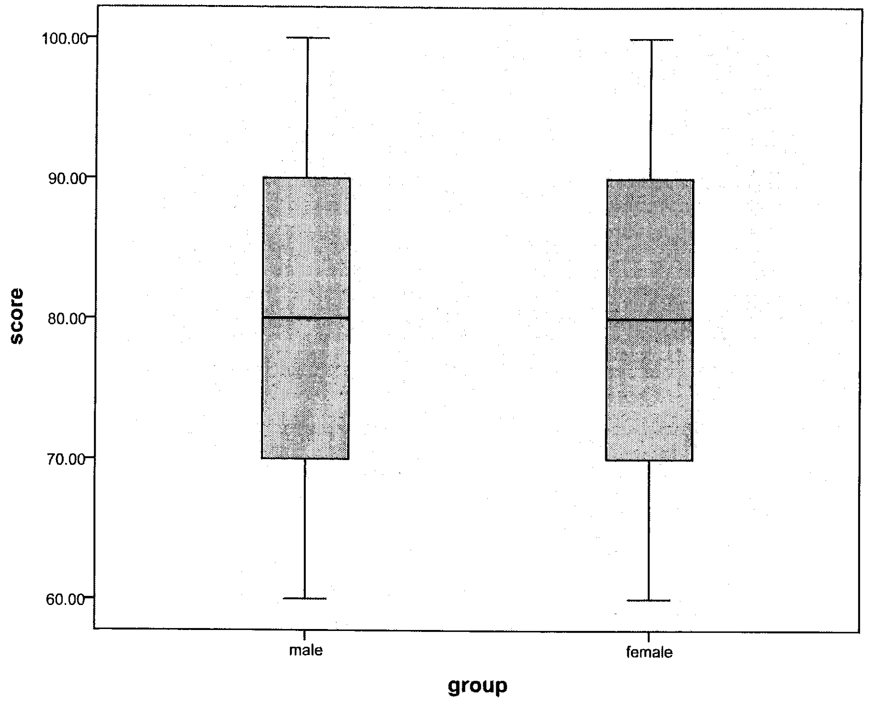

FIGURE 12.7 Box and Whisker Plots Suggesting Equality in Variability Between the Scores of Two Independent Variables

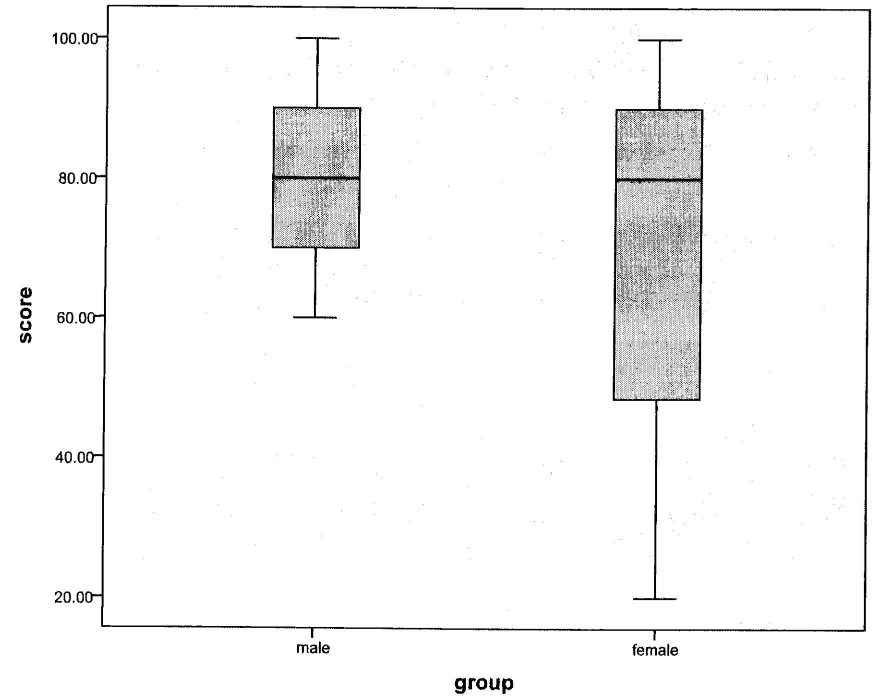

FIGURE 12.8 Box and Whisker Plots Suggesting Difference in Variability

significance was positively or negatively impacted by the sample size. This scenario can be seen when a result is statistically significant but has a small effect size (Type I error: can happen with a large sample size), or when a result is not statistically significant but has a large effect size (Type II error: can happen with a small sample size. For further discussion of Type I and II Error, see Chapter 10). Hence, it is important to plan a study's sample size carefully. Online sample size calculators are available to determine sample size estimates for a main study, using power and effect size data that can be gathered from a pilot study.

One other statistical issue to consider is a basic reminder that the .05 significance is used to calculate one statistic only. If you plan to run multiple ANOVAs, then you need to correct the .05 cutoff so that there is not an overuse of the data set. Statisticians sometimes describe studies that calculate multiple statistics with the same grouping variable but different dependent variables as "fishing" to find results, because there is a possibility that by just casting a line a great number of times (running statistics), you might catch a fish (find something significant). The fishing scenario might be seen in studies that use an independent variable such as gender to determine the effect on individual questions in a questionnaire (30 questions). In this case, there are 30 ANOVAs calculated. To counter this concern, the researcher can sum questions on a questionnaire and calculate one ANOVA instead of 30, or divide the .05 cutoff by the number of statistics to be calculated (in this case 30), thereby providing a more conservative cutoff that takes into account the multiple statistics being calculated (.05/30 = .002).

As a final note, there are many types of ANOVAs, such as an ANCOVA where you need to control for initial differences in group scores by weighting them on the posttest scores, or a repeated measures ANOVA where the same subjects are measured multiple times. Huck's (2012) text, *Reading Statistics and Research*, is a valuable resource that covers each of these advanced ANOVA options.

Inputting Data into a Statistical Program

The example below explains how to use SPSS (Statistical Package for the Social Sciences, available from www.spss.com) to examine data from an ex post facto study. It provides a small set of data with the independent variable being instrument family (woodwinds and brass) and the dependent variable being scores on a performance anxiety measure.

An ANOVA will be used to compare the scores of the woodwind and brass instrumentalists on a performance anxiety test. The inputted data for the example set of four woodwind and four brass instrumentalists shown in Table 12.5 would have two columns: one to demonstrate the independent variable groups and another to demonstrate the dependent variable scores. The independent variable column will have nominal level data: a number to represent those in the woodwind group (labeled 0 in this case) and a number to represent those in the brass group (labeled 1 in this case). Each row shows

TABLE 12.5 Example Data for One Independent and One Dependent Variable (N = 8)

	Group	Score
1.	0	91
2.	1	51
3.	0	83
4.	0	71
5.	1	70
6.	1	60
7.	1	61
8.	0	82

one person's group affiliation label and that same person's score on the performance anxiety test (so Sally the bassoonist may be first across, scoring 91 on the test, and Fred the trombonist may be second across, scoring 51, etc.).

The next step would be to check basic issues for calculating an ANOVA. These issues can be weighed by looking at histogram and box and whiskers graphs. To do this in SPSS, go to the "Analyze" drop-down and then choose "Descriptive Statistics" and "Explore." Place the variable "score" in the Dependent List and the variable "group" in the Factor list and then click on the "Plots" button. From there mark "Factor levels together" under "Boxplots" and" and "Histogram" under "Descriptive." Then click "Continue."

Viewing the box and whiskers graph, as shown in Figure 12.9, the spread, or *homogeneity of variance* across both groups appears consistent.

The histogram graph as shown in Figure 12.10 indicates that the distributions of the subgroup samples look like normal curves (display normality), especially for this example of eight data points. (Note that any real study would have a far larger sample size.)

You next calculate an ANOVA by going to the "Analyze" drop-down and choosing "General Linear Model" and "Univariate." Place the variable "score" in the Dependent Variable box and the variable "group" in the Fixed Factor(s) box. Then click "OK."

The readout will look like what is shown in Table 12.6, documenting that the difference between groups was statistically significant [F (1, 6) = 13.12, p = .01]. The original means can be obtained from the options button, showing that group 0 (woodwinds: mean of 81.75) had a higher overall mean score than 1 (brass: mean of 60.75).

Testing for differences among variables is at the core of many quantitative research designs. However, music teachers, too, can benefit from the procedures outlined in this chapter. For instance, when measuring differences in performance among students, a teacher might consider such ex post facto difference questions as "Do my students perform at different skill levels based on what instrument they have chosen to play, or their reading grades (high/low), or the amount of practice they reported last week (none/some/much)?" Music teachers can also weigh experimental difference questions of interest, such as "If I tried one method book format to introduce sight singing with

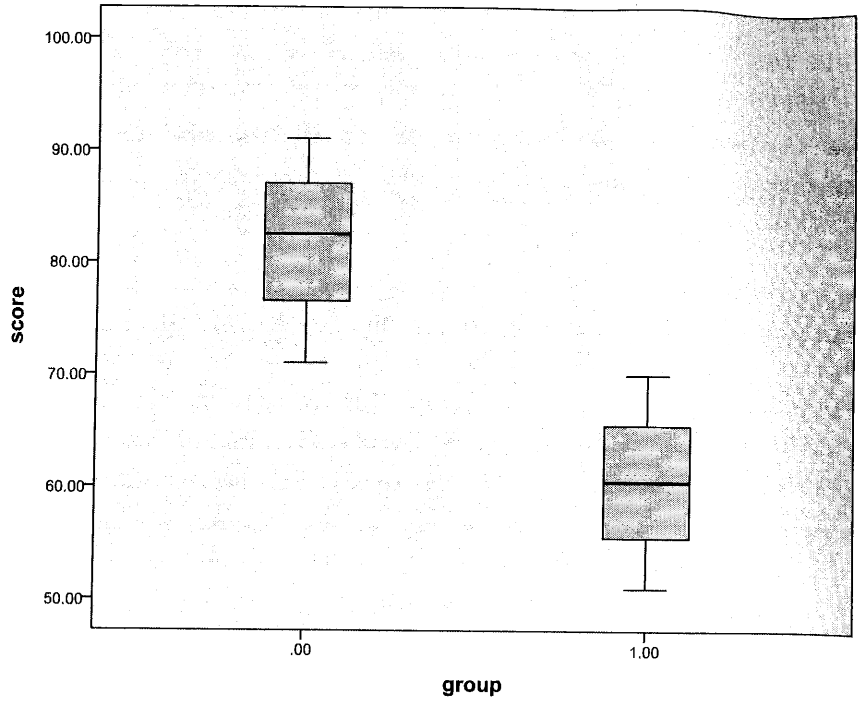

FIGURE 12.9 Box and Whisker Plots for Example Data (N = 8)

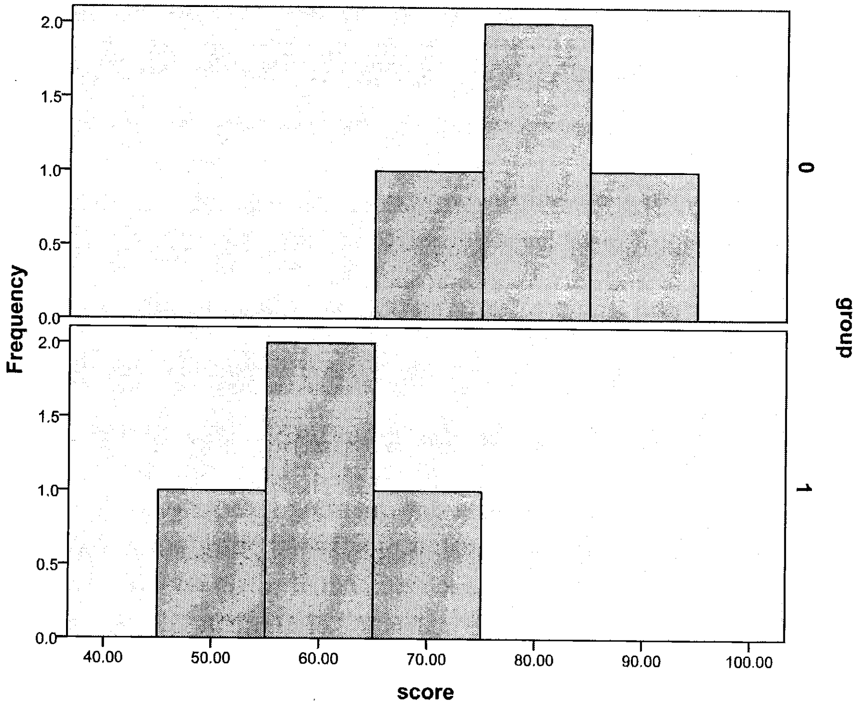

FIGURE 12.10 Histogram of Example Data (N = 8)

TABLE 12.6 Summary Table for Example ANOVA Results (N = 8)

Tests of Between-Subjects Effects

Dependent Variable: score

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Group	882.000	1	882.000	13.115	.011
Error	403.500	6	67.250		

one of my sections of students and another method book to introduce sight singing with the other section, I wonder which method book format would help the students perform better at the end?" The possible list of questions is indeed numerous but always contextual to each teacher's interests and areas of expertise. It is always important to weigh, carefully and logically, all instructional choices so that decisions are not arbitrary or are not clouded by external, unpredictable, and potentially meaningless factors.

Chapter Summary

1. Ex post facto and experimental designs are types of difference studies found in quantitative music education research.
2. Analysis of variance measures differences between groups. Researchers need to check whether the assumptions of the ANOVA are met before using this statistic.
3. It is important for researchers to consider how a study is designed and implemented in order to avoid threats to internal validity. External validity issues can impact the generalizability of a study's finding.

Topics for Further Exploration

1. Types and levels of variables of interest to music educators.
2. Research designs and corresponding statistics.
3. How power, effect size, and sample size work together.

Suggested Assignments

1. Compare males and females on a self-assessment of performance anxiety. Then the class can construct a measure documenting severity of performance anxiety experiences, such as sweaty palms, quickened pulse, etc. and sum together responses on the questions for each individual to get an overall score. Using the SPSS instructions above, an instructor or students can then calculate an ANOVA, with gender being the independent variable, and performance anxiety being the dependent variable.

2. Design an experiment, such as jumping large/small, heavy/light origami frogs as described in "Activity-Based Statistics" (Schaeffer, Gnanadesikan, Watkins, & Witmer, 1996). Using the SPSS instructions above, an instructor or students can then calculate an ANOVA with the student-gathered data.
3. Descriptively or through the use of an ANOVA, use a demographic grouping variable of your choosing (such as grade, instrument, or gender) to compare the achievement of your students on an achievement or classroom test.

Recommended Reading

Frankel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). Boston, MA: McGraw-Hill.

Huck, S. W. (2012). *Reading statistics and research* (6th ed.). Boston, MA: Pearson. Prentice Hall.