Getting Started with Quantitative Methods in Physics Education Research

Lin Ding* and Xiufeng Liu**

* School of Teaching and Learning, The Ohio State University, Columbus, OH 43210
** Department of Learning and Instruction, State University of New York, Buffalo, NY 14260

Abstract:

In this article we provide a brief overview of three groups of quantitative research methods commonly used in physics education research (PER): *descriptive statistics*, *inferential statistics*, and *measurement instrument development and validation*. These quantitative research methods are used respectively in three major types of PER, namely *survey research*, *experimental/quasi-experimental studies,* and *measurement and evaluation studies*. In order to highlight the importance of the close alignment between research questions and selected quantitative research methods, we review these quantitative techniques within each research type from three perspectives: data collection, data analysis, and result interpretation. We discuss the purpose, key aspects and potential issues of each quantitative technique, and where possible, specific PER studies are included as examples to illustrate how these methods fulfill specific research goals.

1. Introduction

Physics education research (PER) uses many forms of inquiry to investigate the learning and teaching of physics. Broadly speaking, there are three types of research methods that are used frequently in PER: qualitative methods, quantitative methods, and mixed methods.[1-3] In a previous volume of *Getting Started in PER*, Otero and Harlow laid out useful guidelines for carrying out qualitative studies and briefly touched on topics related to mixed approaches.[4] In this article, we introduce basic quantitative methods for empirical investigations and discuss their purposes, procedures and potential issues in the context of physics education research. Where possible, we also discuss studies from previous literature to illustrate how various quantitative techniques are used in PER.

Before starting, it is important to note that this article is not intended to be a comprehensive review of all quantitative methods in educational research, nor is it meant to be an introduction to basic statistics. Instead, all quantitative methods discussed in this article are tied closely to theory-oriented research questions of interest to PER.

With these notes in mind, we first start with a brief introduction to (1) the role of quantitative methods in PER, (2) the differences between quantitative methods in PER and those in other areas of physics, and (3) the nature of quantitative research questions. We then discuss respectively in Sections 2–4 some basic PER quantitative methods that are commonly used in three different types of PER studies: (1) survey research, (2) experimental/quasi-experimental studies, and (3) measurement and evaluation studies. Finally, a summary of useful references to other resources is provided in Section 5.

1.1 The Role of Quantitative Methods in Physics Education Research

To best appreciate quantitative studies, we need to understand when and why quantitative methods are involved in physics education research. Since the field of PER emerged mainly from physics and by physicists, there has historically been an emphasis on quantitative research in PER. Generally speaking, a quantitative method is used when a researcher wishes to quantify observations through some statistical techniques in order to better describe, explain, and make inferences about certain events, ideas or actions in physics education.

Quantitative methods have many advantages compared to other research methods. As opposed to qualitative PER studies, quantitative methods allow a researcher to focus on the variables of interest in order to study the relationships or even causal relationships among the variables.[5, 6] Also, because quantitative methods mostly deal with numerical data, it is conceivable that fewer human biases are introduced in processing and communicating information.[5, 6]

However, it is important to remember that these advantages are based on a crucial premise; that is, data collected from empirical PER studies must be numerical or at least quantifiable in some form. Quantifying observations is called measurement. Since numerical data are necessary for quantitative methods, measurement is the foundation and often the starting point of quantitative methods.

1.2 Differences between quantitative methods in PER and those in other physics areas

As we embark on any quantitative PER study, it is helpful to pause and think about the differences between the quantitative techniques used in PER and those used in other areas of physics.

One unique aspect of quantitative PER studies that separates them from those in other fields of physics is subjects of interest. Measurement in traditional physics involves physical entities or properties of physical entities (even as intangible as electric field or temperature). These subjects of interest have been well defined and have a universally agreed-upon set of variables associated with them. Measurement and data analysis of these variables typically follow some well-established procedures that can facilitate a physicist's subsequent efforts in building models of how the physical world works. On the other hand, quantitative PER studies almost always seek to investigate non-physical characteristics of human beings engaged in teaching and learning. These characteristics, such as conceptual understanding, reasoning skills, scientific practices, beliefs, attitudes, and epistemologies, are known as constructs.[7] Researchers often have not yet reached a consensus with regard to the definitions of these abstract constructs, not to mention that they usually lack a set of commonly accepted variables for measurement or analysis. This unarguably makes quantitative studies in PER challenging to conduct properly.

Another unique aspect of quantitative PER studies relates to measurement instruments. In traditional physics, many measurement tools are globally calibrated and standardized. Even with tools that are designed for use in different measurement systems, standard unit conversion still allows quantitative analysis to be readily comparable across diverse situations. However, in PER there is no definitive set of tools that are globally accepted for measurement, although researchers have been striving to reach common ground in this matter.

Even if there were a globally standardized measurement of some variables of interest in PER, quantitative data analysis would still differ drastically from that in other physics areas because of the difference in the nature of quantitative data. Generally speaking, there are four different scales of quantitative data: *nominal*, *ordinal*, *interval*, and *ratio*.[8]

- *Nominal data*, also known as a type of categorical data, is discrete and has no order. For instance, data that indicates student gender is considered as nominal. In practical analysis, we may use the number of "1" to represent male and "0" for female, but it doesn't make sense to say that 1 is greater than 0 in this case.[a]
- *Ordinal data*, another type of categorical data, is also discrete but has order. Consider the ratings in the Colorado Learning Attitudes Survey about Science or the Maryland Physics Expectation Survey.[9] The five rating points, ranging from 1 (strongly disagree) to 5 (strongly agree), indicate an increasing level of agreement. So, a choice of 5 means a higher level of endorsement than a choice of 1, but the difference in quantity between 5 and 4 (agree) may not necessarily be the same as the difference between 4 (agree) and 3 (neutral).[b]
- *Interval data* is continuous on a scale of equal intervals, and has order. A temperature reading measured either in Celsius or Fahrenheit is an example of interval data. Regardless of where it is located on the scale, one degree of temperature change indicates the same amount of thermal energy change in an object. In other

[a] In fact, we may also use symbols or letters to represent normal data. For instance, we can use F to indicate female and M to indicate male.

[b] As with nominal data, ordinal data also can be represented by using symbols, but these symbols follow a certain hierarchical order. For instance, we may use SD, D, N, A, and SA to represent respectively "strongly disagree", "disagree", "neutral", "agree" and "strongly agree". The sequence of these five symbols indicates an increasing level of endorsement.

words, the difference between 1°C and 2°C is the same as the difference between 20°C and 21°C.

- *Ratio data* not only has all the properties of interval data, but also has meaning for an absolute origin (i.e., 0) and a ratio between two values. Mass is a ratio variable. It is legitimate for us to say the mass of a proton is 1836 times that of an electron.

In traditional physics, quantitative data is mostly at the level of interval and ratio.[c] This means that we can perform a number of mathematical operations with the data, and the subsequent results still hold physical meanings. As mentioned above, adding or subtracting temperature readings and calculating the ratio of two masses yield meaningful outcomes. On the other hand, quantitative data in PER are mostly nominal or ordinal; they are seldom interval and never ratio. Strictly speaking, even test scores from the Force Concept Inventory (FCI),[10] Force and Motion Conceptual Evaluation (FMCE),[11] Conceptual Survey of Electricity and Magnetism (CSEM),[12] and Brief Electricity and Magnetism Assessment (BEMA),[13] to name but a few, do not produce interval data. This is because we cannot claim that the difference in understanding of force concepts between two students who score 1 and 2 on FCI is the same as the difference between another two students who score 20 and 21 on the same test.[14-16] Nevertheless, researchers often treat test scores like these as a close approximation to interval data, if the scores of a student population follow a Gaussian distribution. In general, data at the ordinal and nominal levels (also known as categorical data) are typically presented in the form of frequencies and are subject to non-parametric statistical analysis, whereas data at the interval and ratio level (also known as continuous data) allow for calculating means and standard deviations and are amenable to parametric statistical analysis.[17] Simply put, parametric statistics assume a normal distribution of the variable in the population from which the data is collected. In the case of continuous data, when the sample is adequately large, the normality assumption often holds. But when it comes to categorical data or continuous data of small sizes, it is likely that the normality assumption will be violated. In that case, traditional parametric statistical analyses are no longer valid. Since non-parametric statistics do not assume a normal distribution, they can be used in many cases where the traditional parametric statistics are no longer appropriate.

[c] Of course, there are exceptions. For example, electron spin is of discrete states, and therefore is not at the interval level.

Given the above nature of quantitative data in PER, it is conceivable that a satisfactory level of validity and reliability in PER quantitative studies is difficult to achieve, perhaps more so than in other areas of physics. Since discussion about validity and reliability has been covered in Engelhardt's[18] introduction to classical test theory, we avoid repetition by referring readers to the previous volume for more details. However, we want to emphasize that no matter how carefully quantitative data in PER are to be measured and analyzed, we can only use observable results to infer what we are truly interested in, and oftentimes those that are *not* directly observable are what we are truly interested in. This is why validity and reliability in quantitative PER studies are always difficult problems to tackle.

1.3 Nature of Quantitative Research Questions

Depending on the nature of research questions one attempts to answer in PER, a researcher may choose to use quantitative techniques in different manners. Primarily, the researcher has the following three choices to meet his/her needs. One is descriptive use of statistical techniques. This type of quantitative techniques is suitable for *survey research*[18] that aims to identify and document some universal patterns in a large number of subjects. For example, the researcher may want to know, "What is the average performance for students at State University X on the Force Concept Inventory?" Using descriptive statistical techniques can be appropriate for addressing this question.

The second choice is inferential use of statistical techniques. This is particularly useful for *experimental/quasi-experimental studies*, in which individual subjects (in experimental designs) or cohorts of subjects (in quasi-experimental designs) are randomly sampled and assigned into either a control group or treatment groups.[19, 20] By comparing and contrasting the different groups, the researcher can make inferences about the effectiveness of some treatments. An example research question that can be answered by using inferential techniques is, "Do students in interactive engagement classrooms hold more expert-like views about knowledge and learning, as measured by the Maryland Physics Expectations Survey (MPEX),[21] than those in traditional classrooms?" The third type of quantitative technique is instrument development and calibration. These quantitative techniques are regularly used in *evaluation studies*[6, 22] to develop measurement instruments that can match a specific evaluation plan and to establish validity and reliability evidence for the instruments. An example research question that can be answered by this

type of study is: "To what extent do the psychometric properties of the test items indicate that all questions measure the same construct as they are supposed to?" In a previous volume of *Getting Started in PER,* Engelhardt reviewed classical test theory as a tool for developing PER assessment instruments. In this article, we briefly discuss Rasch modeling[23-26] as an alternative quantitative approach to instrument validation.

More details on these three types of quantitative techniques are expanded in the next sections (Sections 2-4). In Section 5, we provide a short summary and some suggestions for how to further pursue quantitative research methods in PER, as well as in a broader educational research context.

2. Descriptive quantitative methods in survey research

Descriptive statistics are commonly used in *survey research* to quantitatively describe some aspect(s) of interest in large samples of a population.[27] When preparing for this type of study and selecting appropriate quantitative techniques for data analysis, a researcher must be clear about the purposes, procedures, and possible issues of his/her work. In what follows, we discuss some important items for consideration regarding data collection, data analysis, and result interpretation in PER survey research.

2.1 Prepare for data collection in survey research

Before collecting data, a researcher must have some clearly articulated research questions or hypotheses he/she wants to pursue. These questions or hypotheses should include specific variables of interest, survey tools, and subject samples.

Variables of interest reflect *what* a researcher wants to observe or measure; for example, the researcher may be interested in investigating students' conceptual understanding of a physics topic or their epistemologies about knowledge and learning of physics. But these variables are still vague and do not yield immediately observable consequences. Therefore, it is crucial to specify a set of observables that can be directly measured. In the example of conceptual understanding, the researcher may explicitly consider measuring students' performance

(scores) on conceptual questions related to a chosen physics topic. However, justification may be needed, through a theoretical framework and/or literature review, to support the link between the observables (student scores) and the unobservable variables (student conceptual understanding).

Once variables of interest are clearly defined, the researcher then needs to consider *how* to carry out the measurement. A key aspect is to find an appropriate survey tool that can serve the purpose of measuring the variables of interest. A number of survey instruments and concept inventories have been created within science education in general and physics education in particular. They cover a wide range of topics, including content learning, general reasoning skills, and epistemologies. A researcher can select appropriate instrument(s) to best suit his/her needs.[d] Another important aspect for consideration is to specify the context under which a survey is used. Depending on the research goals, a survey can be given at different time points, for example, before instruction, during instruction, or after instruction. It can be administered as a paper-and-pencil test, a mail-in questionnaire, an online poll, or in other forms such as interviews. Given the wide variety of ways for using a survey tool, a researcher must think through the options carefully before collecting data.

Also crucial is to identify with *whom* a measurement is conducted. This relates directly to subject samples in a study. Since quantitative results acquired from survey research should ideally reflect the patterns of a population, it is useful to choose representative samples for data collection. Some important factors for consideration include the role of subjects (learners, educators, parents, researchers, administrators, policy makers, etc.), the number of subjects, their academic background, age, gender, ethnicity, nationality, and other factors that may be relevant to the research questions one seeks to answer.

2.2 Select appropriate statistical techniques for data analysis in survey research

[d] If no suitable survey tools exist, one needs to develop his/her own measurement instrument. Section 4 introduces Rasch modeling as a useful quantitative approach to instrument development and validation. Readers can also refer to reference 18 for more details on instrument validation through the classical test theory.

After data are collected, a researcher needs to carefully select appropriate statistical techniques for analysis. Common methods used in descriptive statistics include analyzing the mean and median, standard deviation and standard error, standard scores, correlation, and regression. Except for correlation and regression, which require two or more sets of data, all other statistics can be used to descriptively analyze one data set. In the following, we first discuss the methods for one-data-set situations and then discuss correlation and regression.

2.2.1 Descriptive analysis for one data set

Mean and median perhaps are the most commonly used statistics for describing a set of data collected from a survey. For data at the interval/ratio level, estimating a mean is simply calculating the arithmetic average. Note that we can perform addition and subtraction with data at the interval/ratio level, so calculating means makes sense. The median is the middle number for a string of data that are arranged in the order of increasing or decreasing values. Depending on the data distribution, a median can be close to or far from a mean. If the two are close in value, then the data distribution is somewhat symmetric. Otherwise, the distribution is skewed.

Because nominal/ordinal data have limitations on what mathematical operations can be performed, calculating mean values may not be suitable for nominal/ordinal data. For instance, we may assign a number, say, 1, 2, 3, or 4—a set of nominal values—to each student in four classes that adopt different pedagogies. In this case, it makes little sense to calculate a mean of the assigned numbers among students. For ordinal data it is also problematic to analyze means. Consider a case where we have four cohorts of equal numbers of students at different grade levels—grade 2, 5, 8, and 10. While it is logical to say that grade 5 is higher than grade 2, it certainly is awkward to say that the average grade level for these students is 6.25. That said, in practical applications ordinal data are often approximated to interval data if a normal distribution is assumed. As mentioned before, researchers in PER frequently take test total scores on most PER assessments as interval data. However, caution is needed when it comes to survey questions that use rating scales. Unless the range of rating points is sufficiently wide to allow a possible normal distribution, approximation to interval data in most cases may not be appropriate (see Section 4.1 for more details). For example, when analyzing the Colorado Learning Attitudes about Science Survey (CLASS) to investigate student epistemological beliefs, Adams et al. carefully avoided using ordinal-level

raw data for direct calculation, and instead they collapsed student ratings into two categories and calculated percentages of student responses that matched experts' responses for quantitative analysis.[9]

Standard deviation and standard error both reflect fluctuations in data. However, they contain different meanings. Standard deviation is a measure of variation in the sample from which the data is collected, whereas standard error is an estimate of the accuracy for using a sample mean to represent a population mean. Typically, the larger the sample is, the better the estimate for the population is, and therefore the smaller the standard error is. It follows that the standard deviation (S) and standard error (σ) are related by a factor of the square root of N, where N is the number of subjects in the sample:

$$\sigma = \frac{S}{\sqrt{N}}.$$

It is worth noting that for the same reasons as mentioned above, the standard deviation and standard error are suitable for interval/ratio data but may not be so for ordinal/nominal data.

Standard scores, also known as Z scores, are used to convert raw data into another form that can indicate how many standard deviations each individual data point is above or below the mean. This conversion can allow a researcher to have a firm grip of the relative position of the data points. Standard scores are calculated by taking the difference between each data point (x) and the mean (μ), and then dividing it by the standard deviation (S):

$$Z = \frac{x - \mu}{S}$$

Typically, if a set of data follow a normal distribution,[e] the standard scores are referred to as Z scores. Otherwise, we call them t scores. Again, standard scores are suitable for interval/ratio data, but may not be appropriate for ordinal/nominal data.

What, then, is appropriate for ordinal/nominal data? Descriptive statistics for ordinal/nominal data is often in the form of frequencies. One approach researchers often use is to calculate the percentages of different categories for data at the ordinal/nominal level. For instance, in order to examine the

[e] One can plot a histogram to visually check if the data follows a normal distribution (a bell-shaped, symmetrical curve). If a more definitive answer is needed in judging the normality, one can use the Kolmogorov-Smirnov test for goodness of fit. See reference 17 for more details.

gender distribution in a class, a researcher can count the numbers of male and female students and then calculate the percentages to find out the distribution. If multiple ways to categorize the same sample of subjects exist, as is often the case, a researcher may consider using a contingency table or a matrix to represent data distribution. For example, a researcher may wish to know student distributions in terms of both gender and ethnicity. In this case, a table with rows and columns showing gender and ethnicity respectively can help facilitate this type of data analysis. (See Figure 1.) Such a table is often called a contingency table.[28]

Ethnicity / Gender	Asian	Black	Caucasian	Hispanic	Others
Female					
Male					

Figure 1: A Contingency Table with Rows and Columns Indicating Gender and Ethnicity, Respectively

2.2.2 Descriptive analysis for two or more data sets

Correlation is another common descriptive statistical technique. It is used to analyze the relationships between two sets of data. For interval/ratio data, we often choose the Pearson coefficient to measure the extent to which two sets of data are linearly related. In PER, this technique is frequently adopted to investigate how students' performances on two different assessments correlate with each other. For instance, Thornton et al. administered both the FCI and Force and Motion Conceptual Evaluation (FMCE) to approximately 2000 students in a studio physics course at Rensselaer Polytechnic Institute.[29] Student scores on both assessments were taken as interval data and used for calculating a Pearson correlation. Results showed despite the differences in content coverage and the number of questions per topic in the two tests, there was a strong positive correlation ($r = 0.78$, out of a maximum of 1) between the FMCE and FCI scores, indicating that there is a large overlap in what is measured by the two tests.

It is worth noting that for data at the ordinal or nominal levels, the Pearson correlation is no longer a suitable statistic. Instead, one has to use other measures. For ordinal data, one may consider using Spearman's Rho (ρ) to calculate correlation.[30] This approach generates a rank order correlation to

examine if a high ranking in one ordinal data set corresponds to a similar ranking in the second ordinal data set. As for nominal data, calculation of correlation goes beyond the scope of descriptive statistics and in fact alludes to inferential statistics, such as the Chi-square statistic used for testing association. We defer relevant discussion to Section 3. However, when it comes to special cases in which two sets of nominal data are both dichotomous (either 1 or 0), one can use Phi (φ) to calculate correlation—a simplified Pearson correlation for dichotomous data.[31] For example, Heller and Huffman used this method to calculate Phi correlations between dichotomously scored FCI items.[32, 33] Based on these correlations, they performed a principle component analysis to check if the FCI items fall under six dimensions of content topics as claimed by the FCI designers.

A final caveat on correlation is that it is often sensitive to sample sizes and therefore, when reporting a correlation, researchers may also need to include its significance level to show if it is statistically meaningful, given the sample size. The topic of testing significance involves inferential statistics. See Section 3 for more details on inferential quantitative techniques.

Regression analysis is an extension of correlation analysis and involves two or more data sets. This analysis looks into the relationship between a dependent variable (DV) and one or more independent variables (IVs) so as to model how the DV is affected by the IVs. Among various types of relationships, linearity is the simplest one; it means that a change in each IV is directly proportional to a change in the DV. Graphically, such a relation can be represented by a straight line. Results of linear regression analysis can reveal the strength (correlation coefficient squared) as well as the magnitude (slope of a regression line) of the relationship between the DV and each IV. It is worth noting that strength and magnitude are two different concepts. Strength reflects how much variance in the DV can be accounted for by the IVs, therefore indicating the degree to which the relationship between the DV and IVs can be modeled by a linear regression. Magnitude, on the other hand, reflects how the size of a change in the DV is affected by every unit change in each IV, regardless of whether a linear relationship is warranted or not. Sometimes, the magnitude of the relationship between the DV and each IV is small (small slope of a regression line), but the strength of a linear relationship between them is high (all dots in a close proximity to the line), or *vice versa*.

In physics education research, linear regression is frequently used to describe how a variable of interest changes with other variables. For instance, Kortemeyer[34] studied the relationship between students' approaches to physics, as measured by their online discussion behaviors, and learning outcomes measured by student FCI scores. In this study, Kortemeyer used FCI posttest score as a dependent variable, and FCI pretest score and percentages of student solution-oriented behaviors in online discussion as two independent variables. Student solution-oriented behaviors were identified as discussions that focused on getting a correct answer without dealing with the deep physics of a problem. By conducting a linear regression analysis, Kortemeyer found the following relation:

$$\text{Post FCI} = 7.606 + 0.857 \times \text{Pre FCI} + (-0.042) \times \text{Solution-oriented Behaviors}$$

with 47.9% of the variance in "Post FCI" accounted for by "Pre FCI" and "Solution-oriented Behaviors." As pointed out by the author, the negative slope of –0.042 indicated that for every 10% increase in solution-oriented discussion student posttest FCI score would decrease by 0.42 points, controlling for the FCI pretest score.

As for nonlinear regression, it is beyond the scope of basic descriptive statistics, so we only give it a passing note. In PER, nonlinear regression analysis is not often used. One of a few examples is a study conducted by Henderson et al.[35] They applied logistic regression to seek the relationship between physics instructors' knowledge of research-based instructional strategies and 20 independent variables. Readers can refer to the original paper for more details. Another case that utilizes nonlinear regression is Rasch modeling. Details about Rasch modeling and its roots in logistic regression are briefly introduced in Section 4.

2.3 Interpret analysis results in survey research

Although descriptive statistics are fairly straightforward, interpretation of analysis results still calls for attention. It is important to remember that while the ultimate purpose of applying descriptive analysis is to make generalizations about some aspects of a population, what a researcher has at hand is information about a sample that is hopefully representative of the population. One must be mindful of this gap. When interpreting descriptive statistics, make sure to attend to sample characteristics such as the number of participants, their backgrounds, and the context in which the study is conducted. Where possible, discussions on issues, such as

whether using an alternative similar survey tool will lead to comparable results, are also helpful.

A particular note worth mentioning here is that when dealing with correlation, one should not confuse it with causation. A high correlation can be due to various reasons, including some spurious relationships between two sets of variables, and it does not necessarily indicate a causal relationship. On the flip side, without controlling for confounding factors, even causally related variables can have a rather low correlation, so caution is needed in interpreting correlation statistics.

3. Inferential quantitative methods in experimental/quasi-experimental studies

Experimental and quasi-experimental studies in PER allow a researcher to make claims about the effect of some treatment or intervention through comparisons between two or more events. If such comparisons involve quantitative analyses, inferential statistics then become a useful tool for these types of studies.[27] Since appropriate use of inferential statistics in experimental/quasi-experimental PER is inseparable from careful research designs, we discuss in this section three perspectives, relating to data collection, data analysis, and data interpretation respectively.

3.1 Prepare for data collection in experimental/quasi-experimental studies

In addition to those noted in survey research, (quasi)experimental studies have unique requirements that one needs to take into consideration before starting data gathering. A key issue to bear in mind is that variables of interest need to be clearly specified and be kept separate, as much as possible, from other variables of non-interest that may exert influences on the outcomes. For instance, a researcher may want to compare student exit performance on FCI as a result of two different physics instructions—traditional *vs.* interactive instruction. Conceivably, there are many factors other than instruction types that can influence student FCI scores. Therefore, it is crucial that a comparison is made between two groups of students who have similar academic backgrounds and receive comparable treatments other than different instruction. As such, a difference in student FCI scores, if any, can then be attributed to the difference in instruction with a high degree of confidence.

To ensure that subjects in different groups are comparable and only differ in the variables of interest, a researcher can take several measures. One way is to randomly assign subjects into groups. In an ideal *experimental design*, each individual is randomly assigned either into a control group or a treatment group, and the positive and negative effects in each variable of non-interest will hopefully cancel each other out due to randomness.[36] However, in reality a researcher does not always have the luxury to do so, as assigning human subjects differs vastly from preparing material samples that can be at the researcher's disposal. Oftentimes, our human subjects in PER are constrained by certain boundary conditions, such as classes, schools, and districts. So, in *quasi-experimental designs* we randomly assign each class, school, or district rather than each individual into different groups.[36] In this case, it is useful to conduct *a priori* comparisons between the groups on some variables that the researcher suspects may cause some confounding results—results that are considered as undesirable because of variables beyond the researcher's control. As an alternative to *a priori* comparisons, a researcher may choose to use *post hoc* comparisons to check if the groups he/she has studied are comparable. However, *post hoc* comparisons run risks of being unable to salvage an otherwise rescuable study if *a priori* comparisons were conducted and necessary re-assignments of subjects were carried out beforehand.

Before starting data collection for a (quasi)experimental study, a researcher also needs to pre-determine a significance level, an expected effect size, and/or the number of subjects needed for meeting this effect size and the significance level. In inferential statistics, a comparison between groups is always framed in a pair of hypotheses consisting of a null hypothesis (H_0) and an alternative hypothesis (H_a).[8] The null hypothesis states that there is no significant difference between the groups at the level of α, and conversely the alternative hypothesis states that there is a significant difference. Here α is what we call a predetermined significance level, indicating the maximum probability of getting a statistic (z-value, t-value, F-value, Chi-Square, etc.) less than what is found in the data. A commonly used value for α is 5%, which is equivalent to an odds of 1 out of 20. Note that the 5% alpha value is merely a convention; a researcher should choose a significance level most suitable for his/her research design (also see Section 3.3 for more discussions).

When deciding on which hypothesis to accept or reject, a PER researcher can potentially make two types of errors. One is called Type I error, which overestimates the significance of a difference by mistakenly rejecting the null hypothesis (H_0). The probability of making this type of error is the

same as the significance level of a statistic test α. The other is Type II error, which underestimates the significance of a difference by failing to reject the null hypothesis (H_0). We use β to indicate the probability of making a Type II error. Conceivably, 1- β indicates the probability of making a correct claim of a significant difference by correctly rejecting a null hypothesis, and hence it is also called the power of a statistic test.

While a difference found between groups may be significant, it can take on various sizes of magnitude. The magnitude of a difference is often referred to as effect size.[37] In the case of a 2-group comparison, for example, effect size is calculated as the difference between the means of the two groups divided by the standard deviation of one group (usually of the control group) or both groups. According to Cohen,[37, 38] a difference of an effect size less than 0.2 in the 2-group comparison case is considered as small, between 0.2 and 0.8 as medium, and greater than 0.8 as large. (More information on effect size for other types of comparisons can be found in references 37 and 38.) Sometimes a small difference can be statistically significant; whereas other times a large or medium difference shows no statistical significance at all. These occurrences have much to do with the sample sizes involved in a comparison. If a significant difference is anticipated due to some theory-laden framework, then a researcher needs to be sure that the samples he/she has in the study are sufficiently large to be able to detect such a difference. On the other hand, too large of a sample is unnecessary and can be a waste of time and resources. So, *a priori* analysis is often needed to calculate the minimally sufficient sample size for each group. With a pre-determined significance level and a desired effect size, a researcher can determine the number of subjects in each group needed to carry out the study. There are several free online programs that can help researchers perform *a priori* calculations; these include G*Power analysis[37] and A Priori Sample Size Calculator.[39]

3.2 Select appropriate statistical techniques for data analysis in experimental/quasi-experimental studies

Depending on the goals and the design of an experimental PER study, a comparison can be made either at different time points with the same sample of subjects or across different samples at the same time. Since the statistical techniques required in these two types of comparisons are different, we discuss them separately.

3.2.1 Inferential statistics for comparisons of one sample over time

Experiments or quasi-experiments in PER sometimes require a researcher to take measurement of the same sample at two different time points, between which some type of treatment or intervention is implemented. By making comparisons between the measurements at the two times, a researcher can study the effect of the treatment or intervention. Common techniques for carrying out this kind of comparisons include the repeated *t*-test and the McNemar test.

Repeated *t*-test, also called a paired *t*-test, is suitable for data at the interval/ratio level. It is an extension of the one-sample *t*-test (or univariate *t*-test) that is used for comparing the mean score of a sample with a specific number. In a repeated *t*-test, there are two sets of data associated with a single sample, measured respectively before and after a treatment (called "pre" or "post" hereafter as a shorthand). The pre and post data should be matched for each subject, and those with missing data need to be deleted. The difference between each matched data pair, also known as "gain" (gain = post score – pre score), is key to a repeated *t*-test. By comparing the "gains" with zero, a researcher can choose between a null hypothesis (H_0) stating there is no significant difference between the pre and post measurements, or an alternative hypothesis (H_a) suggesting otherwise.

The repeated *t*-test produces a statistic that has a symmetric bell-shaped distribution (slightly more spread out than a normal distribution) and is called the *t* statistic. The spread of the *t* distribution is dependent on the degree of freedom (d*f*) which simply is the maximum number of values in the data that are allowed to vary. In a repeated *t*-test, the degree of freedom is equal to the total number of subjects minus one (*N*-1). Associated with the *t* statistic and degree of freedom is a *p* value. It reflects the probability of obtaining a *t* statistic with a value no greater than what is found in the data. If the *p* value is smaller than a pre-determined significance level (α), then a researcher can claim that the "gains" are statistically different from zero; in other words, the difference between the pre and post data is significant. This detected significance may allow a researcher to further claim that some treatment implemented between the two measurements has produced statistically meaningful effects—effects not due to measurement error—on the variable of interest (for example, student performance on a physics assessment). Otherwise, the treatment has yielded little measurable effect.

An important assumption about the repeated *t*-test (and in fact about any *t*-test) is that the population from which the sample is randomly selected

follows a normal distribution on the measured variable. Typically, a repeated *t*-test is robust if a sample contains 30 or more subjects. In the case of smaller samples for which a normality assumption does not hold, one may consider using the Wilcoxon signed rank test[40, 41]—a non-parametric version of the repeated *t*-test—for conducting paired comparisons.

In PER, this type of study is frequently used to investigate changes in student learning over time. For instance, Pollock[42] conducted a longitudinal study to measure changes in student conceptual learning over a period of 3 years. In this study, all students first took a tutorial-based introductory physics course on electricity and magnetism (E&M)—namely Physics II—during their freshman year. A subset of the students (N = 38) then took two upper-division physics-major E&M courses—namely Physics 301 and Physics 302—during their junior year. Pollock administered BEMA as both a pre and a post test before and after each of these courses. In an effort to study knowledge retention during the period between Physics II and Physics 301 courses, Pollock used a repeated *t*-test to compare student post-Physics-II BEMA scores with their pre-Physics-301 BEMA scores and found a 5% decrease in student performance. Although this is a small decrease compared to what has been reported in educational-psychological studies in a relatively long time period, Pollock still found it to be statistically significant at the $p = 0.01$ level.

In addition to using gains (arithmetic differences between pre and post data) to express growth (or decrease) over time, researchers in PER also calculate "normalized gain".[43] This is defined as the ratio between an actual change and a potential maximum change:

$$\textbf{Normalized Gain} = \frac{\textbf{Post Score} - \textbf{Pre Score}}{\textbf{Full Score} - \textbf{Pre Score}} \times 100\%$$

An actual change is the same as a gain (gain = post score – pre score), and a potential maximum change is calculated as the difference between a perfect score and a pre score (maximum change = full score – pre score). Since its first introduction to PER by Hake,[43] normalized gain has been widely adopted. Researchers often use normalized gains (or gains) for direct comparison with zero to examine if there is a significant change between student pre and post performances.

McNemar test is a statistical technique suitable for repeated measurements that involve categorical (nominal/ordinal) data with only two possible

values (either 0 or 1, for example). As with a repeated *t*-test, data in McNemar tests must be matched for each subject, and those with missing data need to be deleted. This test focuses on the changes between two measurements of the same sample (i.e., changes from 0 to 1 and from 1 to 0) but ignores those that do not show change (see Figure 2). Based on the numbers of such changes, the McNemar produces a Chi-square statistic (χ^2) and an associated *p* value. If the *p* value is smaller than a prescribed significance level (α), it suggests that the difference between the two measurements is significant. The degree of freedom (d*f*) for the McNemar test is always 1.

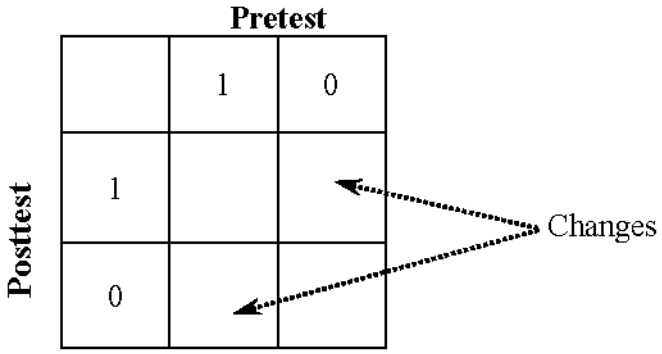

Figure 2: McNemar Test for Repeated Measurement--a Pretest and a Posttest on a Dichotomously Scored Item--for the Same Sample

Note that the McNemar test differs from the univariate Chi-square test, although both are used for testing categorical data of one sample. The former is only suitable for comparing matched data of repeated measurements; whereas the latter is used to test whether the categorical data of a single measurement follows an expected distribution. Data in the McNemar test must be binominal, whereas data in the univariate Chi-square test can have multiple categories/levels. One advantage of a McNemar test is that it essentially is a nonparametric test, so it can be used for analysis of small samples.

In PER, the McNemar test is particularly useful for comparing students' performances on questions that are dichotomously scored (1 for correct and 0 for incorrect). For instance, Stewart, Griffin, and Stewart[44] administered two versions of 10 selected FCI questions to over 650 students, one being the original FCI version and the other containing modified questions with altered contextual features. Each student in this study completed both versions and had two scores. Specifically, half of the students took the original FCI first before taking the modified version, and

the other half did the reverse. As such, the order effect was reduced. To investigate the effect of change in question context on student responses, the authors conducted a McNemar test for the 10 questions and found that the difference in student performance on the two versions was significant at the $p \leq 0.01$ level for all the 10 questions. This result prompted the authors to think that the context change could have noticeable influences on students' responses to individual multiple-choice questions.

3.2.2 Inferential statistics for comparison of two or more samples

When it comes to comparisons between two or more samples, a typical scenario is that one sample is a control group, designed to be a baseline, and the others are experiment groups in which subjects receive different degrees of some treatment or intervention. By comparing these groups, a researcher can make inferences about the effect of the treatment or intervention. Common techniques for inferential statistics include the two-sample *t*-test, one-way ANOVA (*F*-test), two-way ANOVA (*F*-test), ANCOVA (*F*-test), Chi-square test, and multi-way frequency test.

Two-sample *t*-test is a useful technique for comparing two independent samples. Assuming that the two samples differ in some aspect of interest but are comparable otherwise, we can test a hypothesis regarding whether or not the difference between them is significant. This test produces a *t* statistic and an associated *p* value. The degree of freedom of this test is equal to the total number of subjects minus two (*N*-2).

It is worth stressing that in practical application the two-sample *t*-test can be easily confused with a repeated *t*-test. Sometimes researchers mistakenly use a two-sample *t*-test for comparing data that are collected from a single sample. To avoid this common error, make sure to double-check that the data are collected from two separate samples. If not, the two-sample *t*-test is not an appropriate choice. For example, in longitudinal studies data collected from the same group of students at two different time points should be analyzed by using a repeated *t*-test, not a two-sample *t*-test. Another important aspect about the two-sample *t*-test relates to its assumptions about the population and samples. Similar to the repeated *t*-test, the two-sample *t*-test assumes that the populations from which the samples are randomly selected are normally distributed. Additionally, the variances and standard deviations of the two independent samples are assumed to be equal. Generally, a two-sample *t*-test is a robust test when the sample size is reasonably large ($n \geq 30$). In the case of small

samples, one may consider using the Mann-Whitney test[40, 41]—a nonparametric alternative—for making two-group comparisons.

In PER, the two-sample *t*-test perhaps is the most frequently used technique for comparing two independent groups of subjects. For example, Day and Bonn[45] developed a Concise Data Processing Assessment (CDPA). In order to examine how this assessment could be used to distinguish different populations, they administered CDPA to undergraduate students, graduate students, and faculty and conducted a series of *t*-tests to compare their scores. Results showed that there was a significant difference in CDPA scores between 1st year and 4th year undergraduate students, between 4th year undergraduates and faculty, and between graduate students and faculty (with all p values < 0.001). Additionally, there was a significant difference in student scores between the 1st year and 2nd year undergraduates who did not receive instruction on data processing in their previous lab courses ($p < 0.001$). Given these results, the authors came to a conclusion that the CDPA can be used for separating different populations along the novice-to-expert spectrum with regard to their data processing abilities.

ANOVA (*F*-test) is an abbreviation for "analysis of variance." It essentially is a significance test that uses the F distribution—a right-skewed continuous distribution—to detect differences among two or more groups of subjects. The null hypothesis (H_0) of an ANOVA test is that the means of all groups are equal, and the alternative hypothesis (H_a) is that at least two groups have different means. ANOVA calculates both within-groups variance and between-groups variance. Simply put, within-groups variance reveals how data points in each group disperse from their mean, whereas between-groups variance shows how far apart the groups spread from each other. Conceivably, if the dispersion within each group is noticeably smaller than the spread among the groups—in other words, if within-groups variance is considerably smaller than between-groups variance, then the difference among the groups becomes fairly evident. Otherwise, it can be difficult to detect differences among the groups. This idea is strikingly analogous to an imaging system's resolution ability in optics. Recall that Rayleigh calculated the ratio of wavelength to aperture's diameter as a measure for angular resolution. Similarly, ANOVA uses the ratio of between-groups variance to within-groups variance as a measure (F statistic) to determine if these groups are too close to be called different.

Depending on the number of independent categorical variables involved, there are different types of ANOVA, namely one-way ANOVA, two-way ANOVA and multi-way ANOVA. In one-way ANOVA, what is being compared (dependent variable at the interval/ratio level) is considered to be a function of only one categorical predictor (independent variable). So, comparisons are only made between the different categories of this single independent variable. For instance, Smith and Wittmann[46] attempted to compare student performance on the FMCE among three different tutorial-based curricula, namely Tutorials in Introductory Physics,[47] Activity-Based Tutorials,[47] and Open source Tutorials.[48, 49] They used the student FMCE score as a dependent variable and curriculum as an independent variable that contained three categories (classes). By using one-way ANOVA, a comparison of student FMCE scores therefore was made across these three categories (classes) of the independent variable.

In two-way ANOVA, what is being compared is a function of two categorical variables; for instance, a difference in student reasoning abilities may be accounted for by both curriculum and gender. In this case comparisons need to be made between the different categories of both independent variables, namely a comparison among classes that use different types of curricula (Class-1, Class-2…Class-*g*), as well as a comparison between the two genders (male and female). The purpose of these comparisons is to detect the *main effects* of the two independent variables. In practice, before testing the main effects, it is helpful to detect if there is any *interaction* between the two independent variables. (See Figure 3.)

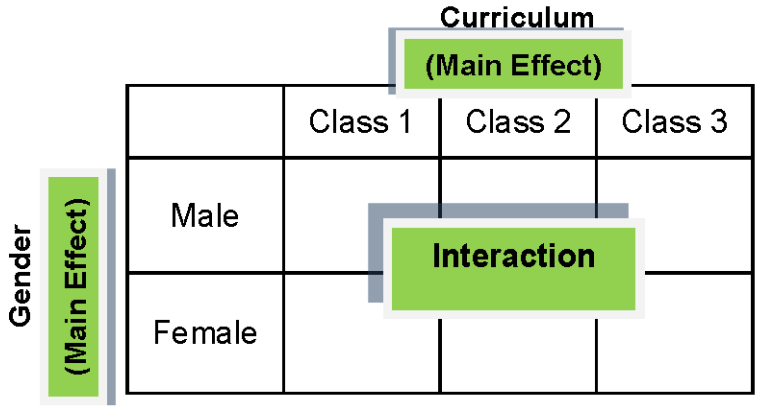

Figure 3: Two-way ANOVA Testing Both Main Effects and Interactions

An absence of interaction implies that the effect of either independent variable on the dependent variable is consistent for the different categories of the other. For instance, suppose that there is no significant interaction between the two independent variables, curriculum and gender. This means that the difference (be it small or large) in the dependent variable (student reasoning ability, for example) between male and female is constant across all the classes. Similarly, the differences in the dependent variable among all the classes remain constant for both genders as well. This absence of interaction allows us to proceed with comparisons for main effects. Otherwise, we have to compare each combination of class and gender separately, as the broad-brush comparisons for main effects do not yield interpretable information. In multi-way ANOVA, what is being compared is a function of three or more independent variables. Since the basic ideas behind multi-way ANOVA are the same as those for two-way ANOVA, we omit further discussion.

In cases of two or more dependent variables, ANOVA becomes multivariate ANOVA or MANOVA. The underlying core ideas for MANOVA are similar to those for ANOVA, but the interpretation of the effect on the dependent variables will be based on a composite variable. Due to space limits, this analysis is not further discussed in this article. Readers can see reference 8 for more details.

One-way ANOVA has two degrees of freedom, one for between-groups and the other for within-groups. The between-groups degree of freedom is the number of groups minus 1 ($df_{\text{between}} = g - 1$), and the within-groups degree of freedom is the total number of subjects minus the number of groups ($df_{\text{within}} = N - g$). If there happen to be only two groups in the independent variable, then one-way ANOVA is equivalent to the aforementioned two-sample *t*-test, and the *F* statistic is simply equal to the *t* statistic squared ($F = t^2$ for two groups).[8]

In two-way ANOVA, there are three degrees of freedom for between-groups (df_{between}) and one degree of freedom for within-groups (df_{within}). Suppose there are g_1 groups in the first independent variable and g_2 groups in the second independent variable. The between-groups degree of freedom is g_1-1 for the main effect of the first independent variable, g_2-1 for the main effect of the second independent variable, and $(g_1\text{-}1) \times (g_2\text{-}1)$ for the interaction between them. The within-groups degree of freedom is the total number of subjects minus the product of $g1$ and $g2$ ($df_{\text{within}} = N - g_1 \times g_2$).

Like any other statistical test, ANOVA has its own assumptions. Ideally, subjects need to be randomly selected and hence are representative of the populations. Also, the population for each group is assumed to follow a normal distribution, and the standard deviation is the same for each group. These are fairly strict assumptions, so in reality they are rarely satisfied. Practically, a researcher in PER should strive for random selection of subjects, because this is considered as the most important condition for a valid ANOVA. Sometimes a researcher may encounter a situation in which the number of subjects in each group is small and the normality assumption may not be legitimate. In cases like this, one may consider using the nonparametric Kruskal-Wallis[40, 41] test as an alternative to one-way ANOVA, and using the nonparametric Friedman's test[40] as an alternative to two-way ANOVA.[f]

In PER, one-way and two-way ANOVA are frequently used for comparing among multiple groups. In a study conducted by Ding et al.,[50] the CSEM was administered as both a pre and posttest to 1535 students enrolled in a calculus-based introductory electricity and magnetism course during a two-year period. These students were from various traditional classes taught by nine instructors using two different textbooks and two different homework delivery systems. In order to test whether or not student performance on the CSEM was different between these nine classes, the authors conducted a one-way ANOVA and found that there was no significant difference between these classes in the pretest scores, posttest scores, gains, or normalized gains (with all p values > 0.47). The authors concluded that despite the variations in instruction, traditional classes generated more or less the same results on student conceptual learning of electricity and magnetism topics measured by CSEM. This result allowed the authors to combine the nine classes into one group for subsequent aggregate analysis.

Two-way ANOVA is also a commonly used quantitative method in PER. For example, Chen et al. studied the effects of online homework and interactive engagement instruction on students' conceptual learning, measured by the Force Concept Inventory.[51] Two categorical independent variables—homework and instruction—were considered. In the homework

[f] It is worth noting that Friedman's test was originally designed for nonparametric analysis of variance involving multiple repeated measures. (Note that its difference from the repeated t-test lies in the fact that the aforementioned repeated t-test can only deal with 2 repeated measurements.) In repeated ANOVA, "subjects" are considered as an additional independent variable.

variable, there were two groups: online homework (OHW) and ungraded homework (UHW). In the instruction variable, there were also two groups—interactive engagement (IE) and non-interactive engagement (NIE). The authors first used student pre-instructional FCI scores as a dependent variable for a two-way ANOVA and found no significant interactions between the two independent variables and no significant main-effects either. Then the authors used student FCI normalized gains as a dependent variable for a two-way ANOVA. This time they detected a significant interaction between the two variables (with a p value at the order of 10^{-6}), suggesting the combined effect of online homework and IE instruction (positively) influenced student learning gains.

ANCOVA (F test) is an acronym for "analysis of covariance." It is a useful technique for analyzing data from quasi-experimental research designs. At the heart of this analysis is a combination of regression and ANOVA. Recall that the independent variables involved in ANOVA are categorical. While keeping these independent variables, ANCOVA adds an additional continuous (interval/ratio) predictor—often referred to as covariate—and assumes a linear relation between this covariate and a dependent variable. This covariate allows researchers to control for a quantity that they suspect has influence on the dependent variable. For example, a researcher may wish to compare student performance on a physics assessment across different classes that use differing pedagogies. However, he/she suspects that math skills may play a key role in the difference of students' physics assessment scores. In this case, the researcher can use ANCOVA for comparison by taking student math examination scores as a covariate and the class as a categorical predictor. Note that more than one covariate can be introduced to ANCOVA. However, for simplicity, we only discuss the situation of one covariate with one categorical variable, which is also known as one-way ANCOVA.

ANCOVA tests the same hypotheses as ANOVA, along with some more. One additional hypothesis in ANCOVA relates to the covariate; it tests whether or not there is a linear relationship between the dependent variable and the covariate. Additionally, ANCOVA tests if there is any interaction between the covariate and other categorical variables. In the absence of interaction, the regression lines between the dependent variable and the covariate are parallel for all the groups of each categorical predictor. In other words, the differences among the groups are constant regardless of the value of the covariate. For instance, in the above example where the researcher introduces math examination scores as a covariate, if he/she finds there is no significant interaction between the covariate (math

score) and the categorical variable (class), then the regression lines for all the classes are more or less parallel to each other. (See Figure 4.) This means that the differences between these classes are constant when controlling for student math examination scores. Otherwise, these regression lines are not all parallel, or to put in another way the differences between the classes are not constant. Typically, the significance of interaction should be tested first before one proceeds with other tests. If no significant interaction is detected, then a researcher can continue with what is discussed in ANOVA for subsequent tests. Otherwise, he/she may consider plotting regression lines to get a visual impression of how the covariate and the categorical variable interact. Such a plot is often called an interaction plot (see Figure 4).

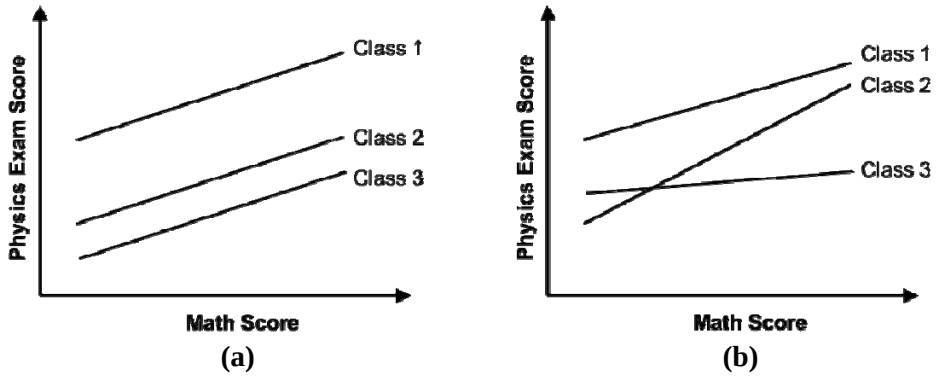

Figure 4: Interaction plots. (a) No interaction (b) with interaction

In one-way ANCOVA, the degree of freedom for testing the covariate is always equal to 1 ($df_{\text{covariate}} = 1$). For testing interaction between the covariate and each categorical predictor, the degree of freedom is equal to the number of groups for that predictor minus one ($df_{\text{interaction}} = g - 1$). Since ANCOVA has introduced a covariate as an additional independent variable, its within-groups degree of freedom is one less than that of ANOVA ($df_{\text{within}} = N - g - 1$). The between-groups degree of freedom remains the same as that of ANOVA ($df_{\text{between}} = g - 1$).

ANCOVA makes the same assumptions about the population data as ANOVA. As usual, random selection of samples and normal distribution of data are important assumptions to consider. In the case of small samples in which the normality assumption does not hold, a researcher may consider using a nonparametric rank analysis of covariance as an alternative.[52, 53]

In PER, ANCOVA is a useful technique for correcting pre-instructional differences among students when making comparisons (therefore suitable for quasi-experimental design). Brewe et al.[54] implemented physics Modeling Instruction for five years at an ethnically diverse university. In order to investigate whether or not this instruction equally benefitted students of different ethnicity and gender, the authors looked into students' conceptual learning measured by the Force Concept Inventory. Based on the literature, the authors believed that the difference in student math background could account for differential learning in physics. So they conducted two ANCOVA tests, using respectively gender (male/female) and ethnic representation (majority/minority) as a categorical independent variable. Both ANCOVA tests involved SAT I-Math scores as a covariate, and post FCI scores as a dependent variable. It was found that after controlling for the SAT I-Math scores, no statistical difference in post FCI performance was detected between the two ethnic representation groups, but there was a significant difference between male and female students. The authors inferred that Modeling instruction was a step toward greater equality in physics education, but at the same time more improvement could be made.

Chi-square test can be used to test if the frequency distributions across different groups of one variable remain the same for each group of a second variable. Both variables in a chi-square test need to be categorical, containing two or more mutually exclusive groups. (For instance, male and female are generally considered as two mutually exclusive groups in the variable of sex.) If the distributions of one variable are identical for each group of the other variable, then the two variables are called statistically independent. For example, a researcher may wish to study if the distributions of student majors in three physics classes are the same. For convenience, she divides student majors into three categories "physical sciences," "life sciences" and "engineering." If the distributions of student majors are the same for each of the three classes, then she can say that the two categorical variables—"student major" and "class"—are statistically independent. Otherwise, there is a statistical dependence between the two. Since this test is essentially used to detect an association between two categorical variables, it is also called the chi-square test of association or chi-square test of homogeneity. The null hypothesis of this test (H_0) is that the two variables are statistically independent, and the alternative hypothesis (H_a) is that they are statistically dependent.

The chi-square test is a nonparametric test; it applies to categorical variables and does not assume normality in the data. However, it still

assumes that the samples are randomly selected from, and hence are representative of, the populations. Also, the number of occurrences for each possible combination of groups in the two variables (the counts in each cell of Figure 5) must be greater than 5. In case of smaller counts, one may use Fisher's exact test[55] as an alternative.

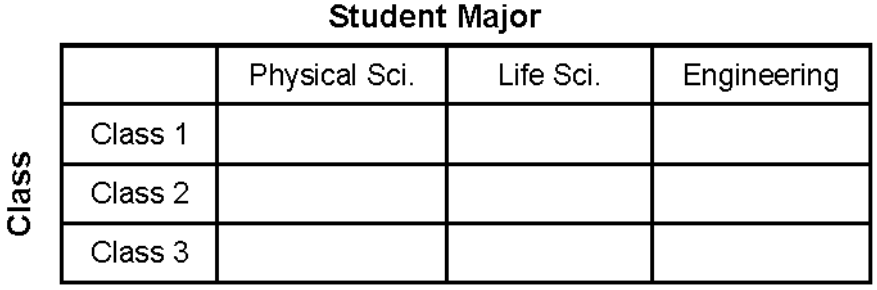

Figure 5: Chi-square Test of Association between Major and Class

A chi-square test produces a statistic that follows a chi-square (χ^2) probability distribution. As with the *t*-test, the precise shape of the chi-square distribution depends on the degree of freedom. Suppose that there are g_1 groups in the first categorical variable and g_2 groups in the second variable. Then the degree of freedom for chi-square test is $df = (g_1 - 1) \times (g_2 - 1)$.

In PER, the chi-square test is a common technique for making comparisons of distributions among different groups of students. Rosengrant, Van Heuvelen, and Etkina[56] investigated student use of free-body diagrams for solving physics problems. They categorized the quality of student-generated free-body diagrams into four levels: 0—no diagram, 1—inadequate diagram, 2—diagram needs improvement, and 3—adequate diagram. They also divided student responses to the physics problems into two groups: correct and incorrect responses. A chi-square test was conducted using diagram quality and response correctness as two categorical variables. The authors found that the distributions of correct and incorrect responses were statistically different for the four levels of diagram quality, and that those who drew adequate free-body diagrams were more likely to solve the problems correctly.

When comparing distributions, it is possible that a researcher may have to deal with more than two categorical variables and wish to test the associations *inter alia*. In such a case, a multi-way frequency analysis is an

appropriate choice. The core ideas of this analysis are similar to those of a chi-square test but contain more complex levels of association. Due to the introductory nature of this article, we only give this technique a passing note without further discussion. Interested readers can refer to Tabachnick and Fidell[57] for more details.

3.3 Interpret analysis results in experimental/quasi-experimental studies

Evaluating the significance of a test is a key component for data interpretation of inferential statistics in PER experimental/quasi-experimental studies. As discussed before, each test generates a statistic and an associated p value. By comparing this p value with a predetermined level of significance α, a researcher can make a decision to either accept or reject the null hypothesis. Typically, the conventional value for α is 0.05, but in case of high-stake situations, a researcher may heighten the standards by making the value smaller. Nevertheless, the opposite rarely happens. Unless there are sufficient reasons to warrant such a decision, one should refrain from loosening the cut-off significance level.

If a researcher detects a significant difference, she/he may wish to further identify where the difference lies. This is not a problem for two-sample comparisons but can be effortful if one has multiple groups. A recommended approach, which is not discussed in the above subsections, is to perform *ad hoc* analyses by making pair-wise comparisons. When doing so, one should use a more stringent value for the significance level α. Consider a case in which there are 7 groups. Among them there are $7\times(7-1)/2=21$ pairs. This means there need to be 21 pair-wise comparisons. If a 5% error probability is allocated to each pair as before, then there will be a total error of 105%, which is not sensible. To overcome this problem, a common practice is to use the conventional value 0.05 divided by the number of possible pairs which can simply be determined by calculating the 2-combination of all groups (i.e., the Bonferroni adjustment).[8] As such, there can still be a total of 5% error in the case of multiple pair-wise comparisons. For example, in the case of 7 groups, there are 21 possible pairs. So, the new significance level α' for pair-wise comparison is $0.5/21 = 0.0024$.

As mentioned earlier, for certain analyses such as two-way ANOVA and ANCOVA, one should examine the significance levels for interactions first before interpreting main effects. One common oversight in this type of analysis is that a researcher may overlook the existence of significant

interactions and jump directly to a therefore unwarranted broad-brush comparison for main effects. It is therefore crucial that a researcher thoroughly document and examine proper statistics before drawing conclusions and making inferences. It is also important to bear in mind that the significance levels generated from any analysis are always associated with specific statistics and degrees of freedom. In order to allow proper interpretations, one should keep a complete note of all relevant information (including statistics, degrees of freedom, and *p* values for both main effects and interactions) in research manuscripts either for his/her own sake or for replication studies by other researchers.

It is worth noting that however powerful a statistical analysis may be, it is after all just a tool that can inform one of what a result is but cannot tell why the result is such. It is the researcher's job to make credible inferences for the reasons that underlie the results and connect them back to the original theoretical framework. To achieve this, a researcher must begin with a series of well thought-out goals and designs. It is important to remember that there is no panacea in statistics that can salvage a poorly designed quantitative study.

4. Measurement Instrument Development and Validation in Evaluation Studies

Empirical PER studies sometimes require a researcher to conduct a systematic evaluation of a program or a curriculum by developing appropriate measurement instruments that serve its specific learning goals. In this type of study—often referred to as an *evaluation study*—a bulk of work involves using quantitative methods for developing measurement instruments to ensure that they are valid and reliable.[g] Some frequently used methods that can serve this purpose include classical test theory,[58-60] Rasch measurement,[14-16, 61] and item response theory.[62, 63] In a previous volume of *Getting Started in PER*, Engelhardt[18] provided a thorough discussion on using Classical Test Theory (CTT) to develop multiple-choice measurement instruments. In this article, we focus specifically on Rasch measurement as an alternative method for developing and validating measurement instruments. Given the limited space, we only

[g] Since some key concepts regarding validity and reliability have been discussed by Engelhardt in the previous volume of *Getting Started in PER*, we avoid repeating and encourage readers to peruse that article for more details.

introduce Rasch measurement as a special case of Item Response Theory (IRT) and leave IRT as a separate topic for future discussion.

In what follows, we first provide a brief overview of what Rasch measurement is and what it can do. To better relate to the previous volumes, we discuss Rasch measurement in comparison with CTT to highlight the similarities and differences between them. We then introduce some key issues in practical applications of Rasch measurement. As before, we discuss the following three aspects: data collection, data analysis, and results interpretation. Here we highlight key issues about Rasch measurement and where possible refer interested readers to other resources for more details.

4.1 Rasch measurement theory *vs.* classical test theory

Rasch measurement is based on a model initially developed by Danish mathematician George Rasch; it is a probabilistic model that utilizes logistic regression to calibrate individual questions in an assessment. As with CTT, the purpose of Rasch measurement is to help a researcher examine psychometric properties of an assessment, so that problematic questions can be identified for revision. Although there are major differences between Rasch measurement and classical test theory as discussed in the following, both can be used to establish validity evidence for a measurement instrument (or informally an assessment) through analyzing the interrelations between the questions therein. For instance, in CTT, individual questions are assumed to be parallel, which means that they should measure the same construct in order for the entire test to be reliable. Ideally, the correlations of individual questions with total test scores should be high (see Engelhardt[18] for more details). Researchers who use CTT to evaluate an assessment therefore strive to seek such evidence. Similarly, in Rasch measurement individual questions are analyzed under the unidimensionality assumption; simply put, all questions should measure the same underlying construct in order for the data to fit the model. Researchers look for fit statistics of individual questions to examine whether or not this unidimensionality requirement is sufficiently met.

However, the two theories differ in some fundamental aspects. First, the conceptual foundations of the two theories are considerably different. CTT assumes that a person's score on an item is a manifestation of a true score that is inherent to the person.[59, 60] Under this conceptual framework, a

measured score (x) can be expressed as a sum of two components: a true score (T) and an error (e):

$$x = T + e$$

CTT treats the error component (e) as a random effect, which on average cancels out if repeated measurements are taken. This is why CTT requires that all questions in an assessment should be parallel. On the other hand, the Rasch measurement theory does not assume that a measured score is a simple linear approximation of a true score inherent to a person. Instead, it conceptualizes the measured score as a probabilistic result of the interactions between a person and an item (question).[61] Presumably, each person possesses some ability that an assessment intends to measure, and items at various difficulty levels are designed to measure the same ability. In Rasch measurement, the probability (Pr) of an observed score (x) is formulated as a function of both person ability (β_n) and item difficulty (δ_i):

$$\Pr(x = 1|\beta_n, \delta_i) = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}$$

The goal of Rasch measurement is to estimate the person ability (β_n) and item difficulty (δ_i) that underlie the observed raw data (x).

Another important difference between CTT and Rasch measurement lies in the nature of the data used in analyses. CTT uses raw data for analysis to generate a list of psychometric properties, including item difficulty, discrimination, Cronbach's alpha, and Ferguson's delta.[64] As mentioned before, raw data collected from multiple-choice tests are at the ordinal level, and therefore performing certain mathematical operations with the raw data may cause problems. For tests containing rating-scale questions or partial-credit questions, this issue becomes even more prominent. Consider a survey of 10 rating questions, each asking students to express their agreement by selecting a response from five choices: strongly disagree, disagree, neutral, agree, and strongly agree. A score is assigned to each choice, ranging from –2 (strongly disagree) to +2 (strongly agree). A traditional approach to estimating a student's performance on the survey is to take the sum (or the average) of the scores on the individual questions. In order for this approach to be valid, one has to assume that the locations of the two extreme choices on a rating scale are the same for each question, and the distance between any two consecutive choices is a constant.[14, 61] As one may realize, these assumptions are virtually impossible to fulfill. Therefore, the traditional approach by summing and/or averaging individual scores for ordinal data is problematic.

Rasch measurement, on the other hand, transforms ordinal-level raw data into two separate sets of interval-scale data through iterative logistic model fitting. One transformed data set is person ability estimates (β_n), and the other is item difficulty estimates (δ_i). These two data sets share the same scale called the *Logit* scale, as it can be derived from taking the logarithm of the ratio between the two probabilities of Pr ($x = 1$) and Pr ($x = 0$).[61] Since both are at the interval level, it is legitimate to make comparisons among persons, items, and between persons and items by using various mathematical operations. For instance, if Rasch analysis generates four person ability estimates $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = 3$, and $\beta_4 = 4$, then we can claim that the difference between 3 and 1 has exactly the same meaning as the difference between 4 and 2, and that it is twice the difference between 4 and 3. Likewise, given a set of four item estimates $\delta_1 = 1$, $\delta_2 = 2$, $\delta_3 = 3$, and $\delta_4 = 4$, we can make similar claims about items in terms of their difficulty levels. Moreover, we can make direct comparisons between person ability (β_n) and item difficulty (δ_i) to find out the probability of a person's correct performance ($x = 1$) on a question.

Perhaps one of the most important features of Rasch measurement that distinguishes it from CTT is the invariant nature of person and item measures. In CTT, all psychometric properties are sample-dependent. For instance, the estimate of an item difficulty level depends on to whom this item is administered. If a group of strong students answer this item, then the results may show this item being too easy. Conversely, if a group of weak students answer this question, then the results may indicate it is a difficult item. Similarly, the estimate of a student score depends on what items are used in the test. Using a set of difficult questions may result in underestimation of student true performance; whereas using a set of easy questions may result in overestimation of student true performance. However, in Rasch measurement person ability and item difficulty estimates are sample independent.[61] This is because Rasch measurement no longer relies on raw data; instead it uses the probabilistic framework to uncover person ability and item difficulty that underlie the raw data. In this case, selecting students of different abilities does not affect the estimates of item difficulty, and similarly pooling different items does not affect the estimates of person abilities. This sample independent feature is also referred to as the invariant property of Rasch measurement.

With the general background about the Rasch measurement theory in mind, we now proceed with a brief discussion on some practical aspects that relate to data collection and data analysis/interpretation in using Rasch measurement for evaluation studies.

4.2 Prepare for data collection in Rasch measurement

As mentioned before, Rasch measures are sample invariant. This puts less restriction on which samples a researcher must include for analysis. That said, in real practice it is highly recommended that a researcher carefully choose an appropriate student audience for testing and where possible include students of various abilities into a sample to increase precision. In Rasch analysis, the precision of estimates for person ability and item difficulty depends largely on the extent to which their distributions match each other. Generally speaking, the more closely they match, the more precise the estimates are.

Another important consideration for data collection is sample size. In fact, there is no consensus on what minimum sample size one should use for Rasch analysis. Some researchers argue that a sample as small as 50 should suffice.[65] Some argue that a minimum sample of 200 and a set of 20 items are necessary for precise estimation of both person ability and item difficulty.[66] That said, in most situations a sample of 100 students can be considered adequate for a test of 10-20 items. If it is a high-stakes test, a minimum of 250 students with 20 items may be needed. In the case of rating-scale surveys or partial-credit tests, extra caution is needed, because more parameters, other than the item difficulty levels, are estimated in these types of assessments. Interested readers can refer to reference 65 for more details.

Also important is selecting an appropriate model for Rasch analysis. Depending on the format of the assessment questions, one can choose to use a dichotomous Rasch model, rating scale model, or partial credit model.[61] Due to the introductory nature of this article, we omit further discussions on these models. Readers can consult reference 61 for more details.

4.3 Analyze and interpret data in Rasch measurement

Fit statistics generated by Rasch measurements can help a researcher identify whether or not individual items fit the unidimensionality assumption and, if not, which items may need to be revised. There are two sets of fit statistics in Rasch analysis; they are item fit statistics and person fit statistics. Item fit statistics allow a researcher to examine whether or not the items in an assessment fit under the unidimensionality assumption and if not which items may be problematic. Person fit statistics can reveal

which individuals may demonstrate inconsistent response behaviors; for example, a high-ability student may incorrectly answer easy questions, or a low-ability student correctly answers difficult questions.

Both fit statistics can be reported in two forms: an un-standardized form called mean square residual (MNSQ) and a standardized form of the *t* statistic. A mean square residual (MNSQ) is the average of squared differences between model expected values and actual observation values. When combined with sample size, MNSQ can be transformed into a *t* statistic. For each fit statistic (MNSQ and *t*), Rasch analysis reports two aspects of measure, namely infit and outfit. Infit statistics (infit MNSQ and infit *t*) put more weight on data points that have a close match between person ability and item difficulty, thus minimizing the effect of outliers; whereas outfit statistics (outfit MNSQ and outfit *t*) give equal weight to all data and hence are less sensitive to outliers.

Typically, MNSQ values within the range of [0.7, 1.3] and *t* statistics within the range of [–2, +2] are considered as acceptable. If fit statistics of an item exceed the upper bounds of the above ranges, then this item may not fit under the unidimensionality condition and can be even haphazard to the overall construct of the assessment (because of too much variation). On the other hand, if the fit statistics of an item fall short of the lower bounds, then the item does not contribute to the assessment (because of too little variation in the data). However, these ranges should not be taken rigidly, as there is no one-size-fits-all rule. Interested readers can refer to reference 67 for more details.

Finally, it is important to remember that as with any other statistical techniques, Rasch modeling can only inform us of problematic items but can never tell us why they are problematic. Again the burden rests with researchers to find out why these items fail to fit the overall construct of an assessment and then revise them by using a well-articulated framework.

5. Summary and Suggestions for Further Reading

We have provided in this article a brief introduction to commonly used PER quantitative methods to help readers get started in this research area. We first discussed what PER quantitative studies are, how different they are from quantitative studies in other fields of physics, and what their typical research questions may look like. Given the different nature of

three types of PER quantitative studies, namely survey research, experimental/quasi-experimental studies, and evaluation studies, we reviewed three major groups of quantitative techniques—descriptive statistics, inferential statistics and Rasch models—that match the goals of the three types of research respectively. Table 1 shows a summary of these quantitative techniques, including their purposes, necessary data set, typical statistical information, and key requirements.

Readers who are interested in learning more about these techniques and others are encouraged to seek additional resources relevant to the topics discussed in this article. The following are some good materials that may help readers gain more in-depth knowledge in this field.

A. Agresti and B. Finlay, *Statistical methods for the social sciences* (Pearson Education, Upper Saddle River, NJ, 2009).

T. G. Bond and C. M. Fox, *Applying the Rasch model: fundamental measurement in the human sciences* (Routledge, Taylor & Francis, N.Y., 2007).

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

M. Hollander and D. A. Wolfe, *Nonparametric statistical methods* (John Wiley & Sons, New York, 1999).

S. W. Huck, *Reading statistics and research* (*6th edition*) (Harper Collins College Publishers, New York, 2011)

X. Liu, *Using and developing measurement instruments in science education: A Rasch modeling approach* (Information Age Publishing, 2010).

B. Tabachnick and L. S. Fidell, *Using multivariate statistics (6th edition)* (Pearson, Boston, 2012).

Table 1 A Summary of Descriptive Statistics, Inferential Statistics and Measurement Instrument Development/validation Techniques

Research Type	Quantitative Method	Sample	Variables	Statistics	Degrees of Freedom
Survey Research	Descriptive statistics	1	1 variable	μ, S, σ, etc.	n/a
	Correlation	1	2 variables	r (interval/ratio) ρ (ordinal) φ (nominal dichotomous)	n/a
Experimental and Quasi-experimental Studies	Repeated *t*-test	1	2 continuous IV	t	$N-1$
	McNemar test	1	2 dichotomous IV	χ^2	1
	Two-sample *t*-test	2	2 continuous IV	t	$N-2$
	One-way ANOVA	2 or more	1 categorical IV; 1 continuous DV	F	$g-1$, $N-g$
	Two-way or multi-way ANOVA	2 or more	2 or more categorical IV; 1 continuous DV	F	g_1-1; g_2-1; $N-g_1\times g_2$
	One-way ANCOVA	2 or more	1 categorical IV; 1 continuous IV; 1 continuous DV	F	$g-1$; $N-g-1$
	Two-way or multi-way ANCOVA	2 or more	2 or more categorical IV; 1 continuous IV; 1 continuous DV	F	g_1-1; g_2-1; $N-g_1\times g_2-1$
	Chi-squared test	2 or more	2 categorical IV	χ^2	$(g_1-1)\times(g_2-1)$
	Multi-way analysis	2 or more	3 or more categorical IV	G^2	$(g_1-1)\times(g_2-1)$ $\times(g_3-1)\ldots$
Evaluation Studies	Rasch measurement	1 or more	n/a	Fit statistics (MNSQ and t)	n/a
	Classical Test Theory	1 or more	n/a	Item & test statistics (see Ref 18)	n/a

Acknowledgement

The authors wish to thank Tom Foster and the other two anonymous reviewers for their insightful comments on the manuscript. We also thank the support of our colleagues in the School of Teaching and Learning at Ohio State University and in the Department of Learning and Instruction at the State University of New York, Buffalo.

References

1. J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (Sage, Thousands Oaks, CA, 2009).
2. B. Johnson and L. Christensen, *Eduational Research: Quantitative, Qualitative, and Mixed Approaches* (Sage Publications, Inc., Thousand Oaks, CA, 2007).
3. H. R. Bernard, *Social Research Methods: Qualitative and Quantitative Approaches* (Sage Publications, Inc., Thousand Oaks, CA, 2000).
4. V. K. Otero and D. B. Harlow, "Getting Started in Qualitative Physics Education Research," in C. Henderson and K. A. Harper (eds), *Reviews in PER: Getting Started in PER* (Am. Assoc. of Phys. Teach., College Park, MD, 2009).
5. G. R. Taylor and M. Trumbull, "Major Similarities and Differences Between two Paradigms," in G. R. Taylor (ed), *Integrating Quantitative and Qualitative Methods in Research*, pp. 235-248 (University Press of America, Lanham, Md., 2005).
6. T. D. Cook and C. S. Reichardt, *Qualitative and Quantitative Methods in Evaluation Research* (Sage Publications, Inc., Beverly Hills, CA, 1985).
7. R. Beichner, "An Introduction to Physics Education Research," in C. Henderson and K. A. Harper (eds), *Reviews in PER: Getting Started in PER* (Am. Assoc. of Phys. Teach., College Park, MD, 2009).
8. A. Agresti and B. Finlay, *Statistical Methods for the Social Sciences* (Pearson Education, Upper Saddle River, NJ, 2009).
9. W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein and C. E. Wieman, "New Instrument for Measuring Student Beliefs About Physics and Learning Physics: The Colorado Learning Attitudes about Science Survey," *Phys. Rev. ST Phys. Educ. Res.* **2**(1), 010101 (2006).
10. D. Hestenes, M. Wells and G. Swackhamer, "Force Concept Inventory," *Phys. Teach.* **30**(3), 141-158 (1992).

11. R. K. Thornton and D. R. Sokoloff, "Assessing Student Learning of Newton's Laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula." *Am. J. Phys.* **66**(4), 338-352 (1998).
12. D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke and A. Van Heuvelen, "Surveying Students' Conceptual Knowledge of Electricity and Magnetism," *Am. J. Phys.* **69**(S1), S12-S23 (2001).
13. L. Ding, R. Chabay, B. Sherwood and R. Beichner, "Evaluating an Electricity and Magnetism Assessment Tool: Brief Electricity and Magnetism Assessment," *Phys. Rev. ST Phys. Educ. Res.* **2**(1), 010105 (2006).
14. W. J. Boone, Townsend, J. S., and Staver, J., "Using Rasch Theory to Guide the Practice of Survey Development and Survey Data Analysis in Science Education and to Inform Science Reform Efforts: An Exemplar Utilizing STEBI Self-efficacy Data," *Science Education* **95**(2), 258-280 (2011).
15. W. J. Boone and K. Scantlebury, "The Role of Rasch Analysis When Conducting Science Education Research Utilizing Multiple-choice Tests," *Science Education* **90**(2), 253-269 (2005).
16. X. Liu, *Using And Developing Measurement Instruments In Science Education: A Rasch Modeling Approach* (Information Age Publishing, 2010).
17. D. J. Sheskin, *Handbook Of Parametric And Nonparametric Statistical Procedures* (Chapman & Hall/CRC, Boca Raton, FL, 2004).
18. P. V. Engelhardt, "An Introduction To Classical Test Theory As Applied To Conceptual Multiple-Choice Tests," in C. Henderson and K. A. Harper (eds), *Reviews in PER: Getting Started in PER*. (Am. Assoc. of Phys. Teach., College Park, MD, 2009).
19. R. E. Kirk, "Experimental Design," in R. E. Millsap and A. Maydeu-Olivares (eds), *Quantitative Methods In Psychology*. pp. 23-45 (Sage Publications, Inc., Thousand Oaks, CA, 2009).
20. C. S. Reichardt, "Quasi-Experimental Design," in R. E. Millsap and A. Maydeu-Olivares (eds), *Quantitative Methods in Psychology*, pp. 46-71 (Sage Publications, Inc., Thousand Oaks, CA, 2009).
21. E. F. Redish, R. N. Steinberg and J. M. Saul, "Student Expectations In Introductory Physics," *Am. J. Phys.* **66**(3), 212-224 (1998).
22. P. H. Rossi and J. D. Wright, Design And Implementation Of Evaluation Models -- Evaluation Research: An Assessment," in D. C. Miller and N. J. Salkind (eds), *Handbook Of Research Design And Social Measurement* (Sage Publications, Inc., Thousand Oaks, CA, 2002).

23. G. Rasch, *Probabilistic Models For Some Intelligence And Attainment Tests* (Reprinted by University of Chicago Press, Chicago, 1960).
24. B. D. Wright, "Sample-Free Test Calibration And Person Measurement," in *Proceedings Of The 1967 Invitational Conference On Testing,* pp. 85-101 (Educational Testing Service, Princeton, NJ, 1968).
25. B. D. Wright, "Solving Measurement Problems With The Rasch Model," *Journal of Educational Measurement* **14**(2), 97-116 (1977).
26. G. H. Fischer and I. W. Molennar, *Rasch Models: Foundations, Recent Developments, And Applications* (Springer-Verlag, New York, 1995).
27. J. R. Fraenkel, N. E. Wallen and H. Hyun, *How To Design And Evaluate Research In Education* (McGraw-Hill Higher Education, New York, 2012).
28. A. Agesti, *An Introduction To Categorical Data Analysis* (John Wiley & Sons, Inc., Hoboken, NJ, 2007).
29. R. Thornton, D. Kuhl, K. Cummings and J. Marx, "Comparing The Force And Motion Conceptual Evaluation And The Force Concept Inventory," *Phys. Rev. ST Phys. Educ. Res.* **5**(1), 010105 010101-010108 (2009).
30. A. Agresti, *Analysis Of Ordinal Categorical Data* (John Wiley & Sons, Hoboken, NJ, 2010).
31. J. Miles and M. Shevlin, *Applying Regression & Correlation: A Guide For Students And Researchers* (Sage Publications, Inc., Thousand Oaks, CA, 2004).
32. P. Heller and D. Huffman, "Interpreting The Force Concept Inventory: A Reply To Hestenes And Halloun," *Phys. Teach.* **33**(8), 503-511 (1995).
33. D. Huffman and P. Heller, "What Does The Force Concept Inventory Actually Measure?" *Phys. Teach.* **33**(3), 138-143 (1995).
34. G. Kortemeyer, "Correlations Between Student Discussion Behavior, Attitudes, and Learning," *Phys. Rev. ST Phys. Educ. Res.* **3**(1), 010101 (2007).
35. C. Henderson, M. H. Dancy and M. Niewiadomska-Bugaj, "Use Of Research-Based Instructional Strategies In Introductory Physics: Where Do Faculty Leave The Innovation-Decision Process?" *Phys. Rev. ST Phys. Educ. Res.* **8**(2), 020104 020101-020112 (2012).
36. W. R. Shadish, T. D. Cook and D. T. Campbell, *Experimental And Quasi-Experimental Designs For Generalized Causal Inference* (Wadsworth, Belmont, CA 2002).
37. J. Cohen, *Statistical Power Analysis For The Behavioral Sciences* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).

38. J. Cohen, "A Power Primer," *Psychological Bulletin* **112**(1), 155-159 (1992).
39. http://www.danielsoper.com/statcalc3/calc.aspx?id=1
40. M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods* (John Wiley & Sons, New York, 1999).
41. E. L. Lehmann and H. J. M. D'Abrera, *Nonparametrics: Statistical Methods Based On Ranks* (Springer, New York, 2006).
42. S. J. Pollock, "Longitudinal Study Of Student Conceptual Understanding In Electricity And Magnetism," *Phys. Rev. ST Phys. Educ. Res.* **5**(2), 020110 020111-020118 (2009).
43. R. R. Hake, "Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey Of Mechanics Test Data For Introductory Physics Courses," *Am. J. Phys.* **66**(1), 64-74 (1998).
44. J. Stewart, H. Griffin and G. Stewart, "Context Sensitivity In The Force Concept Inventory," *Phys. Rev. ST Phys. Educ. Res.* **3**(1), 010102 010101-010106 (2007).
45. J. Day and D. Bonn, "Development Of The Concise Data Processing Assessment," *Phys. Rev. ST Phys. Educ. Res.* **7**(1), 010114 010111-010114 (2011).
46. T. I. Smith and M. Wittmann, "Comparing Three Methods For Teaching Newton's Third Law," *Phys. Rev. ST Phys. Educ. Res.* **3**(2), 020105 020101-020108 (2007).
47. L. C. McDermott, P. S. Shaffer and University of Washington Physics Education Group, *Tutorials In Introductory Physics* (Prentice Hall, Upper Saddle River, NJ, 2002).
48. D. Hammer, "Student Resources For Learning Introductory Physics ," *Am. J. Phys.* **68**(S1), S52-S59 (2000).
49. A. Elby, "Helping Physics Students Learn How To Learn," *Am. J. Phys.* **69**(S1), S54-S64 (2001).
50. L. Ding, N. W. Reay, A. Lee and L. Bao, "Effects Of Testing Conditions On Conceptual Survey Results," *Phys. Rev. ST Phys. Educ. Res.* **4**(010112), 1-6 (2008).
51. K. K. Cheng, B. A. Thacker, R. L. Cardenas and C. Crouch, "Using An Online Homework System Enhances Students' Learning Of Physics Concepts In An Introductory Physics Course," *Am. J. Phys.* **72**(11), 1447-1453 (2004).
52. A. Lawson, "Rank Analysis Of Covariance: Alternative Approaches," *Journal of the Royal Statistical Society. Series D (The Statistician)* **32**(3), 331-337 (1983).
53. S. F. Olejnik and J. Algina, "A Review Of Nonparametric Alternatives To Analysis Of Covariance," in *Annual Meeting of the American*

Educational Research Association, 68th, pp. 46 (New Orleans, LA, 1984).
54. E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez and P. Pamela, "Toward Equity Through Participation In Modeling Instruction In Introductory University Physics," *Phys. Rev. ST Phys. Educ. Res.* **6**(1), 010106 010101-010112 (2010).
55. B. S. Everitt, *The Analysis Of Contingency Tables* (Wiley, New York, 1992).
56. D. Rosengrant, A. Van Heuvelen and E. Etkina, "Do Students Use And Understand Free-Body Diagrams?" *Phys. Rev. ST Phys. Educ. Res.* **5**(1), 010108 010101-010113 (2009).
57. B. Tabachnick and L. S. Fidell, *Using Multivariate Statistics* (Pearson, Boston, 2001).
58. R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, D. C., 1980).
59. E. E. Ghiselli, J. P. Campbell and S. Zedeck, *Measurement Theory For The Behavioral Sciences* (W. H. Freeman, San Francisco, 1981).
60. P. Kline, *A Handbook Of Test Construction: Introduction To Psychometric Design* (Methuen, New York, NY, 1986).
61. T. G. Bond and C. M. Fox, *Applying The Rasch Model: Fundamental Measurement In The Human Sciences* (Routledge, Taylor & Francis, N.Y., 2007).
62. G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi and V. McCauley, "Testing the Test: Item Response Curves and Test Quality," *Am. J. Phys.* **74**(5), 449-453 (2006).
63. L. Ding and R. Beichner, "Approaches to Data Analysis of Multiple-choice Questions," *Phys. Rev. ST Phys. Educ. Res.* **5**(020103), 1-17 (2009).
64. P. V. Engelhardt and R. J. Beichner, "Students' Understanding of Direct Current Resistive Electrical Circuits," *Am. J. Phys.* **72**(1), 98-115 (2004).
65. J. M. Linacre, "Sample Size and Item Calibration Stability." *Rasch Measurement* **7**(4), 328 (1994).
66. D. L. Streiner and G. R. Norman, *Health Measurement Scales: a Practical Guide to Their Development and Use* (Oxford University Press, Oxford, 1995).