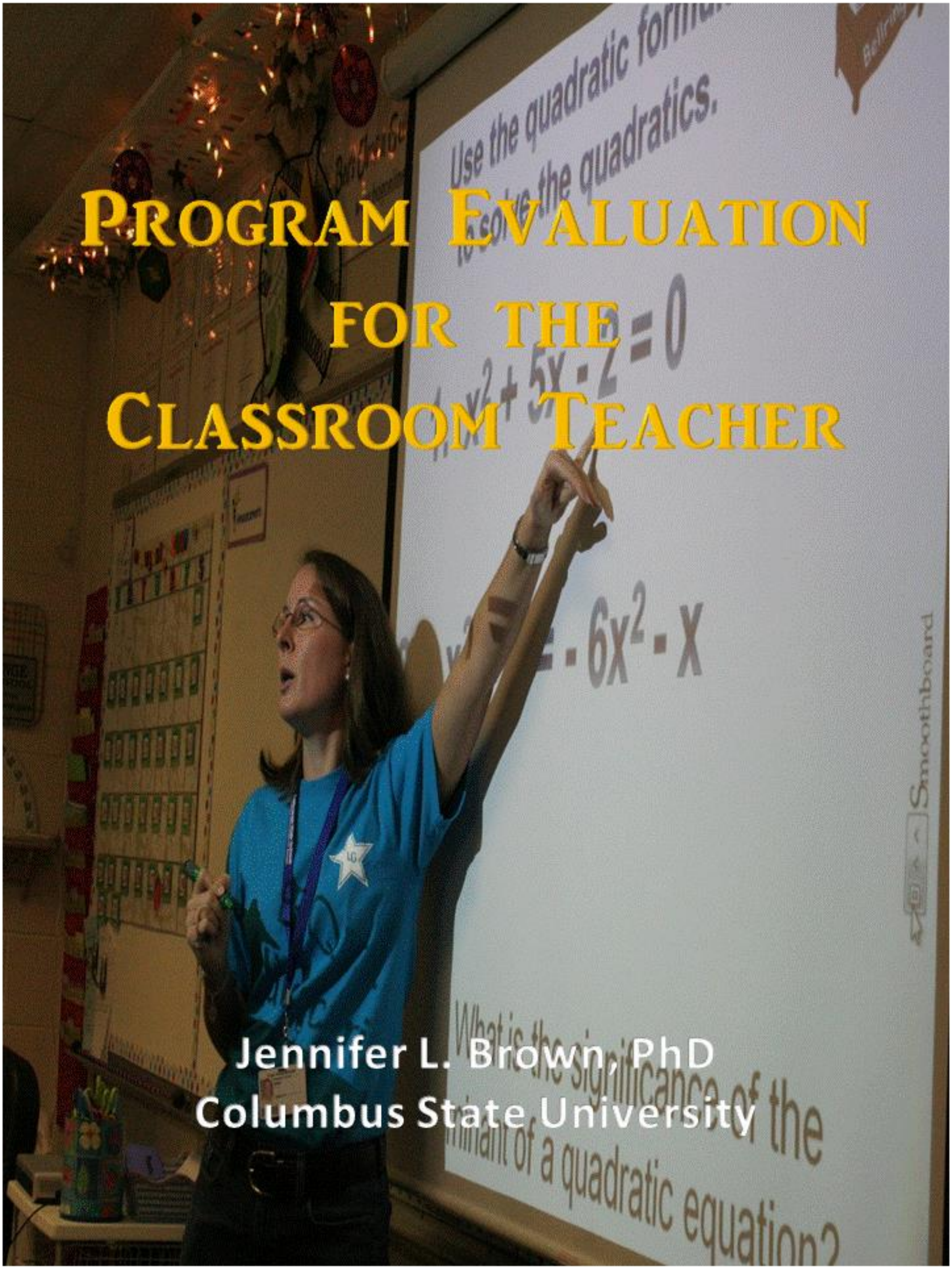
2016

Program Evaluation For The Classroom Teacher

Jennifer L. Brown
Columbus State University

Program Evaluation for the Classroom Teacher



Jennifer L. Brown, PhD
Columbus State University

Peer Reviewed by

Mike Edmondson, PhD
Columbus State University

Supplemental materials, including the demonstration and guided practice datasets, are available at http://www.bugforteachers.com/prog_eval.html.

About the Author

Dr. Jennifer L. Brown began her career as a secondary special education teacher after graduating with her B.S.Ed. in Mental Retardation from the University of West Georgia in 1998. In 2002, she earned her M.Ed. in Interrelated Special Education and National Board Certification. Three years later, Dr. Brown graduated with her Ed.S. in Special Education: Curriculum and Instruction. She took an educational sabbatical from 2006 until 2008 to work on her Ph.D. in Educational Psychology at Auburn University. At Auburn, she worked as a graduate research assistant for Drs. Gerald and Glennelle Halpin during which she gained extensive experience with program evaluation. Her responsibilities included evaluating three large-scale projects (i.e., two externally-funded grants through Alabama Department of Public Health and Alabama United Prevention Services and one internally-funded grant through Auburn University's College of Engineering). After graduating in 2008, Dr. Brown returned to the classroom as a secondary math teacher until she joined the faculty at Columbus State University in 2011. Currently, she is an Associate Professor of Educational Foundations within the Department of Teacher Education.

TABLE OF CONTENTS

CHAPTER 1

OVERVIEW OF PROGRAM EVALUATION

Since the passage of No Child Left Behind (NCLB) in 2001, there has been a vast shift to emphasis on accountability and data driven decisions. It is unfortunate, but, prior to NCLB, numerous educational decisions were made without regard to concrete data or impact on student learning. In the K-12 setting, there are numerous programs for increasing mathematical proficiency, writing across the curriculum, teaching character education, and boosting standardized test scores. Countless hours are devoted to these programs through teacher training and student instructional time; however, little to no effort is given to the evaluation of these programs. Is it worth the loss of instructional time to teach students how to diagram a sentence if the writing examination scores are not improving? Typically, this type of evaluation question is not addressed at the K-12 level.

The purpose of this book was to provide a basic foundation in educational research and illustrate how educational research aligns with program evaluation. As an educational psychologist who was trained in educational research and program evaluation, program evaluation offers numerous practical benefits for the classroom teacher. The content of this book is meant to show you the usefulness and practicality based on my experiences as a program evaluator and classroom teacher.

What is Program Evaluation?

Program evaluation is the systematic collection of data about the activities and outcomes of a program. After data analysis, decisions can be made about whether or not to continue the program, improve its effectiveness, and/or modify the future program implementation (Patton, 2002). A basic understanding of research methods is required to plan and conduct a program evaluation. Program evaluation is similar to traditional educational research (e.g., quantitative/qualitative research questions and identifying cause and effect relationships), but here are a few differences (Suvedi & Morford, 2003). Table 1 depicts a few of those differences (Mathison, 2008).

Table 1

Differences between Educational Research and Program Evaluation

Criteria	Educational Research	Program Evaluation
Motivation	Advance knowledge	Solve practical problems
Objective	Seek conclusions and credible explanations	Lead to decisions and determines worth or value
Criteria	Degree to which results are without error and generalizable	Degree of accuracy, credibility, feasibility.

There are two purposes for program evaluation: formative and summative. Similar to the terms used with classroom assignment, formative evaluations occur during program implementation in order to improve the process or procedure, and summative evaluations occur after the program has ended in order to evaluate outcomes. **Formative evaluations** are used to determine the quality or effectiveness of a program and to indicate strengths or weaknesses, which provides the program staff with formative feedback. With **summative evaluations**, the purpose is to determine the quality of the program after the program has ended; however, it also serves as a method to make decisions about the future of the program (Suvedi & Morford, 2003). Usually, formative evaluations are conducted by internal evaluators, and summative evaluations are conducted by external evaluators (Fitzpatrick, Sanders, & Worthen, 2004). See Table 2 for the advantages and disadvantages of internal and external evaluators according to Suvedi and Morford (2003).

Table 2

Advantages and Disadvantages of Internal and External Evaluators

Options	Advantages	Disadvantages
Internal Evaluators	• Familiar with the organization • Established credibility within the organization	• Potentially biased • May lack evaluation skills
External Evaluators	• Specialized program evaluation skills • Unbiased	• Lacks knowledge of the organization • Limited access to information and people • Potential for extra expense

Throughout this textbook, I will use the implementation of a secondary mathematics curriculum as an example of a program that needs to be evaluated. This hypothetical secondary mathematics curriculum will have an engineering focus. Each unit across all four courses (i.e., geometry, algebra II, pre-calculus/trigonometry, and advanced placement calculus AB) will have NCTM Standards-based expectations, at least one engineering connections (e.g., chemical, civil, electrical, or mechanical engineering), mathematical concepts involved with the unit topic, instructional goal(s), key terms, any required equipment needed for the unit, and a performance assessment. The performance assessment at the end of each unit will be a cumulating activity for the students to apply the mathematical concepts to the engineering field. The program evaluation proposal for this curriculum is presented in Appendix I.

With the implementation of the mathematics curriculum, a formative evaluation could assess the attitudes and instructional methods of the teachers by monitoring professional development workshops and weekly classroom observations. The midterm benchmark examinations could provide formative evaluation information during the academic year. All of these data sources could provide ongoing feedback about the curriculum implementation process, including strengths and weaknesses. A summative evaluation could include assessment of the students' mathematical proficiency with the final benchmark examinations. Other summative evaluations could include the results of the state's graduation exit examinations and the Advanced Placement Calculus Examination. These assessments evaluate the long-term outcomes of the curriculum implementation and the impact on student learning.

Many evidence-based programs are demonstrated at various professional development venues; however, when implementation occurs, there is uncertainty about whether the program was effective because the program was not evaluated in order to determine effectiveness. When planning for a program evaluation, a series of topics should be addressed prior to program implementation to assess the full impact on student learning. The steps include:

(a) meeting with all stakeholders,

(b) identifying evaluation purpose,

(c) designing the evaluation plan,

(d) collecting the data,

(e) analyzing and interpreting the data,

(f) writing the evaluation report.

Each of these steps will be discussed as you move through this textbook.

Step 1: Meeting With All Stakeholders

To begin, who are stakeholders? **Stakeholders** can be any individual or group that has a "stake" or interest in the outcome of the program evaluation (Suvedi & Morford, 2003). With the secondary mathematics curriculum example, the stakeholders could be students, teachers, administrators, district office personnel, and community leaders. If the evaluation team was external to a school system, the following procedure would be followed. For application purposes, each procedural step will be illustrated with a hypothetical secondary mathematics curriculum, which the evaluation team has been hired to evaluate.

1. Meet with the superintendent of schools and the local school board during a caucus meeting to discuss curriculum implementation and evaluation.
2. Meet with the curriculum director at the local county office to discuss curriculum implementation.
3. Meet with school principal to discuss general school culture and plans for curriculum implementation (e.g., professional development and textbook adoption).
4. Meet with the assistant principals and registrar to discuss scheduling and personnel, which may pertain to curriculum implementation and evaluation.
5. Meet with the secondary mathematics teachers to discuss curriculum implementation and evaluation.
6. After the initial meetings, contact the program developer to obtain a copy of the curriculum and other evaluations.
7. If available, contact persons at other school systems who have implemented the mathematics curriculum to get their perspective and possible program evaluations.
8. Search the literature for studies using the mathematics curriculum or similar curricula.
9. Review the curriculum, program evaluations, and literature. Determine if the curriculum aligns with the state and school system's standards and National Council of Teachers of Mathematics (NCTM) standards.

By following these procedures, the evaluation team can determine the target population, assess the current needs, determine the rationale for the evaluation, clarify intended outcomes, and assess stakeholders' reaction to the intended program (Killion, 2002).

Chapter 2

Identifying the Evaluation Purpose

Step 2: Identifying the Evaluation Purpose

An **evaluation purpose** is similar to the purpose of a research study. What do the stakeholders who are requesting the evaluation want to know? Usually, the stakeholders want to know if the program was effective and achieved its goals and objectives. After identifying the purpose, the questions that need to be answered should be identified. Typically, these questions derive from the goals and objectives of the program. Continuing with the illustrative example, the local school board and superintendent have requested an evaluation of the mathematics curriculum. During the planning phrase, a logic model will be created for the stakeholders by the program evaluators. The **logic model** serves as a blueprint for the program, including the inputs, activities, short-term objectives, and long-term objectives. **Inputs** are any funding sources and/or resources provided to support the program. **Activities** are any services, materials, and/or events associated with the program's implementation. **Short-term objectives** are the immediate impact of the implementation activities, and **long-term objectives** are the enduring impacts of the program (Frechtling, 2002). See Figure 1 for the logic model example using the secondary mathematics curriculum. Notice, the short-term and long-term objectives are clear and measurable.

From the logic model, the evaluation questions can be formulated (Fitzpatrick et al., 2004). Using the curriculum example, to assess the implementation activities, which would a series of formative evaluations, one of the evaluation questions could be "Have professional development sessions, conducted with the implementing teachers, promoted a successful curriculum implementation?" As a summative evaluation, another question to assess one of the long-term outcomes could be "Have Graduation Exit Examination: Mathematics Subtest scores changed in comparison to scores before implementation?"

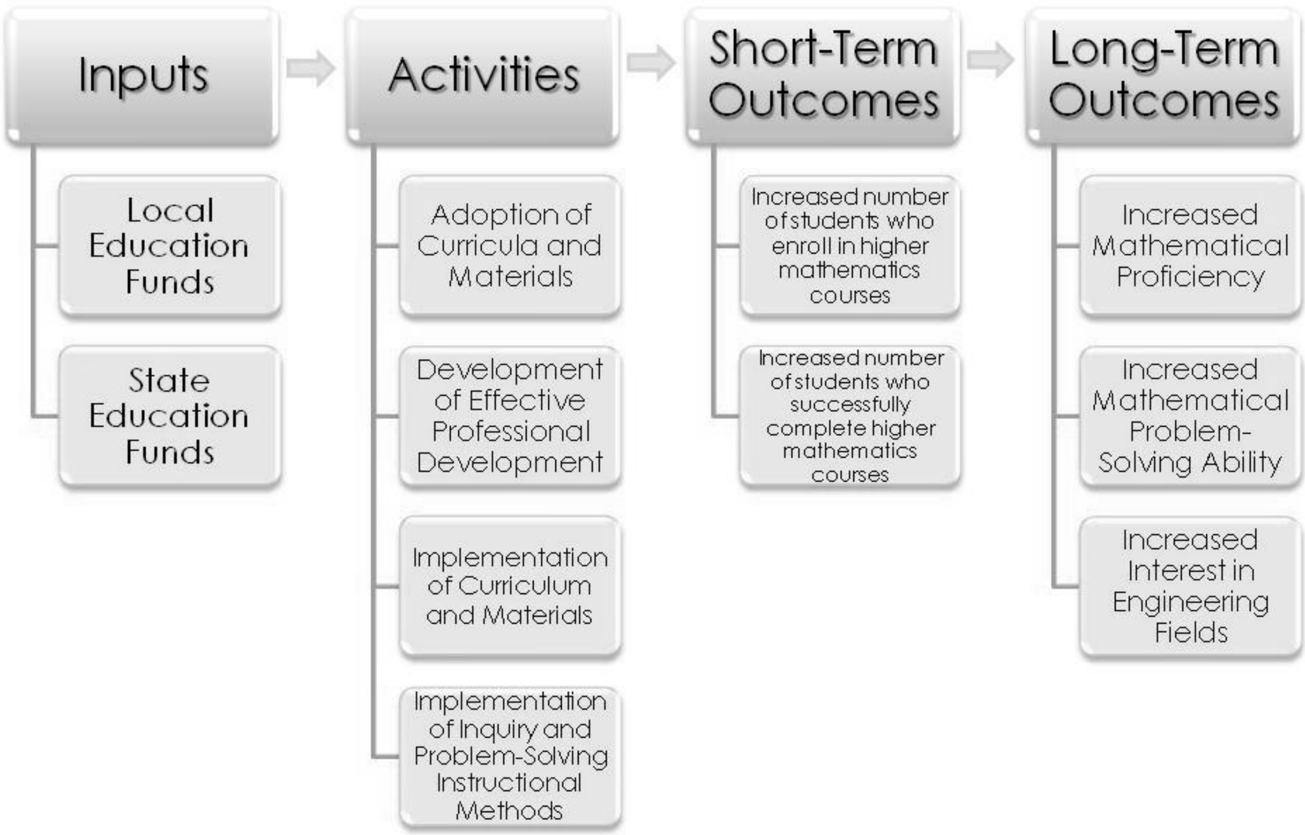


Figure 1. Logic model for the secondary mathematics curriculum implementation.

CHAPTER 3

DESIGNING THE EVALUATION PLAN

Step 3: Designing the Evaluation Plan

An **evaluation plan** is systematic plan that is used to answer your research questions. When planning, you must consider the research design, sampling, program implementation process, and data collection procedures. Depending on the purpose of your program evaluation, there are some questions to consider before designing the evaluation plan (Killion, 2002).

Formative Evaluation

- How well is the program working?
- How is its implementation aligned with the intended plan?
- Does it meet standards of operation?
- Are the components in place as planned?

Summative Evaluation

- Does the program produce results?
- Does it have impact?
- What unintended effects, if any, are occurring?

Design. **Research design** is a strategy for conducting the research or program evaluation in this case. There are various designs, both causal and descriptive, can be considered when designing an evaluation plan. Another design consideration is whether or not to utilize a quantitative or qualitative approach. To determine the appropriate approach, you will need to match the approach to the program's goals and objectives and fit the approach to your audience. For example, if the program's design will utilize predetermined measures for assessment, then a quantitative approach would be best. For example, a longitudinal program evaluation using descriptive research may show trends in the data with the same sample over a period of time. (See Program Evaluation Report Example #4 in the Appendix G.) If the goal of the program evaluation is to elicit participants' experiences, particular with small sample sizes, then a qualitative approach would be best. The qualitative approach may be used to describe and analyze a targeted program, process, or procedure and provide further insight. Figure 2 displays a flowchart of different types of research designs that can be utilized for program evaluation.

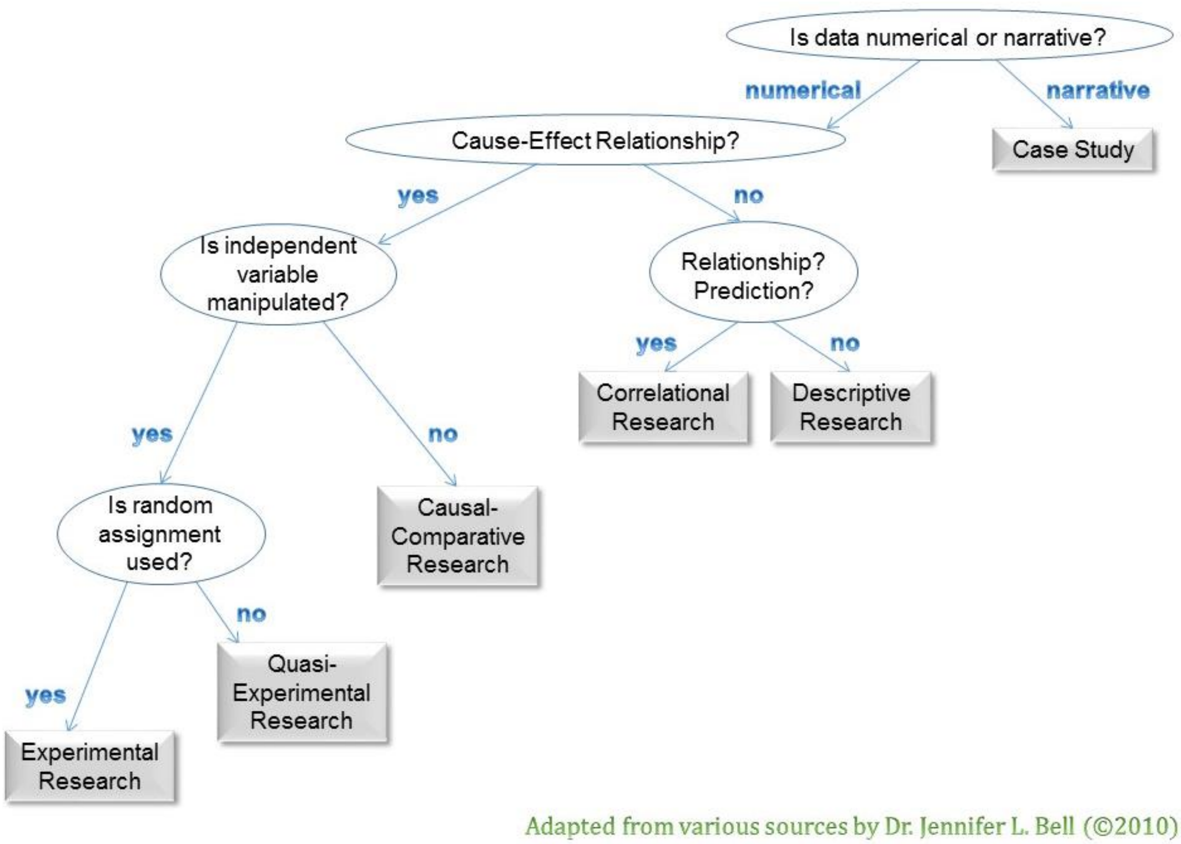


Figure 2. Types of research designs for program evaluation.

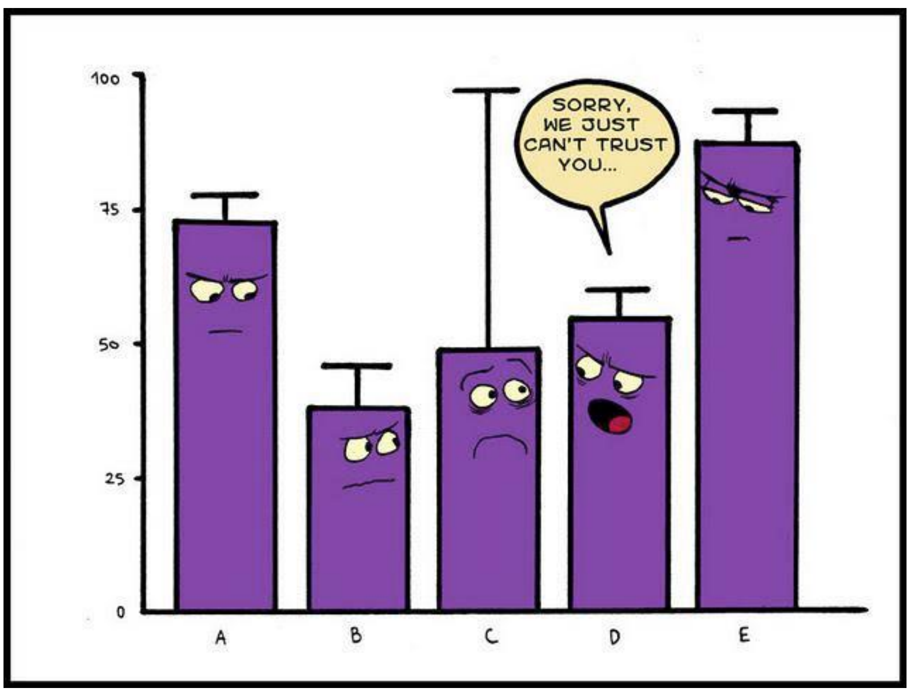
Qualitative Approach

- **Case Study** – This research design will occur when the program evaluator wants an extensive study of a group of individuals. For example, what was the impact of the mentorship program on new teacher mentees? This research question is too multifaceted for a simple quantitative survey. You would want to speak with the mentors and mentees to paint a complete picture of program impact.

Quantitative Approach

- **Descriptive Research** – This research design will answer the question, "How much exists?" For example, what was the average final grade in 9th grade English? You could collect the final grades and calculate the mean. If you would like to examine the effects across time, we refer to those designs as longitudinal, but the results would be reported as descriptives unless you are able to track the same group across time, which is difficult because of attrition. (See Program Evaluation Report Example #4 in Appendix G.)
- **Correlational Research** – This research design will answer the question, "What is the relationship between two variables?" For example, what was the relationship between 10th grade End-Of-Course-Tests in math and final grade in 10th grade math class? Remember, if a relationship exists, then it does not mean causation.
- **Comparing Groups** – The last three research designs, causal-comparative, quasi-experimental, and experimental, involve comparing groups, which allows the program evaluator to determine if one variable caused another variable to change. There are a few distinct differences among the three designs. For **causal comparative research**, the program evaluator will utilize pre-existing groupings. In other words, the conditions of the sample will not be manipulated. For example, using the secondary mathematics curriculum example, what was the effect of the secondary mathematics curriculum on End-Of-Course Tests? Often, evaluation teams will use

pre-existing data to determine if changes occur as a result of an intervention. With the curriculum example, a student sample with similar characteristics will be selected to serve as a comparison group, or control group, with the intervention group (Fitzpatrick et al., 2004). For **quasi-experimental research**, the conditions of the sample will be manipulated. A stakeholder decides which students will be in Group A and what intervention they will receive; however, student placement will not be randomly assigned. The study will occur in the "natural" setting. For example, does Ms. Smith's class perform better using a cognitive strategy for solving word problems compared to Mr. Jones' class? For **experimental research**, a stakeholder will manipulate the conditions and randomly assign students to the groups. For example, did the afterschool tutoring program improve reading levels? Typically, in educational research when comparing groups, causal-comparative and quasi-experimental are the most utilized research designs. They are most appropriate because it is too difficult to have random assignment with the nature of our business.



Types of Sampling

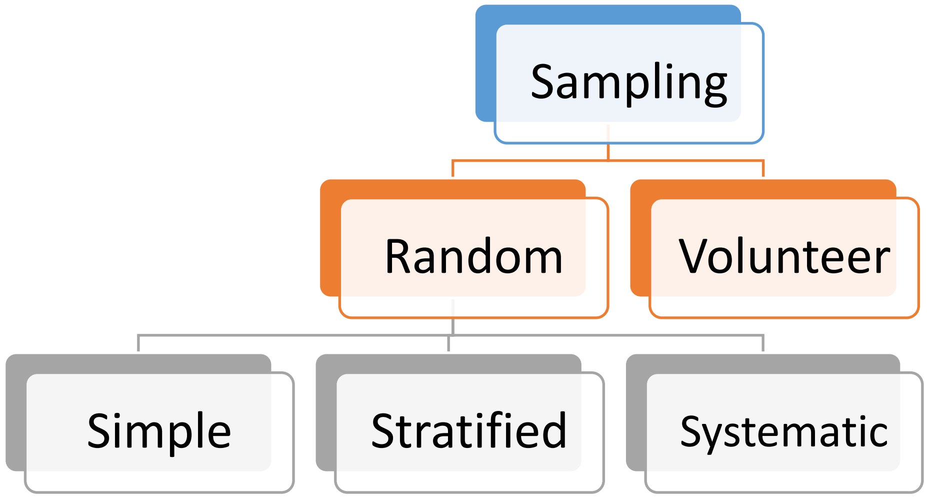


Figure 3. Types of sampling.

Sample is a subset of a targeted population. A **targeted population** is the entire pool of observations who participated in the program activities. There are two basic types of sampling: random and volunteer. See Figure 3. With **random sampling**, each person has an equal chance of being selected. Underneath random sampling, there are three sub-categories: simple, stratified, and systematic. **Simple random sampling** is where every person is thrown into the pot so to speak then will be selected for participation, **stratified random sampling** is where the persons will be selected based on a given characteristics (e.g., gender or racial classification), and **systematic random sampling** is where every nth person will be selected from a list (e.g., alphabetize list of 10th grade students with a high school). With **volunteer sampling,** each person will be selected by convenience and self-selected, which is how we typically sample in educational research.

These sampling techniques typically refer to quantitative research. In the world of qualitative research, purposeful sampling is utilized. **Purposeful sampling** is selecting persons based on the context of the evaluation, which can be explored extensive to uncover or confirm the concepts (Miles & Huberman, 1994).

Chapter 4

Collecting the Data

Step 4: Collecting the Data

The data collection phase offers many questions for the program evaluator to answer. During the evaluation planning phase, you determined what data will be collected to answer the research questions. As the evaluator, you need to think about the accessibility of the data and how these data will be collected consistently to answer those research questions. Table 3 presents typical data collection sources with comments about accessibility (Wall, n.d.).

Table 3

Typical Data Collection Sources

DATA SOURCE	EXAMPLES	COMMENTS
Activity Logs and Archival Documents	• Attendance records • Discipline referrals • Library book checkout records • Time spent logged into a computer • Visitation log for the school counselor • Number of students admitted to post-secondary education • Skill checklists • Essays • Review of performance ratings • Report cards • Standardized test scores	• Typically, these data are pre-existing, which makes them easily accessible.
Focus Groups	• Small group meeting to determine reasons for school violence • Small group meetings to assess academic achievement • Small group meetings to identify factors that promote positive self-esteem	• Use this method of data collection when you want to explore factors in depth, such as how and why. • Typically, the duration of focus groups can range from 45 to 90 minutes. • The list of protocol questions should be written and structured prior to the meeting.

Table 3 (*continued*)

DATA SOURCE	EXAMPLES	COMMENTS
Interviews	• Interview students about obstacles for making career decisions • Interview parents to assess the health habits of their children • Interview teachers about the strengths and weaknesses for a particular textbook • Interview college admission staff to make judgements about the level of preparation of high school students	• Use this method of data collection when you want to probe more deeply about certain attitudes, behaviors, feelings, and why actions are taken.
Observations	• Observations of behaviors in the school cafeteria • Observations of student interactions with others outside the academic classroom. • Classroom observations for teachers who attended a professional development workshop.	• Use this method of data collection when you want to get answers to questions that deal with "what and how many". • Observers will utilize a checklist to document the behaviors, but they will need training to ensure consistency.
Pre-existing/ Published Surveys and Measures	• Work ethics inventories • School climate surveys • Interest inventories • Personality inventories	• These data sources can save you time and effort, but they may not directly relate to your evaluation questions.
Locally Developed Surveys	• Survey teachers about what they think about a particular curriculum • Survey students about their feelings about bullying • Survey counselors about non-traditional career interests • Survey administrators about the disciplinary referral process	• Use this method of data collection when you want to answer "what, how, and why" questions. • They can include open-ended items to address the "why" questions.

Second, after you have determined how will the data be collected, Wall (n.d.) suggests this data collection action plan to outline the key components of the process. (See Figure 4.)

Research Question	Data Needed	Data Source
From Whom	**When**	**By Whom**

Figure 4. Data collection action plan template (adapted by J. Brown).

Using the curriculum evaluation model as an example, the longitudinal study will occur over a 5-year period. The secondary curriculum will be implemented in phases, which begin with Geometry and continue through Advanced Placement Calculus. To determine the level of mathematical proficiency, the students who enroll in the course during the year prior to curriculum implementation will take both of the benchmark examinations (i.e., mid-term and final). The scores from these students will be compared with the scores from the students who participate in the curriculum implementation. For example, Tables 4 and 5 display the timeline for assessment and data collection. As the program evaluator, you would develop this timeline and share it with the stakeholders and any individuals who may assist you with the data collection process. This proactive communication can ensure consistent data collection, particularly for longitudinal designs.

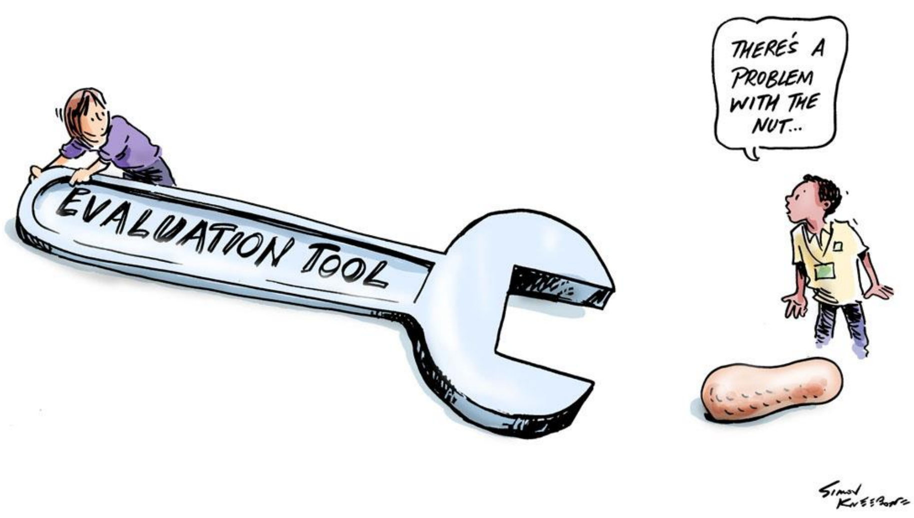


Table 4

Implementation Schedule and Evaluation Instruments for Each Year by Course

	Geometry	Algebra II	Pre-Calculus/ Trigonometry	AP Calculus
Year 0	**Comparison Assessments:** Benchmark Examinations			**Data Collection:** Results of AP Calculus Examination
Year 1	**Implementation:** Geometry Curriculum **Assessments:** Benchmark Examinations	**Comparison Assessments:** Benchmark Examinations		**Data Collection:** Results of AP Calculus Examination
Year 2	**Assessments:** Benchmark Examinations	**Implementation:** Algebra II Curriculum **Assessments:** Benchmark Examinations	**Comparison Assessments:** Benchmark Examinations	**Data Collection:** Results of AP Calculus Examination
Year 3	**Assessments:** Benchmark Examinations	**Assessments:** Benchmark Examinations	**Implementation:** Pre-Calculus/ Trigonometry Curriculum **Assessments:** Benchmark Examinations	**Comparison Assessments:** Benchmark Examinations **Data Collection:** Results of AP Calculus Examination
Year 4	**Assessments:** Benchmark Examinations	**Assessments:** Benchmark Examinations	**Assessments:** Benchmark Examinations	**Implementation:** AP Calculus Curriculum **Assessments:** Benchmark Examinations **Data Collection:** Results of AP Calculus Examination

Table 5

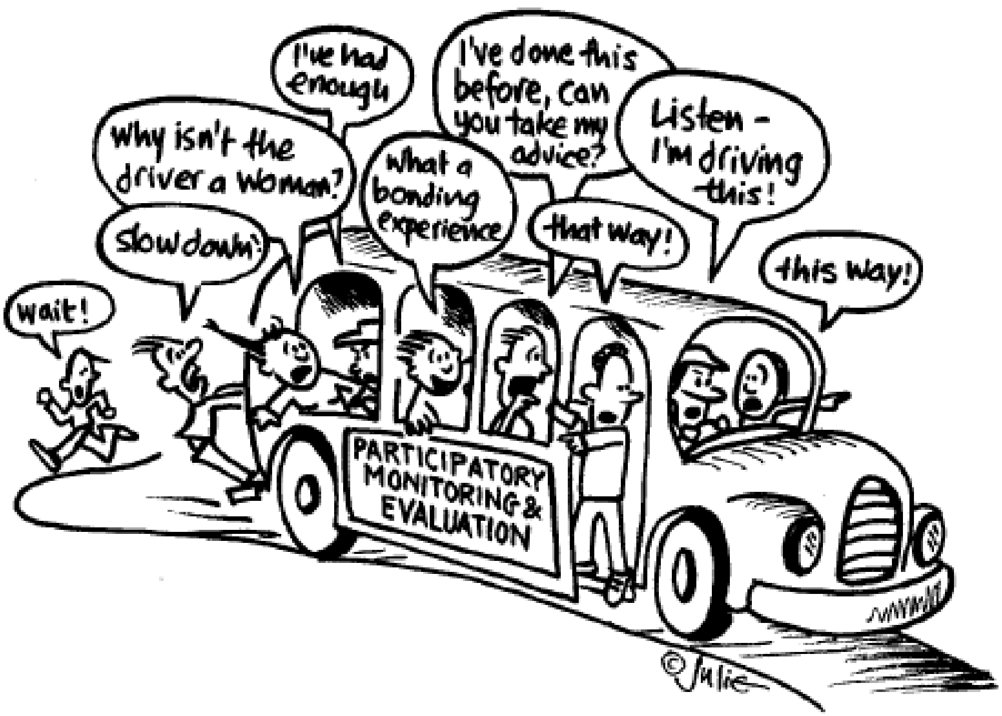
Evaluation Schedule and Instruments for Each Year by Stakeholder

	High School Graduates	Implementing Teachers	Professional Development Participants
Year 0	**Data Collection:** • Results of Graduation Exit Examination: Mathematics Subtest	**Assessments:** 1. Qualitative Interviews: Pre-planning, mid-term, end of course, and post-planning 2. Weekly implementation monitoring checklist 3. Demographic Surveys	**Assessments:** • Exit Surveys
Year 1	**Data Collection:** • Results of Graduation Exit Examination: Mathematics Subtest	**Assessments:** 1. Qualitative Interviews: Pre-planning, mid-term, end of course, and post-planning 2. Weekly implementation monitoring checklists 3. Demographic Surveys	**Assessments:** • Exit Surveys
Year 2	**Data Collection:** • Results of Graduation Exit Examination: Mathematics Subtest	**Assessments:** 1. Qualitative Interviews: Pre-planning, mid-term, end of course, and post-planning 2. Weekly implementation monitoring checklists 3. Demographic Surveys	**Assessments:** • Exit Surveys
Year 3	**Data Collection:** • Results of Graduation Exit Examination: Mathematics Subtest	**Assessments:** 1. Qualitative Interviews: Pre-planning, mid-term, end of course, and post-planning 2. Weekly implementation monitoring checklists 3. Demographic Surveys	**Assessments:** • Exit Surveys
Year 4	**Data Collection:** • Results of Graduation Exit Examination: Mathematics Subtest	**Assessments:** 1. Qualitative Interviews: Pre-planning, mid-term, end of course, and post-planning 2. Weekly implementation monitoring checklists 3. Demographic Surveys	**Assessments:** • Exit Surveys

Beginning with Year 1, the new curriculum will be implemented in all Geometry classes. For summative evaluations, a final benchmark examination will be given every 9 weeks to assess mathematical proficiency based on course content and performance standards. As a source of comparison, the students who are enrolled in Algebra II will be assessed using the two benchmark examinations (i.e., mid-term and final). For Years 2, 3, and 4, the same assessments and information will be collected as the curriculum is phrased into the remaining high school courses (i.e., Algebra II, Pre-Calculus/Trigonometry, and AP Calculus). Other data collections from the Registrar's Office will include 9-week course grades and attendance for each implemented course. Attendance assists with determining the **reach**, which is the extent to which the targeted population received the scheduled intervention dosages, and **dosage**, which is the amount of program activities received by the students. If the students did not attend class, then they are unlikely to benefit from the curriculum content.

One of our evaluation questions was "Have professional development sessions, conducted with the implementing teachers, promoted a successful curriculum implementation?" To collect data for these activities, at each professional development workshop, all participants will complete an exit survey to determine the effectiveness of the session and to determine future professional development needs. To monitor the application of knowledge gained during the professional development workshops, weekly informal observations using a checklist will monitor the implementation process in the classroom. At least one of the following people will conduct these observations: School Principal, Assistant Principal, Curriculum Director, or Assistant Curriculum Director. This data collection will assist with determining **fidelity**, which is the extent to which the implementation of program activities followed standardized procedures.

A formative, or process, evaluation will be conducted to assess the attitudes and instructional methods of the teachers throughout the implementation process. A demographic survey will collect information regarding education level, certification areas, and years of experience in public education. Qualitative interviews with the implementing teachers will ascertain their perceptions and gather feedback for program improvements. The series of interviews will be conducted during pre-planning, mid-term, end of the course, and postplanning. Adults are more likely to reject the new knowledge that contradicts their beliefs. The data gathered during these interviews will evaluate existing knowledge, beliefs, and motivations and will determine the extent to which the implementing teacher have ownership in the curriculum implementation process (Klingner, Ahwee, Pilonieta, & Menendez, 2003). See Appendix I and review the secondary mathematics curriculum's program evaluation program example.

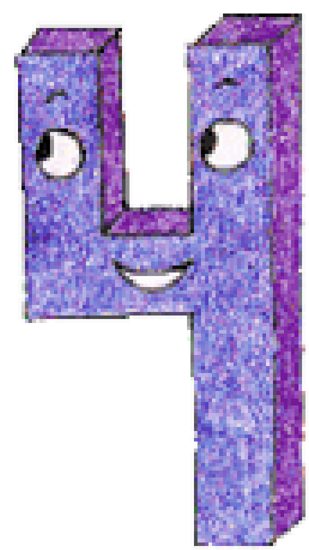


Chapter 5

Analyzing and Interpreting the Data

From our previous chapters, there are two types of data, quantitative and qualitative. In this chapter, we will begin with analyzing and interpreting quantitative data. With program evaluations, the findings should be interpreted then reported in a user-friendly format without statistical jargon. Your average stakeholder will view the statistical terms and symbols as a foreign language. As an educational psychologist, I was trained to utilize multiple statistical techniques ranging from simple descriptives to structural equation modeling. While the data analyst side of me wants to utilize upper-level statistics, I know that simpler is better. This "over" analysis is a common error. The purpose of program evaluation is not to illustrate your statistical knowledge and skills. Rather, it is to convey the findings to the stakeholders, which fit their needs and concerns.

There are two basic types of data: categorical and continuous. With **categorical data**, you are counting "things" (e.g., gender). Think about whether or not the "thing" can be placed in an individual box or can the "thing" be counted. With **continuous data**, you have a range of numbers on a continuum (e.g., test scores). In Figure 5, a flowchart for determining independent and dependent variables is presented. An **independent variable (IV)** is the variable, or observational characteristic, which is not dependent on other observations as the name implies. Sometimes, the IV is referred to as the grouping variable if more than one group exists within the study. A **dependent variable (DV)** is the variable



Step 5: Analyzing and Interpreting Quantitative Data

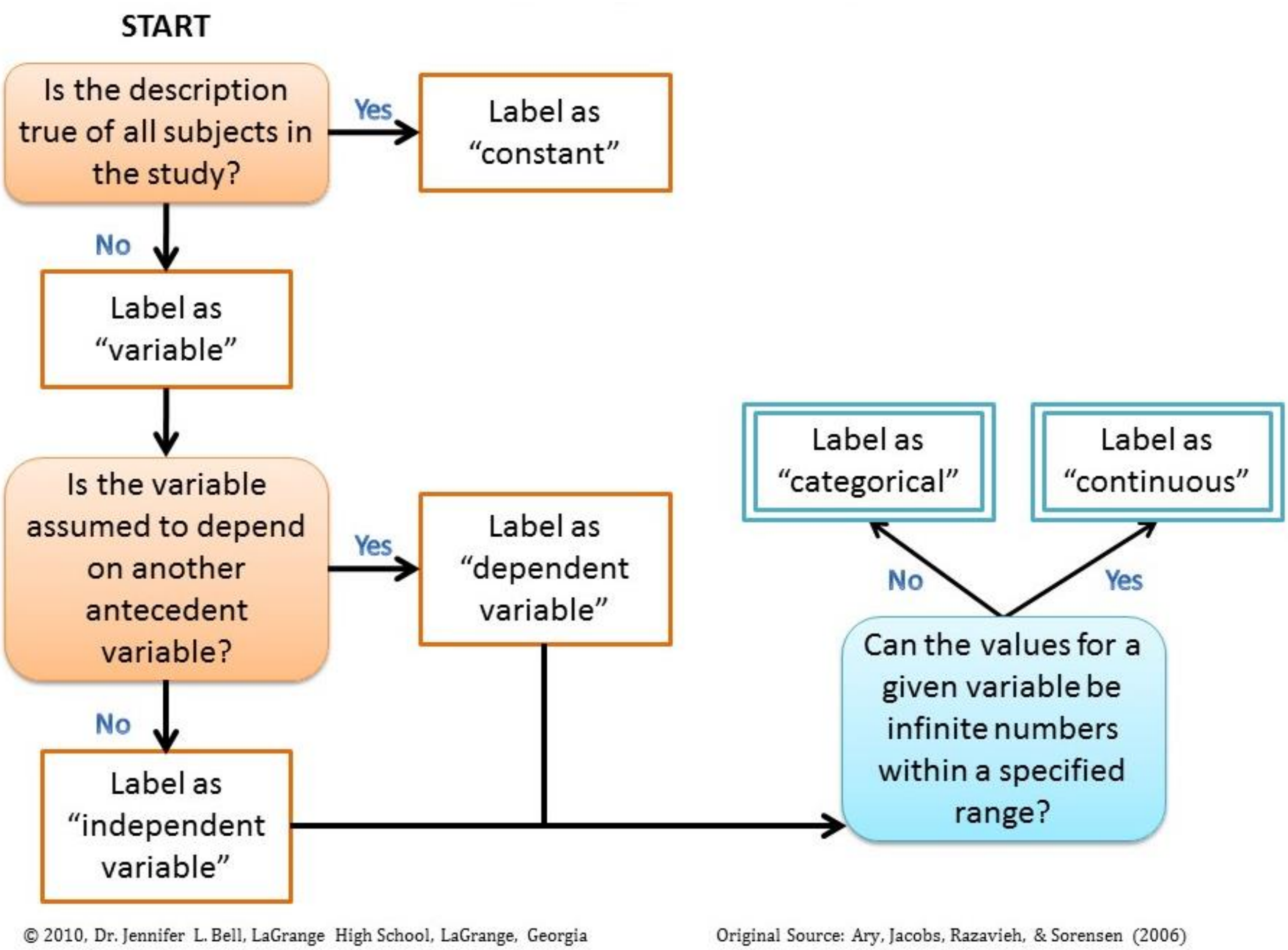


Figure 5. Flowchart for classifying variables and constants.

that is dependent upon another characteristic or observation. (*Note*: These two basic types can be broken down further when working in the field of educational research. Categorical data includes the nominal level of measurement. Continuous data include ordinal, interval, and ratio levels of measurement. For the purposes of program evaluation, we will stay with the two basic types of data.) See Figure 6 for a flowchart to determine a variable's level of measurement. There is a debate in educational research about whether or not ordinal data should be analyzed using the same statistics as interval and ratio data. In program evaluation, rating scales (e.g., *Strongly Agree* to *Strongly Disagree*), which are considered ordinal data, are analyzed using statistics for continuous data.

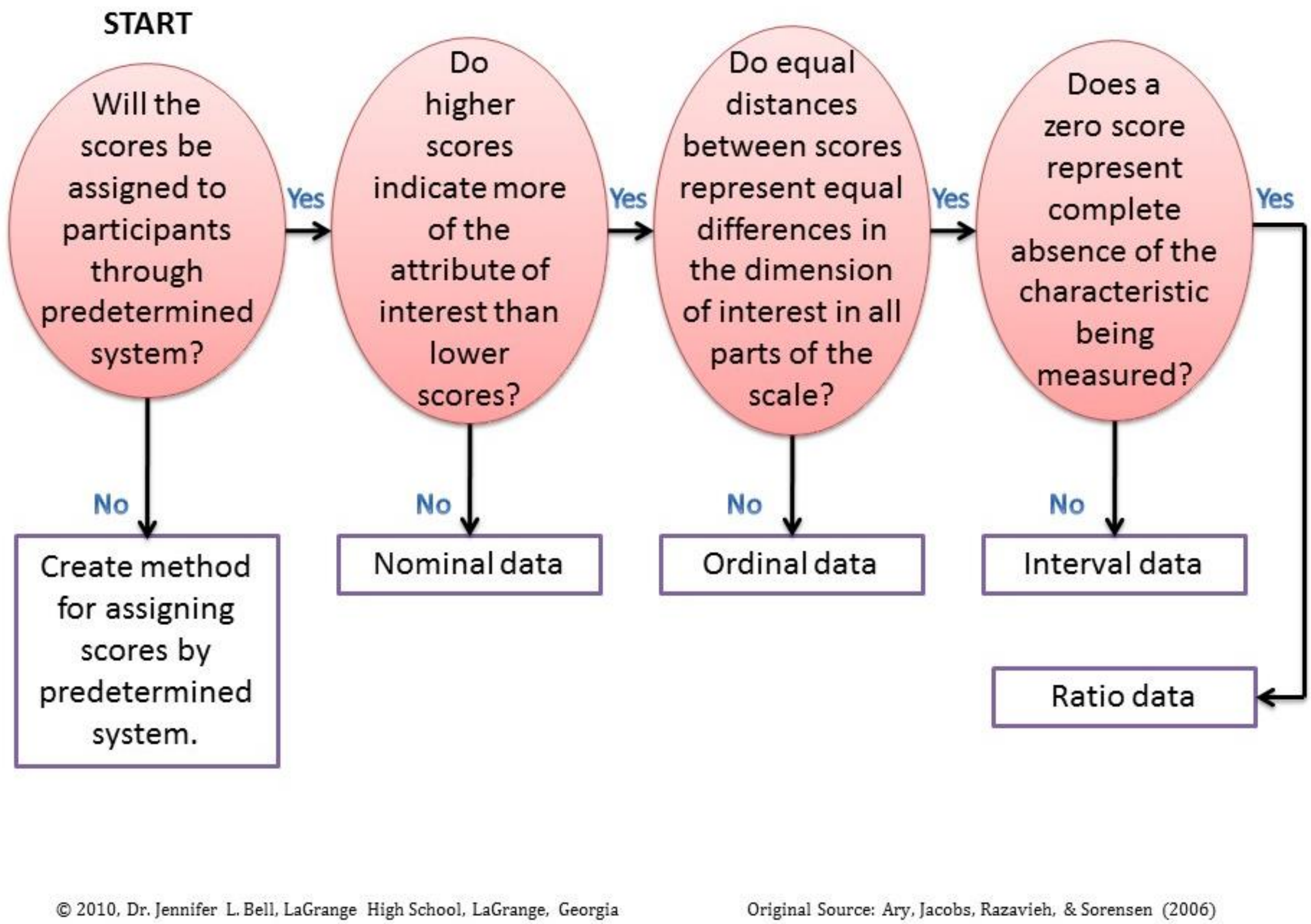


Figure 6. Flowchart for determining scales of measurement.

PRACTICE

For each of the following scenarios, indicate the sample, IV, and DV.

1. Do third-grade students who finger spell their spelling words perform better on their weekly spelling tests than those students who do not finger spell?
2. The purpose of this study was to examine the impact of gender on standardized science assessments among 11th-grade students.

ANSWERS: 1. Sample: third-grade students; IV: group (control/treatment); DV: spelling ability as measured by weekly spelling tests 2. Sample: 11th-grade students; IV: gender (male/female); DV: science achievement as measured by standardized science assessments.

PRACTICE

For each of the following measures, indicate whether they would considered categorical or continuous data.

1. Socioeconomic status
2. Final averages
3. SAT scores
4. Racial classifications
5. Attendance
6. Rankings after math team competition
7. Height
8. Shoe size

ANSWERS: 1. Categorical (e.g., 15 low, 12 middle, and 10 high SES) 2. Continuous 3. Continuous 4. Categorical (e.g., 10 white and 10 black students) 5. Categorical (e.g., 40 days present and 5 days absent) 6. Categorical (e.g., 2 – 1^{st} places, 1 – 2^{nd} place, and 4 – 3^{rd} places) 7. Continuous 8. Categorical (e.g., 5 size 7 shoes, 1 size 9 shoe, and 2 size 10 shoes).

Summarizing the Quantitative Data

In educational research, if you have categorical data, you will "count" the "things" in each category, which is referred to as frequency counts. If you have continuous data, you will run descriptive statistics, which is the numerical summary of the data. Descriptive statistics can be broken into two categories: Measures of Central Tendency and Measures of Dispersion. **Measures of Central Tendency** tell you the center of the data. **Measures of Dispersion** tell you spread of the data or how much variation exists. Figure 7 defines the two measures of central tendency, median (*M*) and mean (*Mdn*), and the two measures of dispersion, range and standard deviation (*SD*). Each of these measures are affected by outliers, except the median. As a good rule

OUTlier
the value that is significantly **out**side the range of the other values in the dataset

of thumb, you can compare the mean and the median. If there are no outliers, the numbers should be similar. Your standard deviation is another good indication of outliers. Large standard deviations (i.e., increased spread in the data) indicate fewer data points are clustered around the mean. Typically, in program evaluation, data points that are more than two standard deviations from the mean are considered outliers.

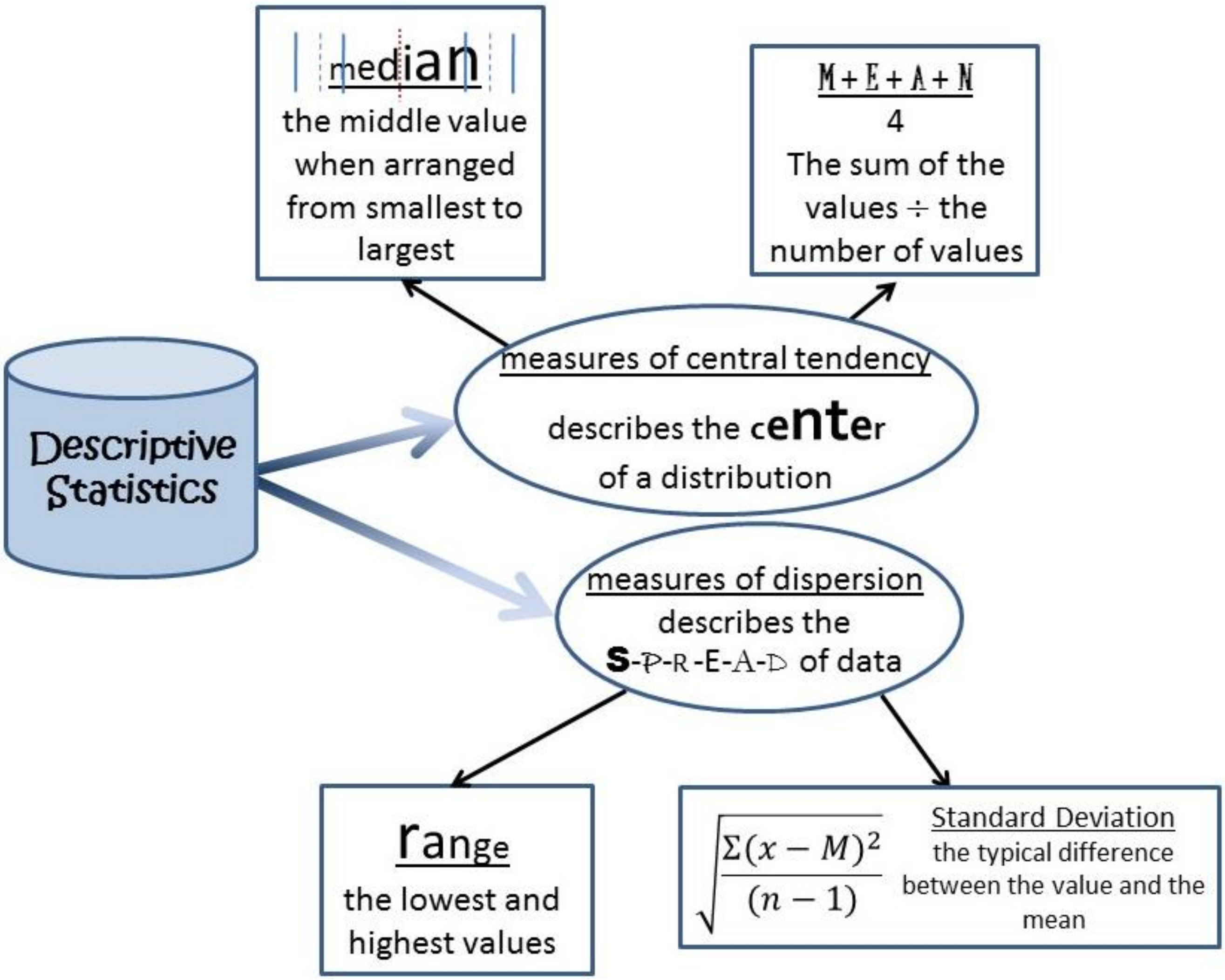


Figure 7. Types and definitions of descriptive statistics.

PRACTICE

Use following table to answer the questions about descriptive statistics. Notice, in APA-formatted tables, you should use *n* for frequency count, *M* for mean, *Mdn* for median, and *SD* for standard deviation. Also, the title of the table should be italicized.

Frequency and Descriptives for Original and Retake Scores by Core Department

				Original					Retake		
Department	*n*	*M*	*Mdn*	*SD*	Min	Max	*M*	*Mdn*	*SD*	Min	Max
English	301	54.59	57	16.45	0	94	73.28	75	18.95	0	103
Math	551	55.60	60	17.15	0	95	56.97	60	23.96	0	100
Science	691	50.74	55	18.40	0	95	60.70	64	25.02	0	116
Social Studies	296	53.70	57	16.26	0	91	70.26	74	22.35	0	100

1. Which core department had the most improvement from the original and retake assessment?
2. Which core department had the better retake scores? Provide a rationale.
3. Which core department had more variation in their original scores? Provide a rationale.
4. Which core department had less variation in their retake scores? Provide a rationale.
5. Which core department had more students participate? Provide a rationale.

ANSWERS: 1. English (73.28 – 54.59 = 18.69) 2. English; the mean and median retake scores were higher compared to the other departments. 3. Science; the standard deviation was higher compared to the other departments, and the median differed from the mean. 4. English; the standard deviation was lower compared to the other departments, and the median was similar to the mean. 5. Science; the frequency (*n*) for participants was larger compared to the other departments.

PRACTICE

Let us practice analyzing the descriptives for a small dataset.

Five students take a math quiz with 15 items. Here are the number of correct items for each student.

7, 8, 8, 9, 13

1. What is the mean? (**Answer:** 9)
2. What is the median? (**Answer:** 8)
3. What is the range of scores? (**Answer:** 7 to 13)
4. What is the standard deviation? (**Answer:** 2.35 – *Note*: You will not need to compute standard deviation by hand, but it helps to see where the number derives.)

x	$(x – M)$	$(x – M)^2$
7	7 – 9. = -2	4
8	8 – 9 =-1	1
8	8 – 9 =-1	1
9	9 – 9 = 0	0
13	13 – 9 = 4	16
	Σ	22
	$\Sigma/(n – 1)$	5.5
	$\sqrt{\Sigma/(n-1)}$	2.35

5. Are there any outliers? (**Answer:** No, 13 is within two standard deviations of the mean, and the mean and median are fairly similar. 9 + 2.35 + 2.35 = 13.7)

Most home and school computers have Microsoft Excel as an available program option; however, I prefer SPSS for data analysis. SPSS is available as a 2-week trial version from IBM. Also, it is available on campus in some of the computer labs. I will demonstrate the steps for analyzing the data with the various statistics using Excel's Analysis ToolPak and SPSS. The directions for how to load the Analysis ToolPak are listed below. If you would like more information about setting up a database in Excel, there is a packet available on my website (http://www.bugforteachers.com/prog_eval.html).

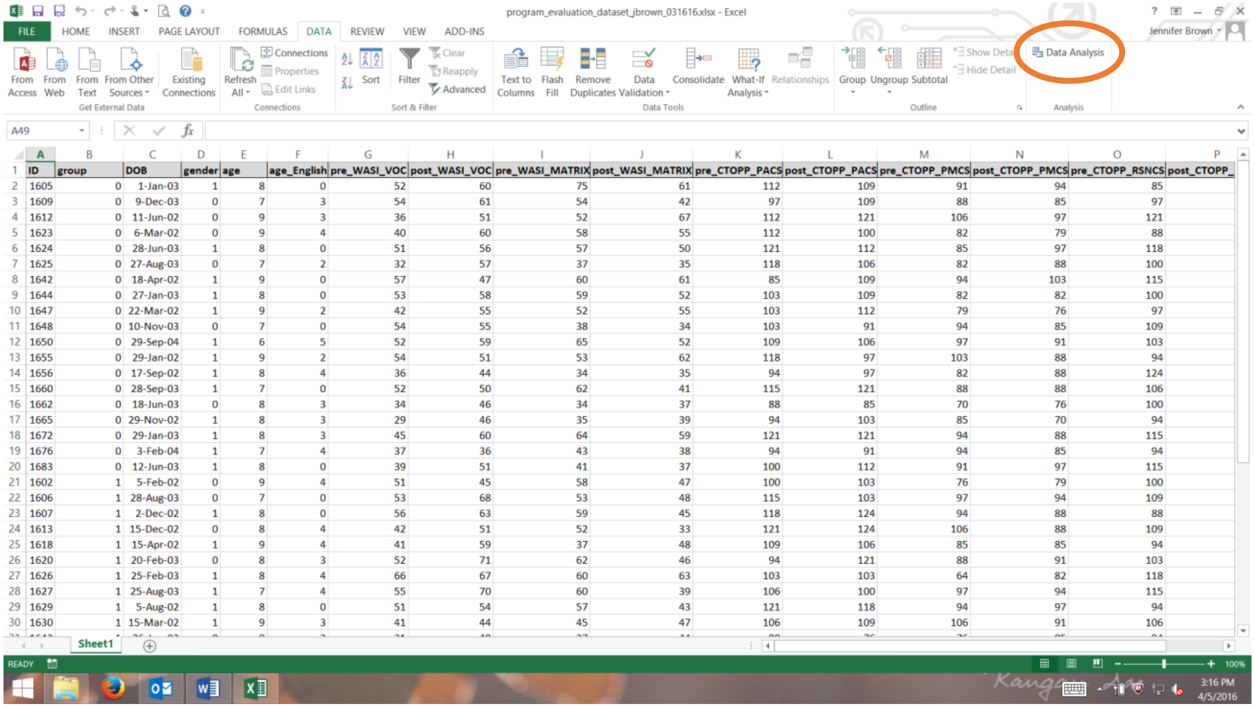
The musical training quantitative dataset will be utilized for demonstrating the various statistical analyses. (The Excel and SPSS files are available for download from http://www.bugforteachers.com/prog_eval.html.) The original dataset was retrieved from Slater et al. (2014). The study examined the effects of a musical training program on phonological awareness with 42 bilingual (Spanish/English) students from a low-income area in California. The study had a control and experimental group. See Appendix A for the background information, measures, variable names, and labels. After the demonstration activities and practice activities, I will include an interpretation based on the produced output. Sometimes, I will include commentary for educational purposes in dark orange after the output and/or interpretations. After each analysis demonstration, there will be a "Why?" section to reinforce the purpose and application of the previously demonstrated technique.

READINGSTATS.COM/SIXTH/INDEX.HTM

If you are looking for additional assistance, this website by Sky Huck offers interactive quizzes, online resources, e-articles, and common misconceptions for a variety of topics related to reading statistics and research (Huck, 2012).

How to Upload the "Analysis ToolPak" in Excel

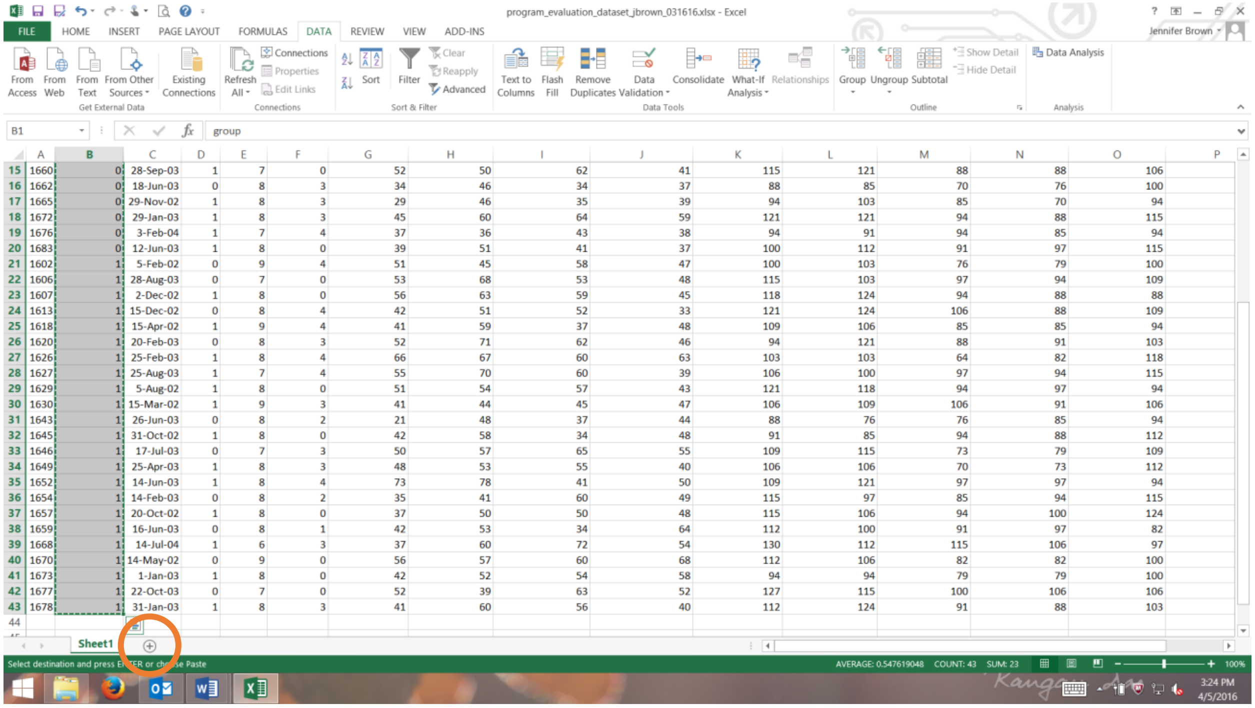
1. Select the **File** tab.
2. Select *Options*.
3. On the pop-up screen, select *Add-Ins*.
4. In the "Manage" box, select **Excel Add-ins**.
5. Select **Go**.
6. In the "Add-Ins available" box, select the box beside "Analysis ToolPak".
7. Select **OK**.
    a. If "Analysis ToolPak is not listed in the "Add-Ins available" box, select **Browse** to locate it.
    b. If prompted to install "Analysis ToolPak" on your computer, select **Yes**.
8. The "Data Analysis" command is available on the **Data** tab.



How to Analyze Frequencies in Excel

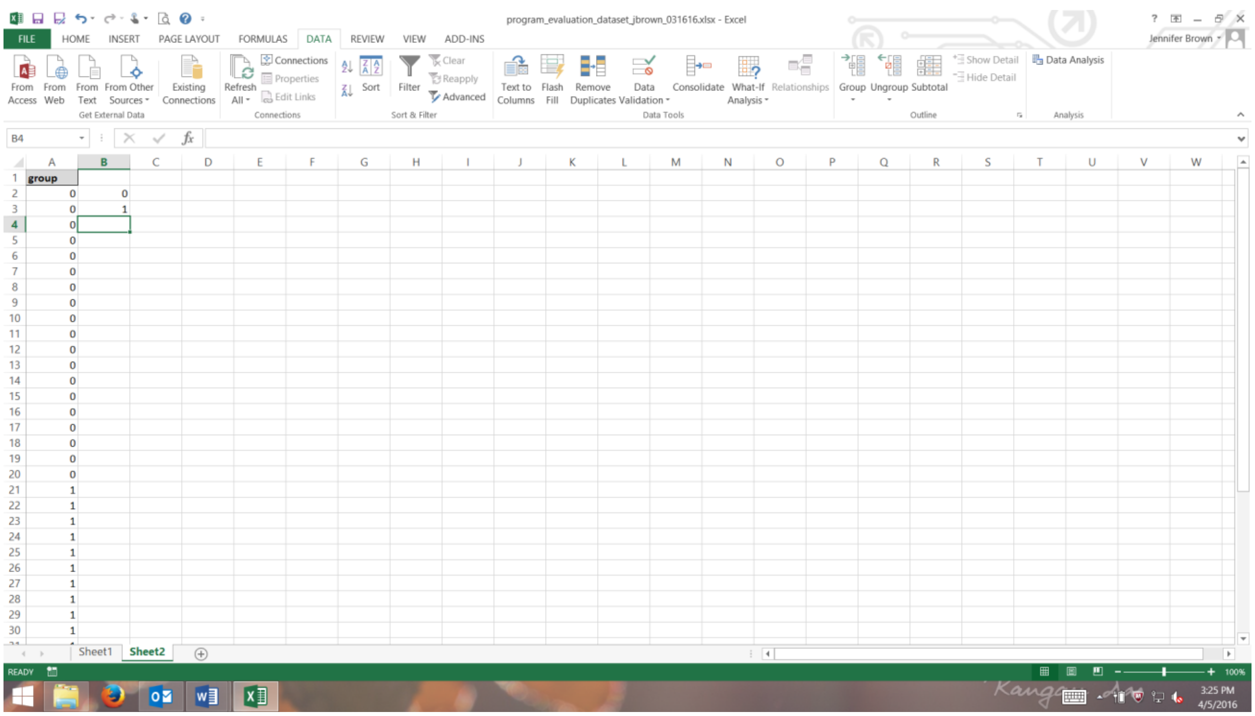
Open the "program_evaluation_Excel_musical_training" dataset in Excel.

1. Copy all of the data within the desired column including the header. (For this example, you should copy the *group* column from B1 to B43.)
2. Open a new worksheet by selecting the + in the lower left corner.



3. Paste the data into column A.

4. In column B, enter the data labels utilized within column A. (For this example, you should enter "0" in B2 and "1" in B3.)

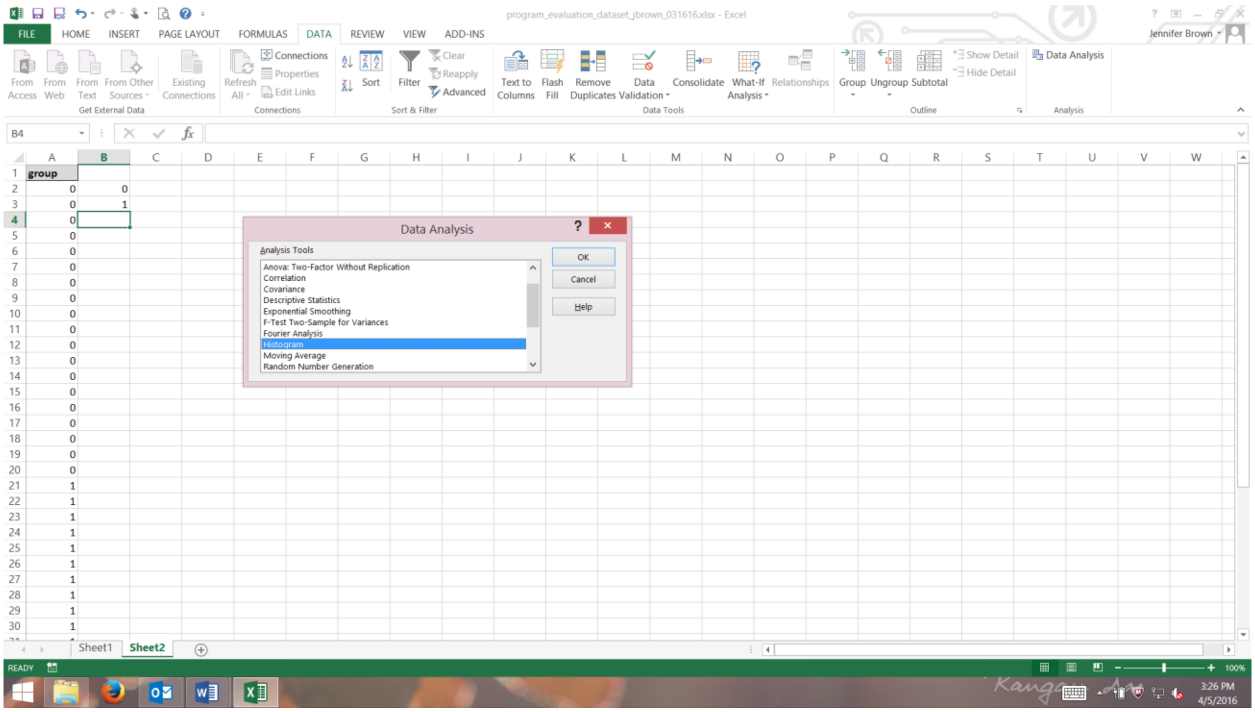


5. Select the **Data** tab.

6. Select “Data Analysis”.

7. In the pop-up window, select “Histogram”.

8. Select **OK**.



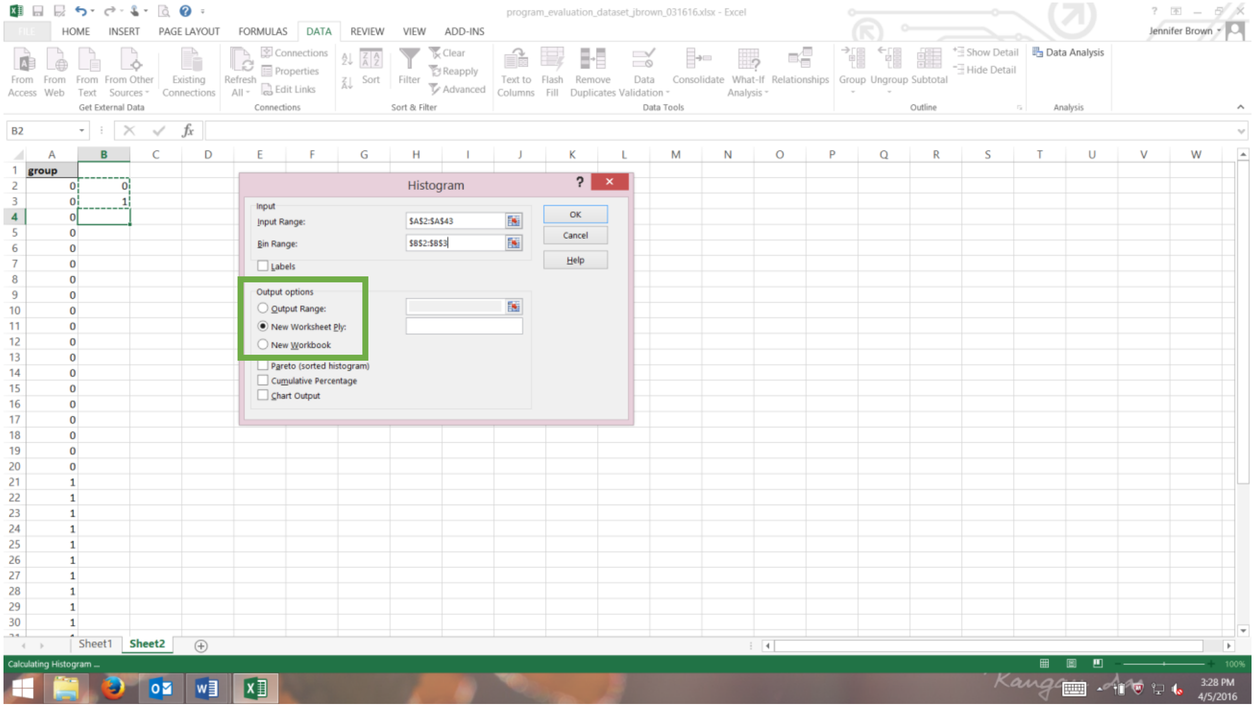
9. Click inside the “Input Range” box.

10. Highlight all of the data in column A. (For this example, you should highlight A2 through A43.)

11. Click inside the “Bin Range” box.

12. Highlight all of the categories in column B. (For this example, you should highlight B2 and B3.)

13. Make sure the radial beside “New Worksheet Ply” is checked under “Output”.



14. Select **OK**. (A new worksheet will open with the output.)

Bin	*Frequency*
0	19
1	23
More	0

Participant Demographics (Categorical Data)

There were 19 participants in the control group and 23 participants in the experimental group.

Group	*n*	%
Control	19	45.2
Experimental	23	54.8
Total	42	100.0

(*Note:* To calculate the percentages, divide the number of participants in each group by the total number of participants, *n* = 42. Then, multiply by 100.)

Why?

Why are we analyzing categorical data with frequency counts?

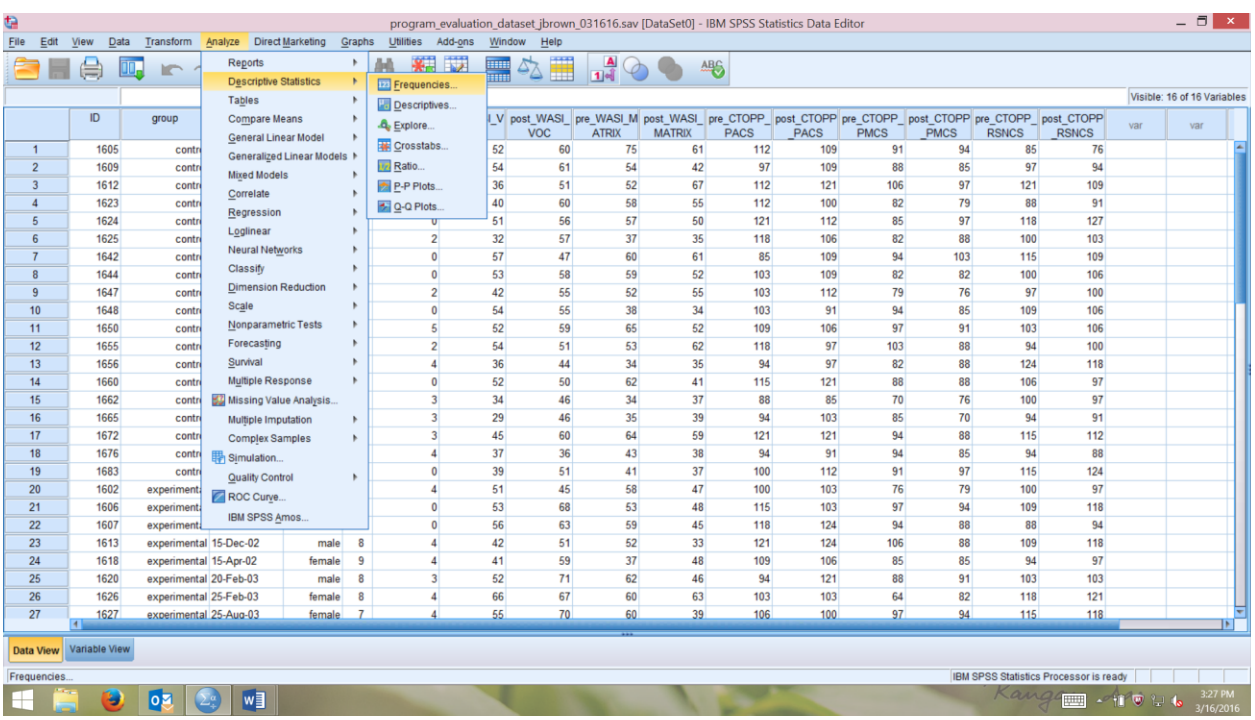
Frequency counts are utilized to “count” values within a given variable. With this previous example, we counted the number of participants in the control and experimental groups. Typically, this analysis technique is conducted with categorical or nominal data (e.g., gender, racial classification, or grade level). It is not appropriate to analyze this type of data with descriptives (e.g., mean, median, standard deviation, or range). You cannot have a gender of 1.5.

How to Analyze Frequencies in SPSS

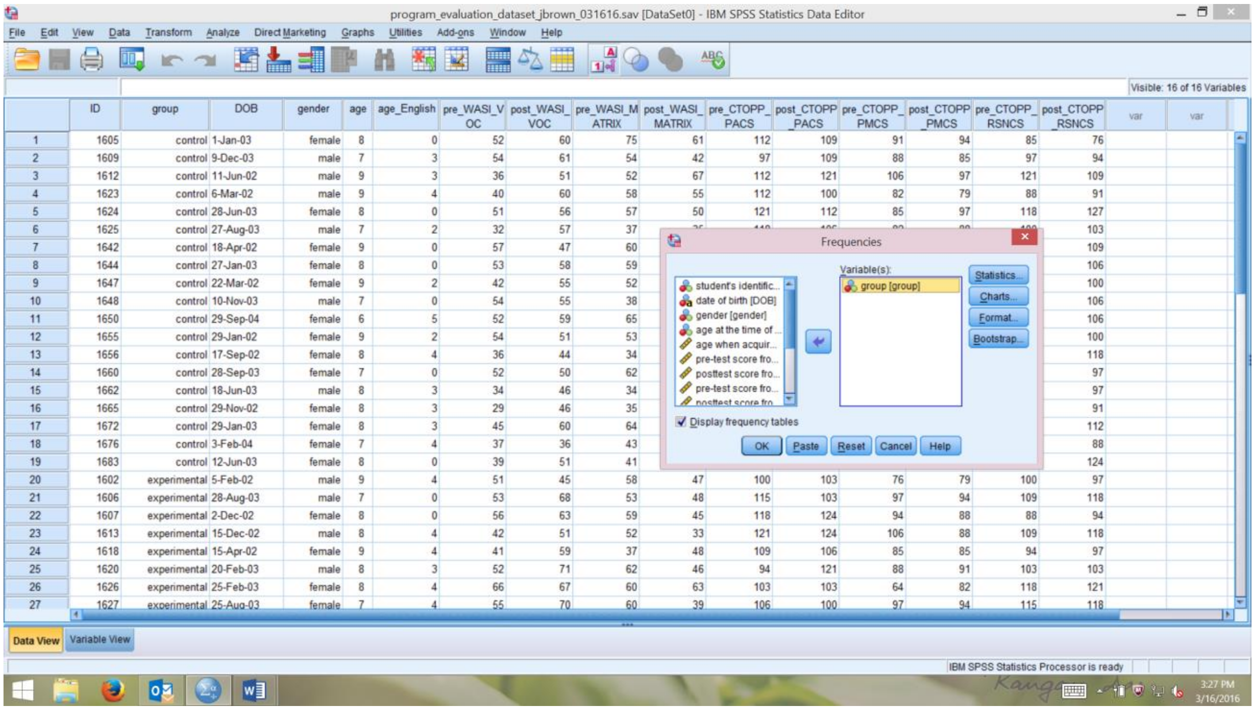
Open the “program_evaluation_SPSS_musical_training” dataset in SPSS.

Participant Demographics (Categorical Data)

1. Analyze → Descriptive Statistics → Frequencies



2. Select and move the following variable into the box: *group*. (*Note*: Highlight using left mouse and select the arrow icon.)



3. Select **OK**. (The analysis will appear on the output screen.)

SPSS Output

group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	control	19	45.2	45.2	45.2
	experimental	23	54.8	54.8	100.0
	Total	42	100.0	100.0	

PRACTICE

You try with the *gender* variable. Compare your output the following output.

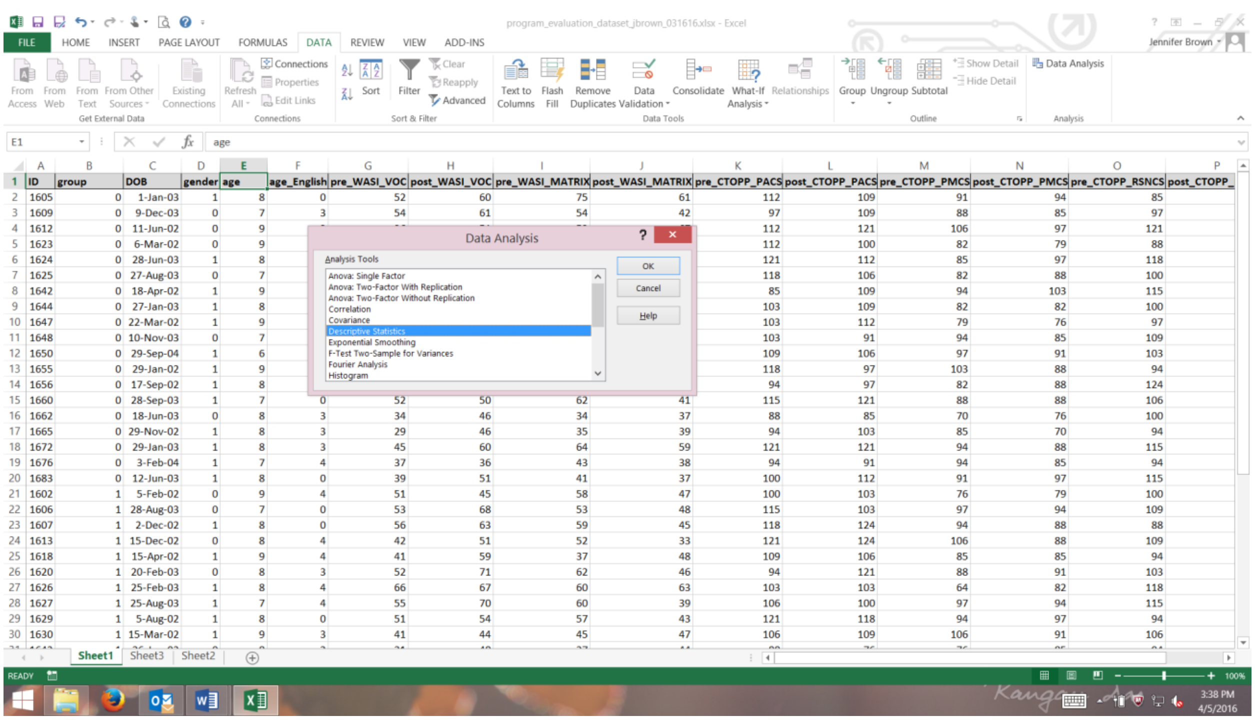
gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	16	38.1	38.1	38.1
	female	26	61.9	61.9	100.0
	Total	42	100.0	100.0	

Of the 42 participants, there were 16 (38.1%) males and 26 (61.9%) females.

How to Analyze Descriptives in Excel

1. Select the **Data** tab.
2. Select "Data Analysis".
3. In the pop-up window, select "Descriptive Statistics".



4. Select **OK**.

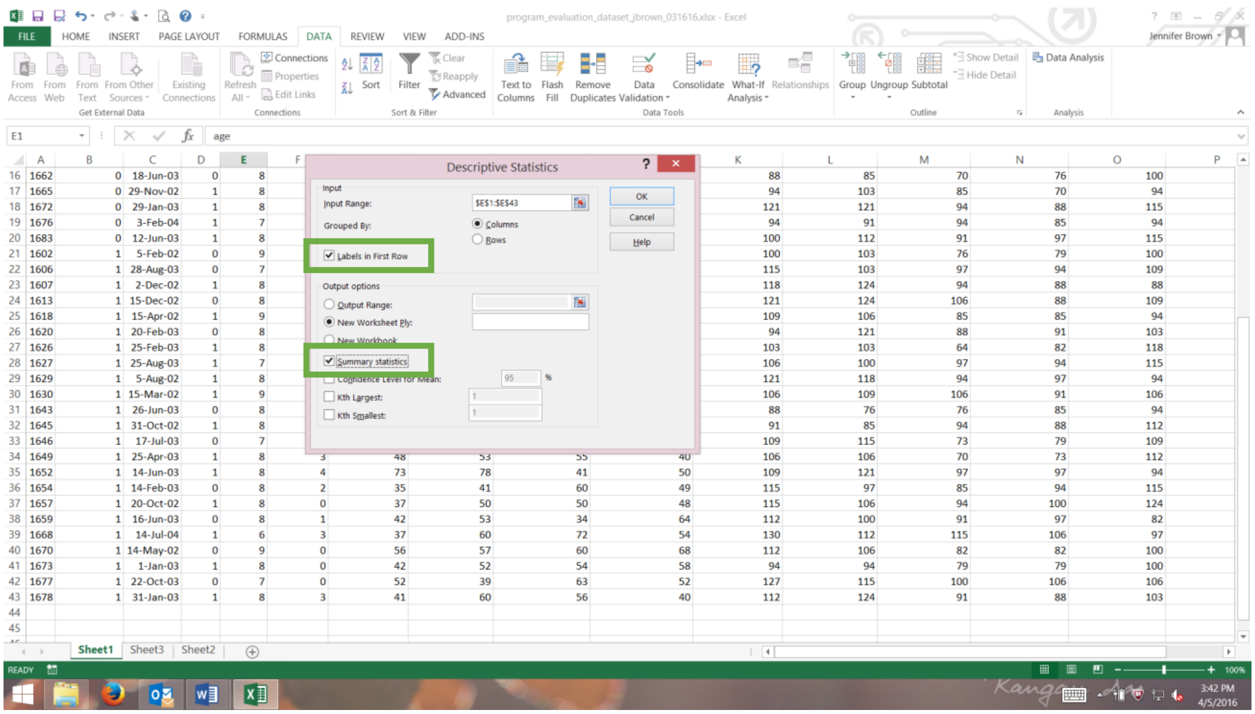
5. Click inside the "Input Range" box.

6. Highlight all of the data in the *age* column. (For this example, you should highlight from E1 to E43.)

7. Check the box beside "Labels in First Row".

8. Check the box beside "Summary Statistics".

9. Make sure the radial beside "New Worksheet Ply" is checked under "Output".



10. Select **OK**. (A new worksheet will open with the output.)

Excel Output

age	
Mean	7.904762
Standard Error	0.121973
Median	8
Mode	8

Standard Deviation	0.790478
Sample Variance	0.624855
Kurtosis	0.044025
Skewness	-0.44818
Range	3
Minimum	6
Maximum	9
Sum	332
Count	42

Participant Test Scores (Continuous Data)

The mean age was 7.90 years with a standard deviation of 0.79. The median age was 8 years with a minimum of 6 years old and a maximum of 9 years old.

Why?

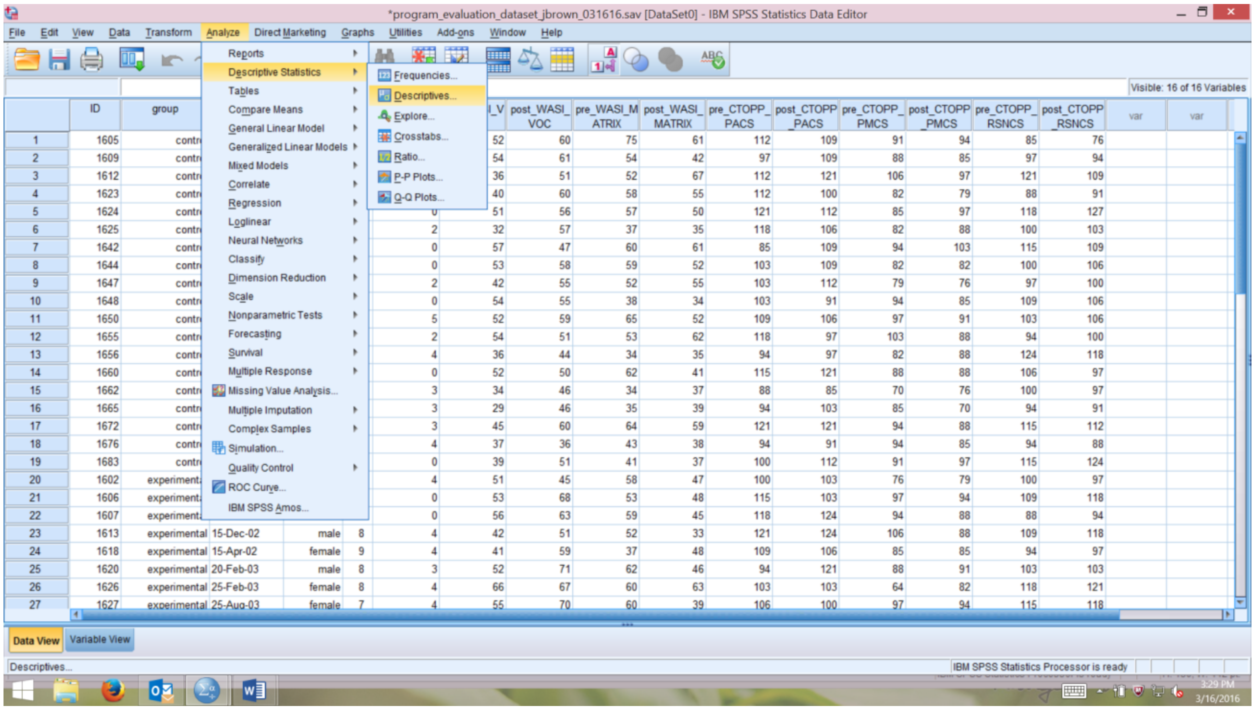
Why are we analyzing continuous data with descriptives?

Descriptives serves as a method to summarize the data. For the reader, it would be too cumbersome to see a long list of frequency counts, which could be quite lengthy if you consider the possible grade frequency counts for one classroom assessment. In addition, there would be no value in the information for the reader.

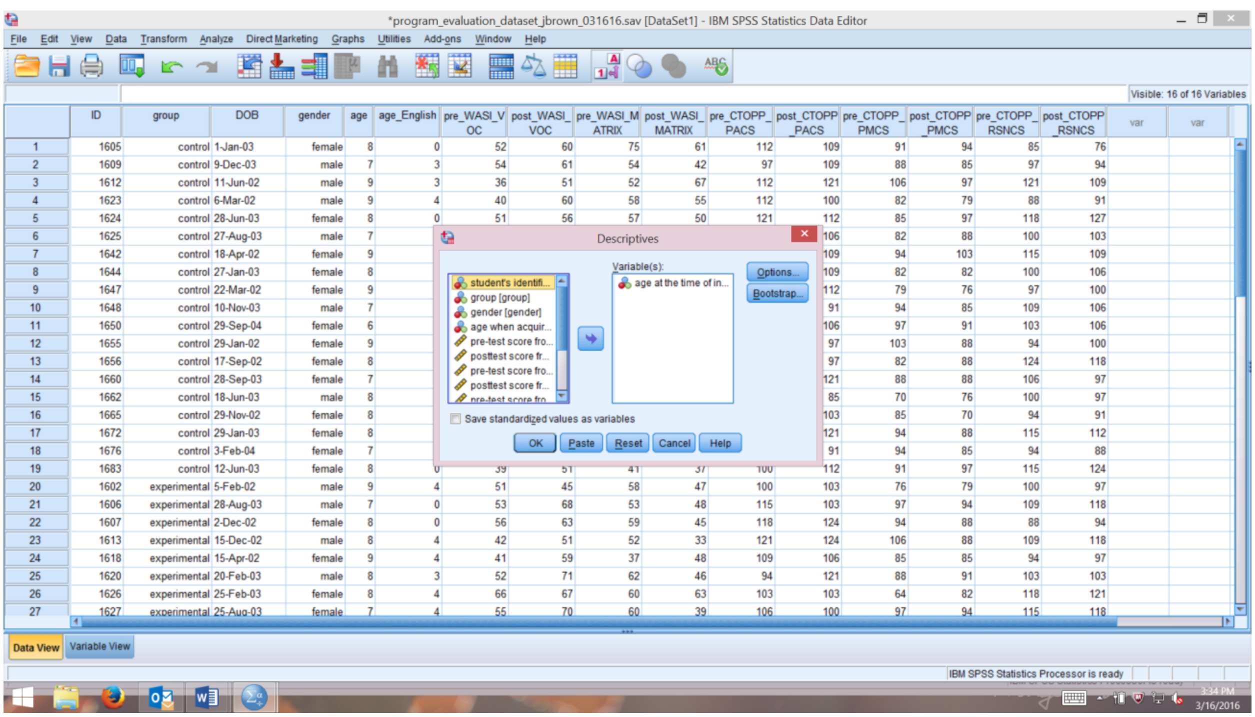
How to Analyze Descriptives in SPSS

Participant Test Scores (Continuous Data)

1. Analyze → Descriptive Statistics → Descriptives



2. Select and move the following variable into the box: *age*. (*Note:* Highlight using left mouse and select the arrow icon.)



3. Select **OK**. (The analysis will appear on the output screen.)

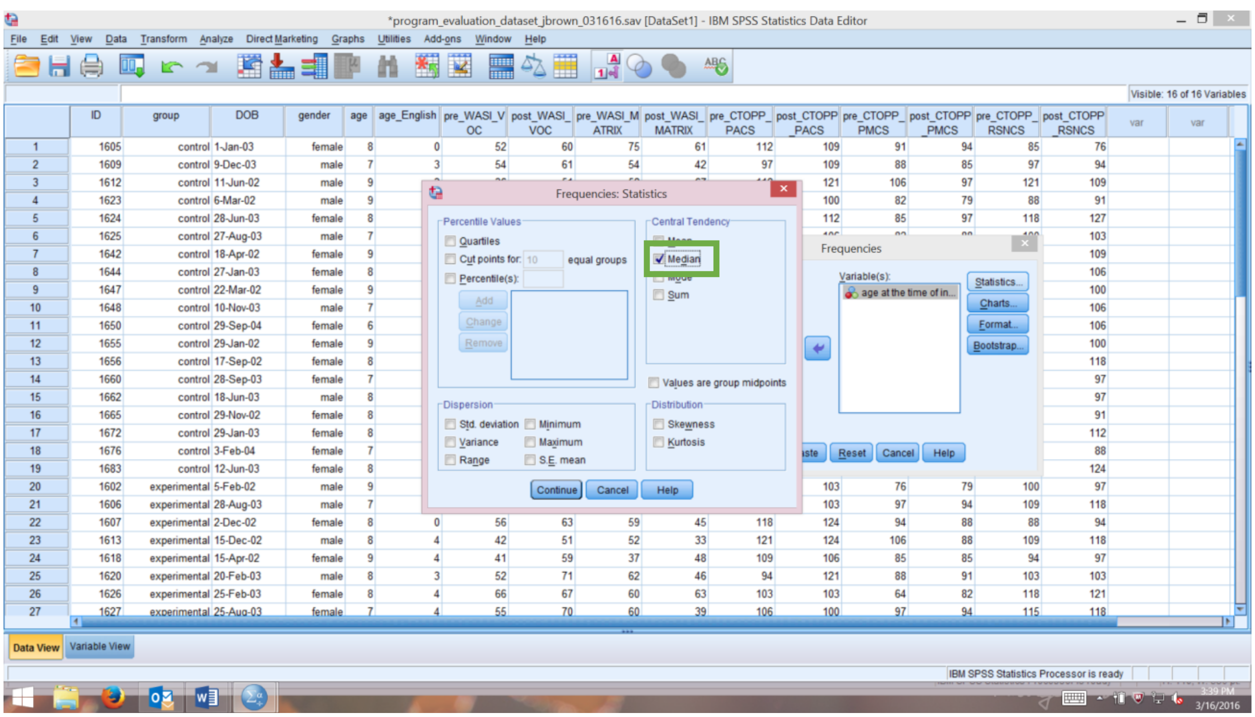
SPSS Output

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
age at the time of intervention	42	6	9	7.90	.790
Valid N (listwise)	42				

If you want to determine the median of the data, you must run it through "Frequencies".

- Move the *age* variable into the box.
- Select the **Statistics**.
- Add a checkmark beside "median" underneath Central Tendency in the pop-up window.
- Select **Continue**.
- Select **OK**.



SPSS Output

Statistics

age at the time of intervention

N	Valid	42
	Missing	0
Median		8.00

PRACTICE

You try with the *age_English* variable. Compare your output the following output.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
age when acquired English	42	0	5	2.02	1.689
Valid N (listwise)	42				

Statistics

age when acquired English

N	Valid	42
	Missing	0
Median		2.50

The mean age when the participants acquired English was 2.02 years with a standard deviation of 1.69. The median age was 2.50 years with a minimum of 0 years old and a maximum of 5 years old.

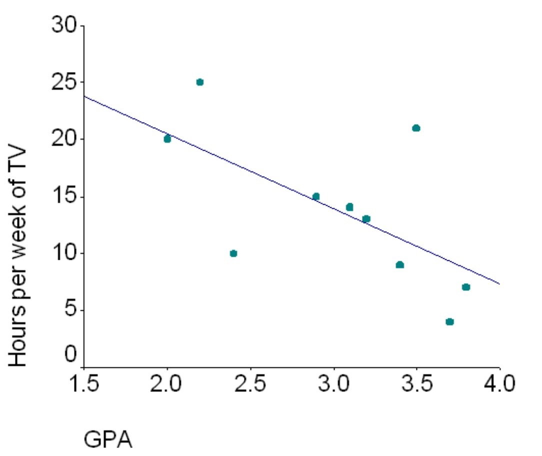
Another common approach to analyzing categorical and continuous data is tables and graphs. The visual representations are easily viewed and can serve as a source for comparison when sharing the findings with stakeholders. In some of the previous practice activities, I provided APA-formatted tables with frequency and descriptive data results. More tables are available in the program evaluation example reports located in the Appendices, particularly Program Evaluation Report Example #4 in Appendix G. In general, it is recommended to utilize bar (histograms) and pie graphs for representing categorical data and line and scatterplots for continuous data. When writing program evaluation reports, I utilize the chart function in Excel to create my graphs.

With most program evaluations, data analysis includes basic descriptives, which include means, standard deviations, ranges, frequency counts, and percentages; however, it depends on the audience of the evaluation. (See Program Evaluation Report Examples #1 through #4 in Appendices D through G.) Using the curriculum example, descriptives will assess exit surveys from the professional development workshops, weekly observations, 9-week course grades, and class attendance.

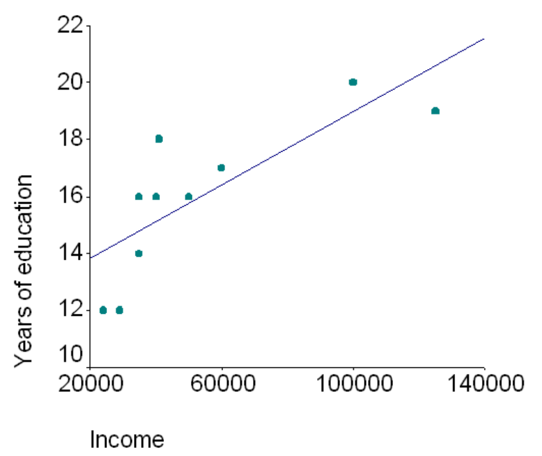
Measure of Association: Bivariate Correlation

With two continuous variables, you will conduct a Pearson Product Moment Correlation, or Pearson *r*, to determine if a relationship exists, which is symbolized with an italicized lowercase *r*. This statistical procedure does have the assumption that the continuous data are linear instead of curvilinear. (See Correlational Research Design.) The correlational coefficients range from -1.00 to 1.00. A **negative correlation** or relationship indicates one value increases as the other one decreases. A **positive correlation** or relationship indicates one value increases as the other one increases. See the illustrations below to visualize the correlations. Notice, the zero relationship looks like a child dropped Cheerio's on the floor. It has no resemblance of a linear formation.

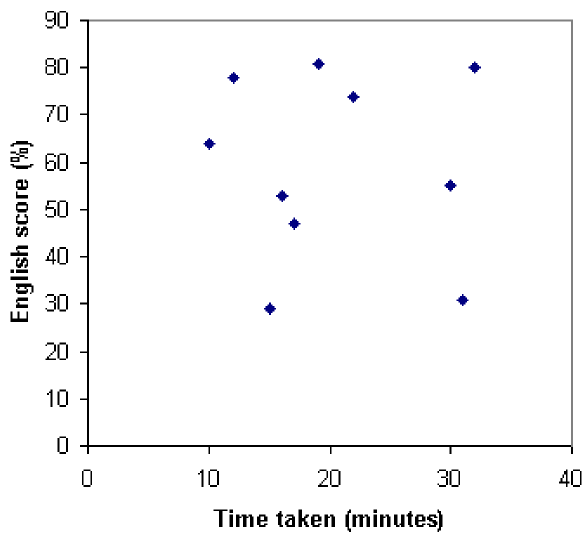
Negative Correlation



Positive Correlation



No Correlation

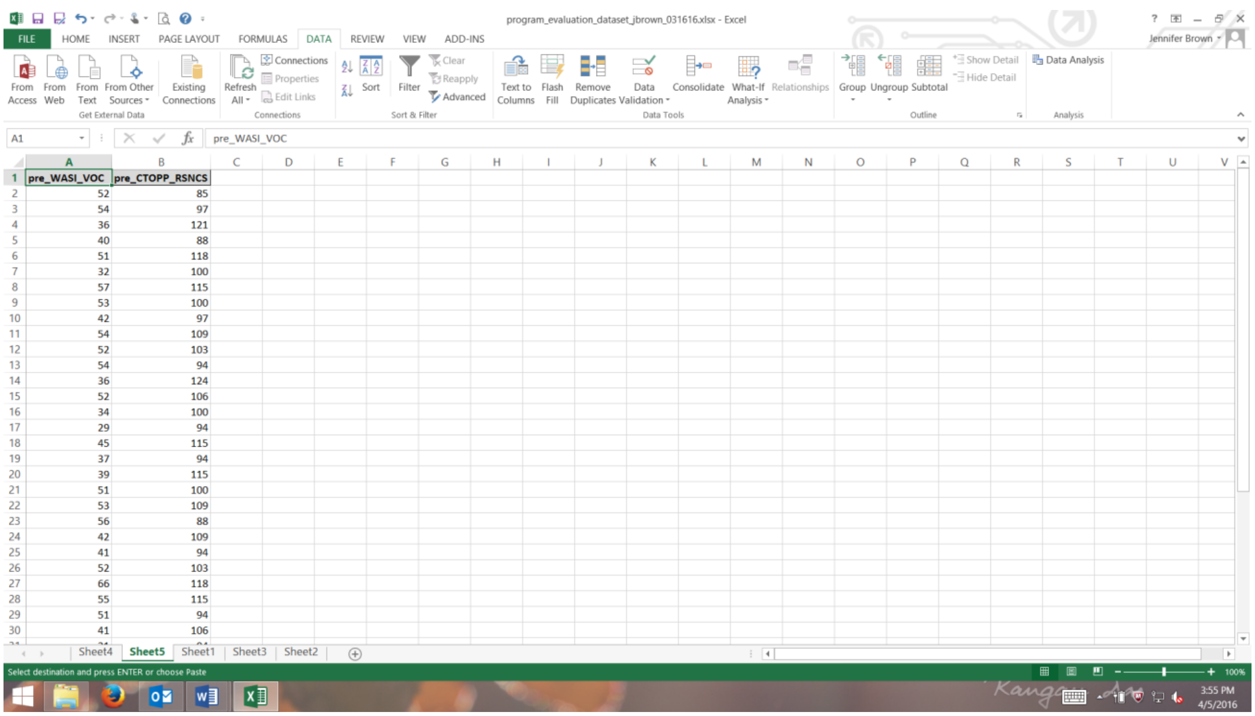


To interpret the correlational coefficients in educational research, you will use Jacob Cohen's (1988) guidelines.

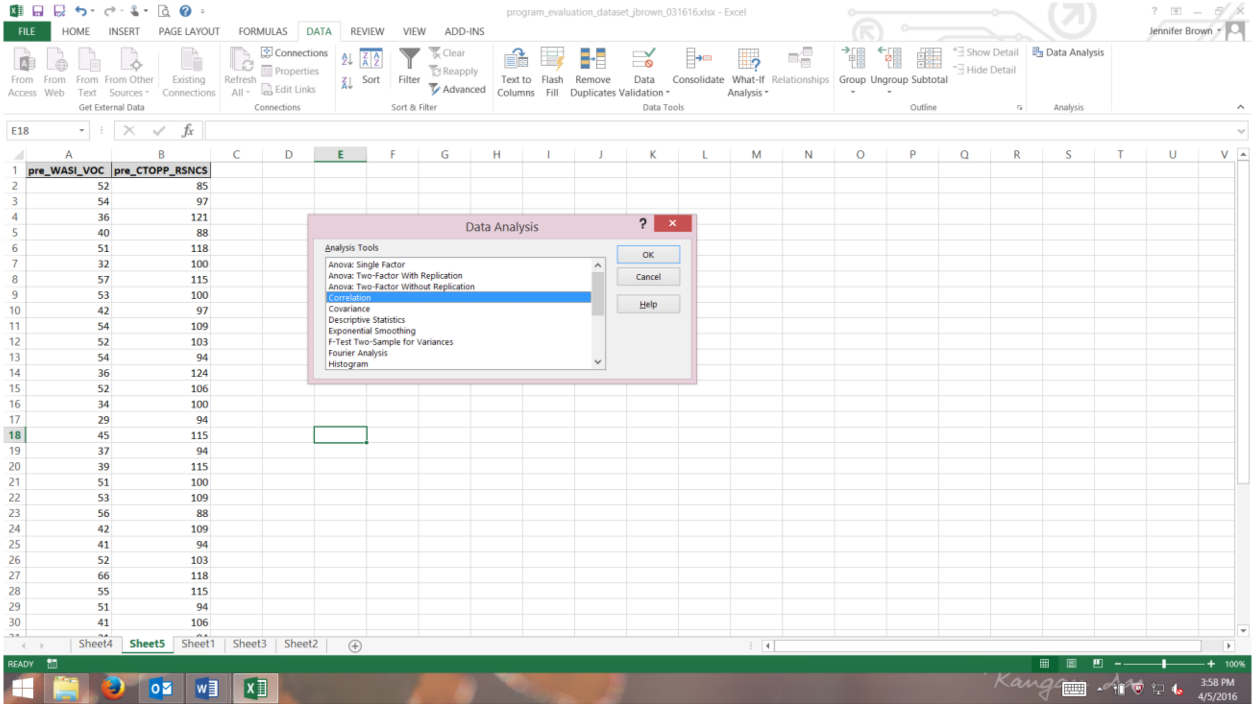
Between ± .10 and ± .30 – weak relationship
Between ± .30 and ± .50 – moderate relationship
± .50 and greater– strong relationship

How to Conduct a Pearson *r* in Excel

1. Copy all of the data within the desired column including the header for the first variable, *pre_WASI_VOC*. (For this example, you should copy from G1 to G43.)
2. Open a new worksheet by selecting the + in the lower left corner.
3. Paste the data into column A.
4. Copy all of the data within the desired column including the header for the second variable, *pre_CTOPP_RSNCS*. (For this example, you should copy from O1 to O43.)
5. Paste the data into column B.



6. Select the **Data** tab.
7. Select "Data Analysis".
8. In the pop-up window, select "Correlation".
9. Select **OK**.

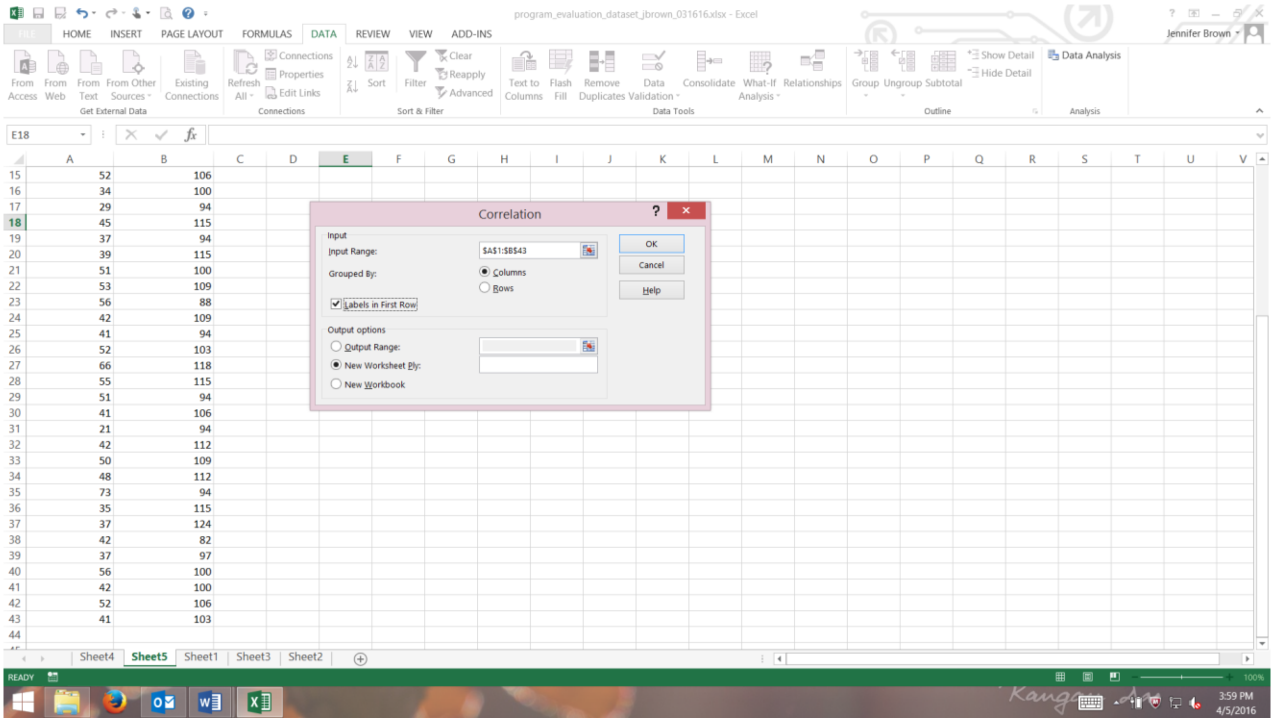


10. Click inside the "Input Range" box.

11. Highlight all of the data in the *pre_WASI_VOC* column and in the *pre_ CTOPP_RSNCS* column. (For this example, you should highlight from A1 to B43.)

12. Check the box beside "Labels in First Row".

13. Make sure the radial beside "New Worksheet Ply" is checked under "Output".



14. Select **OK**. (A new worksheet will open with the output.)

Excel Output

	pre_WASI_VOC	*pre_CTOPP_RSNCS*
pre_WASI_VOC	1	
pre_CTOPP_RSNCS	0.003241811	1

A Pearson *r* was conducted to determine the relationship between the pre-test scores on the Wechsler's Vocabulary subtest and the CTOPP Rapid Symbolic Naming subtest. There was not a relationship between these two variables (*r* = .00).

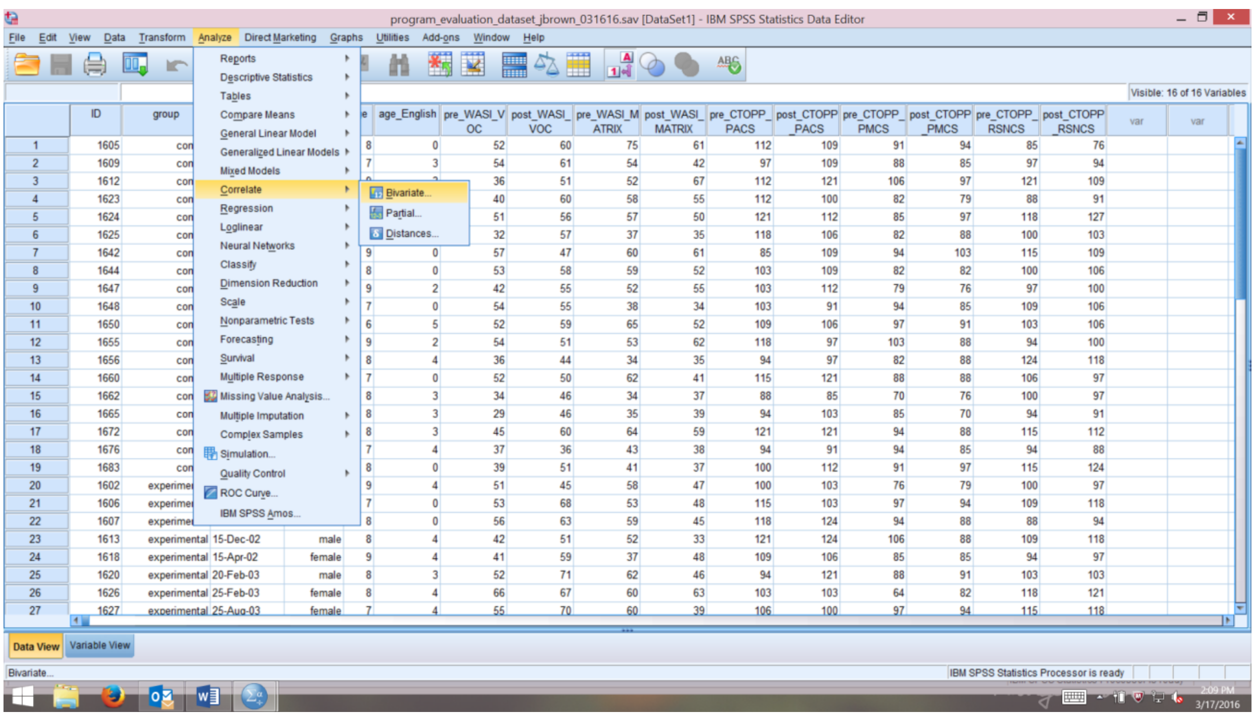
Why?

Why are we analyzing these variables with a Pearson *r*?

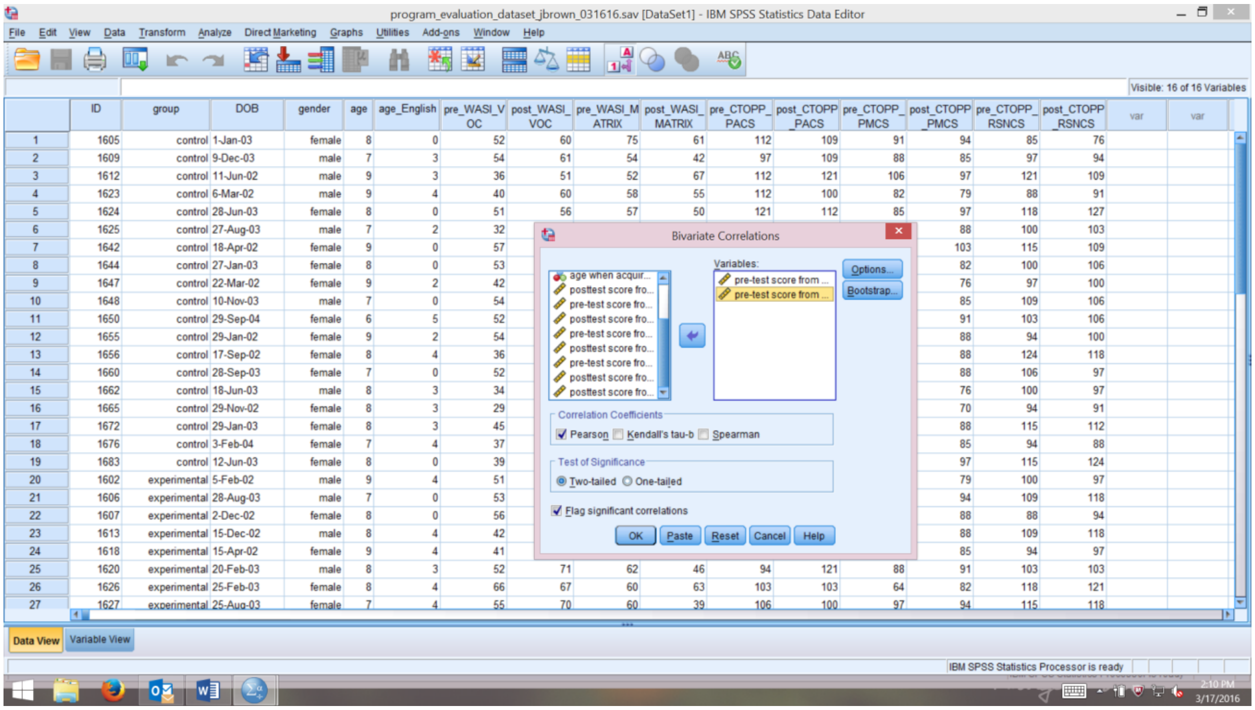
First, both of these variables are continuous. Second, we are conducting the Pearson *r* to determine if a relationship exists. Remember, a relationship does not mean causation. If a relationship exists, the Pearson *r* will indicate the strength of the relationship.

How to Conduct a Pearson *r* in SPSS

1. Analyze → Correlate → Bivariate



2. Select and move the following variables into the box: *pre_WASI_VOC* and *pre_CTOPP_RSNCS*. (*Note:* Highlight using left mouse and select the arrow icon.)



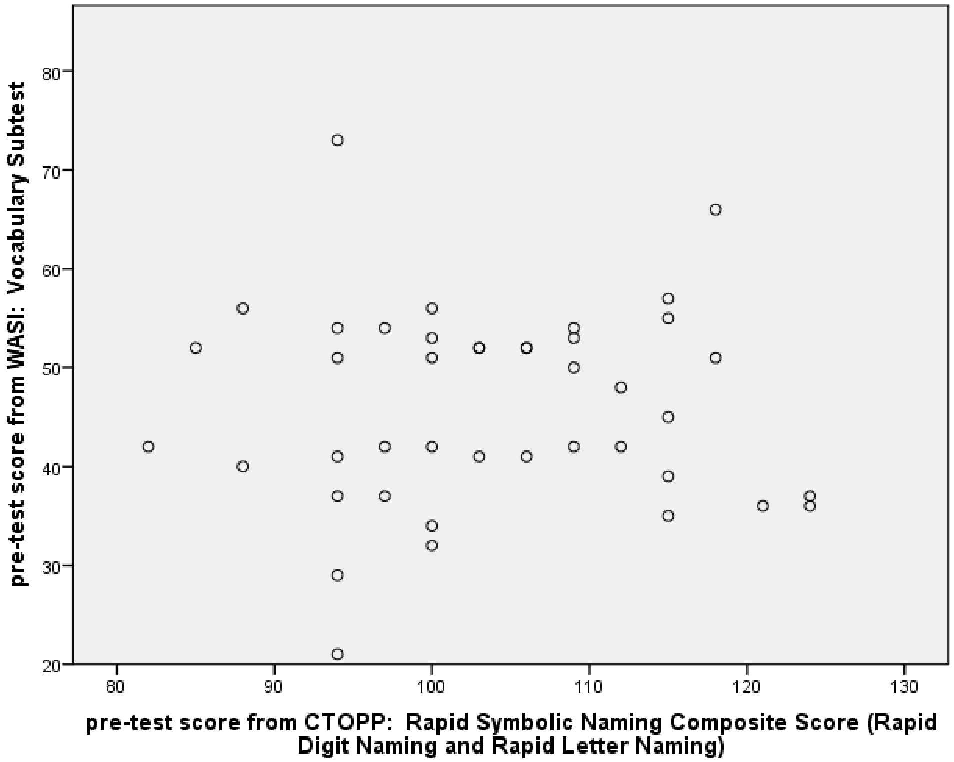
3. Select **OK**. (The analysis will appear on the output screen.)

SPSS Output

Correlations

		pre-test score from WASI: Vocabulary Subtest	pre-test score from CTOPP: Rapid Symbolic Naming Composite Score (Rapid Digit Naming and Rapid Letter Naming)
pre-test score from WASI: Vocabulary Subtest	Pearson Correlation	1	.003
	Sig. (2-tailed)		.984
	N	42	42
pre-test score from CTOPP: Rapid Symbolic Naming Composite Score (Rapid Digit Naming and Rapid Letter Naming)	Pearson Correlation	.003	1
	Sig. (2-tailed)	.984	
	N	42	42

There was not a relationship between the Wechsler's Vocabulary subtest score and the CTOPP Rapid Symbolic Naming Composite Score (r = .00; p = .98).



See the green rectangle on the output. Notice, the diagonal cells are the exact same numbers. If you view the scatterplot above, you will see the dots have no resemblance of a linear formation.

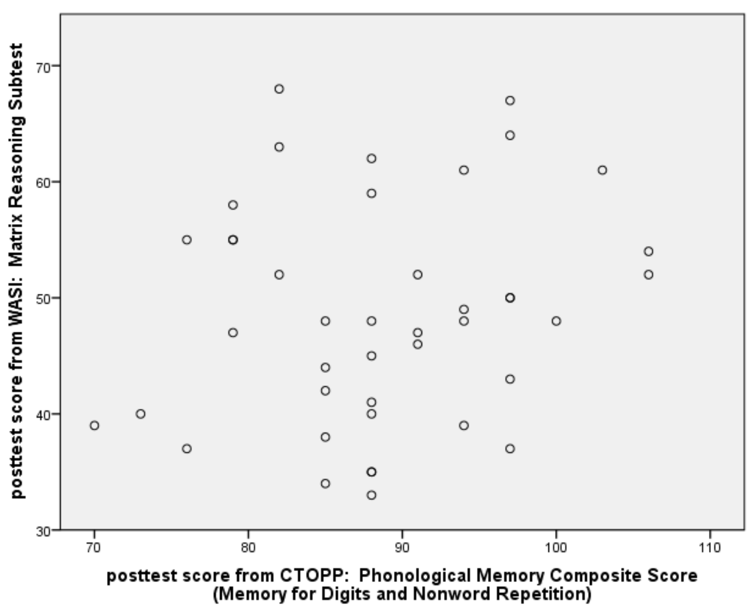
PRACTICE

You try with the *post_WASI_MATRIX and post_CTOPP_PMCS* variables. Compare your output the following output.

Correlations

		posttest score from WASI: Matrix Reasoning Subtest	posttest score from CTOPP: Phonological Memory Composite Score (Memory for Digits and Nonword Repetition)
posttest score from WASI: Matrix Reasoning Subtest	Pearson Correlation	1	.177
	Sig. (2-tailed)		.262
	N	42	42
posttest score from CTOPP: Phonological Memory Composite Score (Memory for Digits and Nonword Repetition)	Pearson Correlation	.177	1
	Sig. (2-tailed)	.262	
	N	42	42

A positive, weak relationship existed between the posttest scores from the Wechsler's Matrix Reasoning Subtest and the CTOPP Phonological Memory (r = .18; p = .26).



Examining the scatterplot above, you can see a slight linear formation, hence the weak relationship.

Inferential Statistics

Inferential statistics are used to analyze sample data, then the findings are generalized back the targeted population. For a categorical independent variable and a continuous dependent variable, you would conduct either a *t*-test or an Analysis of Variance (ANOVA), which are considered **parametric statistics**. A *t*-test and an ANOVA are calculated using the means and variance. **Variance** is the square of the standard deviation for each group and indicates the spread of the individual data. There are a few assumptions one makes before conducting these statistical procedures:

1) Population data would have a normal distribution;

2) The DV data are continuous.

Typically, you would conduct a *t*-test with sample sizes (i.e., less than 20 participants per group), and it can only be utilized with two levels or groups. Datasets with larger sample sizes and/or more than two levels or groups should be analyzed with an ANOVA. (*Note:* In the social sciences, we accept a criteria of .05 as statistically significant, meaning we are 95% confident that the results did not occur by chance. This criteria is referred to as alpha level.)

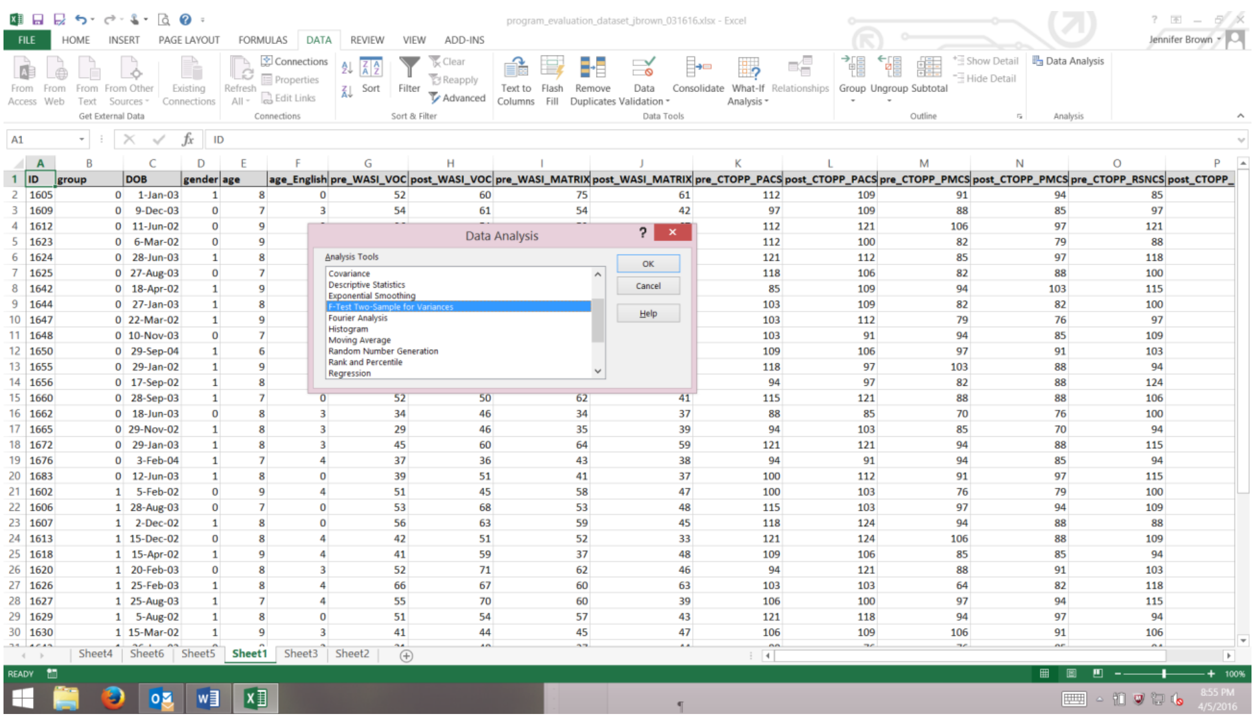
Levene's Test for Equality of Variances

Levene's Test for Equality of Variances is a statistical test to determine if the groups have equal variance among the scores, which is one of the above assumptions. This equal variance is referred to as **homogeneity of variance**. If the assumption was not met, then you would need to use the corrected *t*-test value. In Excel, you have to conduct the Levene's Test before conducting the *t*-test. Based on the statistically significance, you will select "…Assuming Equal Variances" or "…Assuming Unequal Variances". In SPSS, if the Levene's test value was statistically significant, then you would need to use the "Equal variances not assumed" row, which uses an algorithm to adjust the calculations to offset the unequal variance among groups. If the value was not statistically significant, then you would need to use the "Equal variances assumed" row.

How to Conduct a Levene's Test for Equality of Variances*

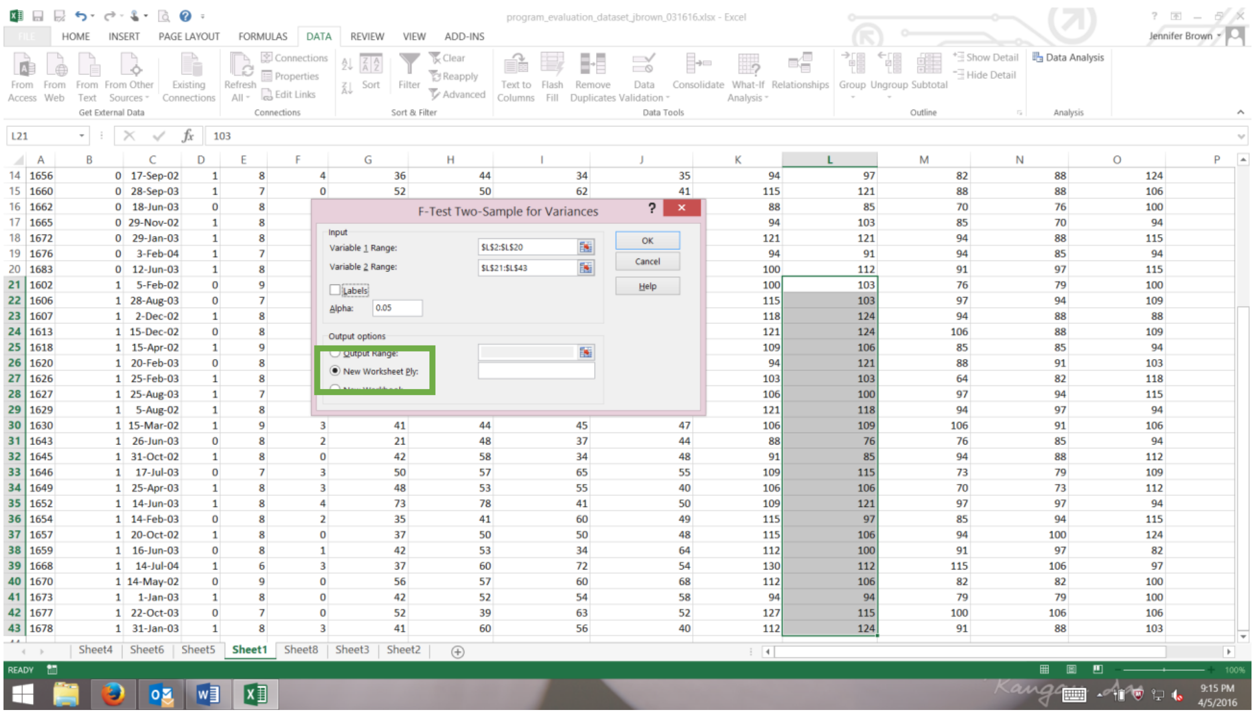
*You need to make sure the data are sorted by group before analyzing.

1. Select the **Data** tab.
2. Select "Data Analysis".
3. In the pop-up window, select "F-Test Two-Sample for Variances".



4. Select **OK**.
5. Click inside the "Variable 1 Range" box under "Input".
6. Highlight all of the data in *post_CTOPP_PACS* column for the control group. (For this example, you should highlight L2 through L20.)

7. Click inside the “Variable 2 Range” box under “Input”.
8. Highlight all of the data in *post_CTOPP_PACS* column for the experimental group. (For this example, you should highlight L21 through L43.)
9. Make sure the box beside “Labels” is NOT checked.
10. Make sure the “alpha” level is 0.05.
11. Make sure the radial beside “New Worksheet Ply” is checked under “Output”.



12. Select **OK**. (A new worksheet will open with the output.)

Excel Output

F-Test Two-Sample for Variances		
	Variable 1	*Variable 2*
Mean	105.8421053	107.3043478
Variance	106.4736842	154.4940711
Observations	19	23
Df	18	22
F	0.689176506	
P(F<=f) one-tail	0.213145951	
F Critical one-tail	0.461153892	

Levene's Test for Equality of Variances was not statistically significant (p = .21).

(*Note:* The *p*, or significance, value was above the alpha level of .05; therefore, we can assume equal variance exists among the groups.)

Why?

Why do we need to conduct Levene's Test for Equality of Variance?

One of the assumptions for inferential statistics is equal variance among groups. This analysis procedure allows you to determine if this assumption was met. We will discuss what to do if the assumption was not met within the upcoming pages.

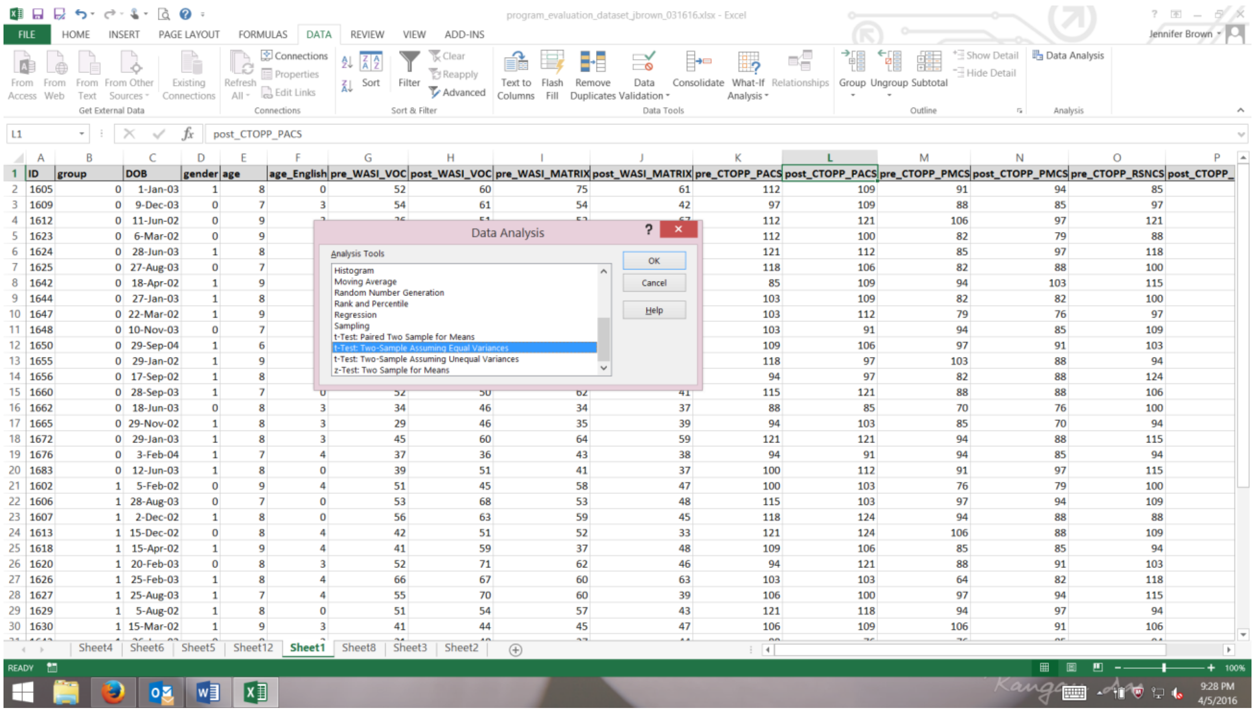
How to Conduct an Independent *t*-test in Excel*

*independent refers to the data being collected from unique units (e.g., participants).

1. Select the **Data** tab.

2. Select "Data Analysis".

3. In the pop-up window, select “t-Test Two-Sample Assuming Equal Variance”. (*Note*: You select this option based on the Levene’s Test value, which was not statistically significant.)



4. Select **OK**.

5. Click inside the “Variable 1 Range” box under “Input”. (*Note*: You need to make sure the data are sorted by group before analyzing.)

6. Highlight all of the data in *post_CTOPP_PACS* column for the control group. (For this example, you should highlight L2 through L20.)

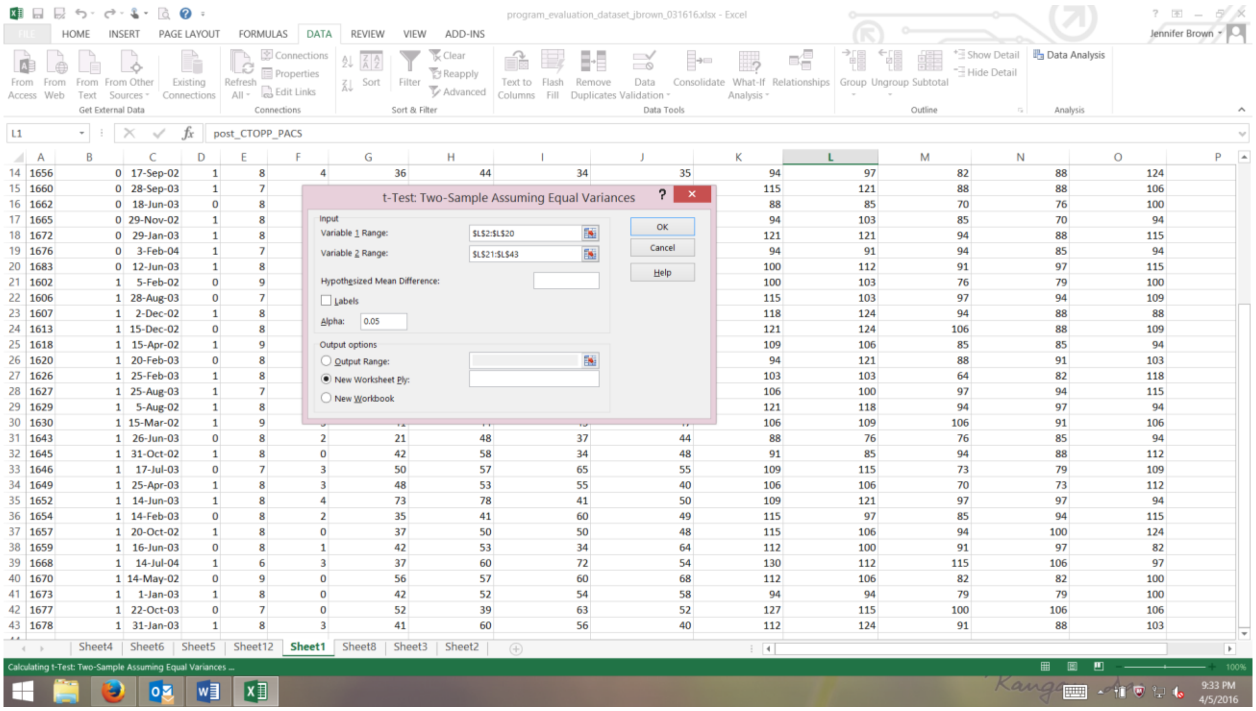
7. Click inside the “Variable 2 Range” box under “Input”.

8. Highlight all of the data in *post_CTOPP_PACS* column for the experimental group. (For this example, you should highlight L21 through L43.)

9. Make sure the box beside “Labels” is NOT checked.

10. Make sure the “alpha” level is 0.05.

11. Make sure the radial beside “New Worksheet Ply” is checked under “Output”.



12. Select **OK**. (A new worksheet will open with the output.)

Excel Output

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	105.8421	107.3043
Variance	106.4737	154.4941
Observations	19	23
Pooled Variance	132.8849	
Hypothesized Mean Difference	0	
df	40	
t Stat	-0.40916	
P(T<=t) one-tail	0.342301	
t Critical one-tail	1.683851	
P(T<=t) two-tail	0.684602	
t Critical two-tail	2.021075	

There was not a statistically significant difference between the groups for the posttest score of the CTOPP Phonological Awareness subtest, $t(40) = -0.41$; $p = .68$.

The independent *t*-test value is highlighted with the green rectangle. Notice, the results include the *t*-test value, degrees of freedom, and the significant value, and an italicized lowercase *t* is the symbol for a *t*-test. (*Note*: For a two-tailed *t*-test, meaning the group difference can go either direction, the difference is considered statistically significant if the *t*-test value is in the top 2.5% or bottom 2.5% of its probability distribution, which results in a *p*-value less than .05. For a one-tailed t-test, meaning the group difference will be unidirectional, the difference is considered statistically significant if the *t*-test value is in the top 5% or bottom 5% of its probability distribution, but you must select the direction prior to conducting the statistic.)

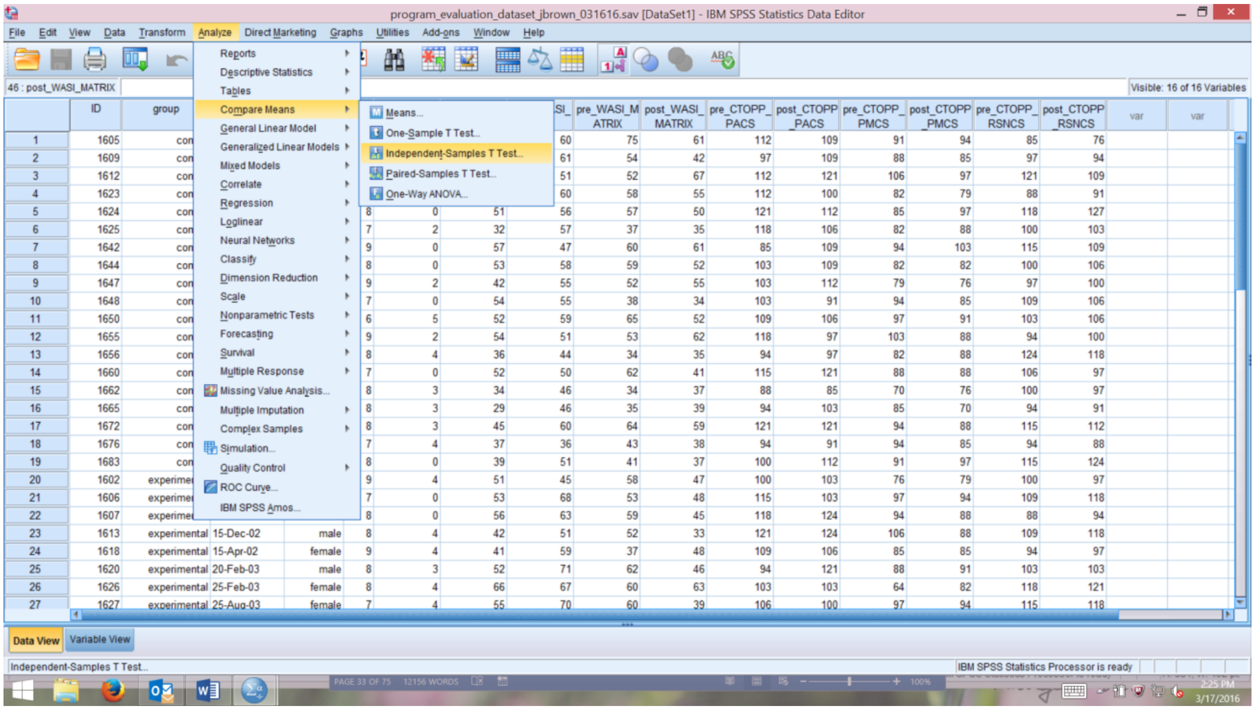
Why?

Why did we select an independent *t*-test to analyze these data?

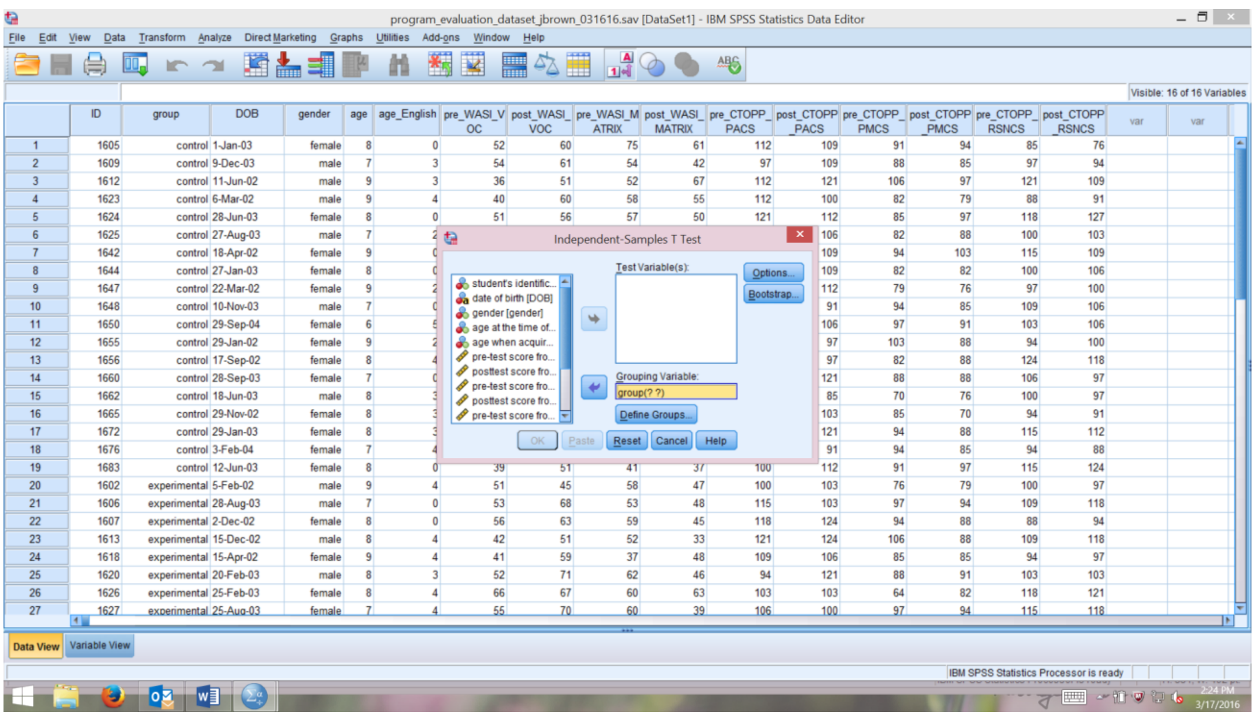
First, the independent *t*-test requires a categorical IV with two levels or groups and a continuous DV. Group served as the IV, which contained a control group and an experimental group, and phonological awareness served as the DV as measured by the posttest score of the CTOPP: Phonological Awareness subtest. By conducting this statistic, you are determining if a difference exists between the posttest scores for the control and experimental groups.

How to Conduct an Independent *t*-test in SPSS

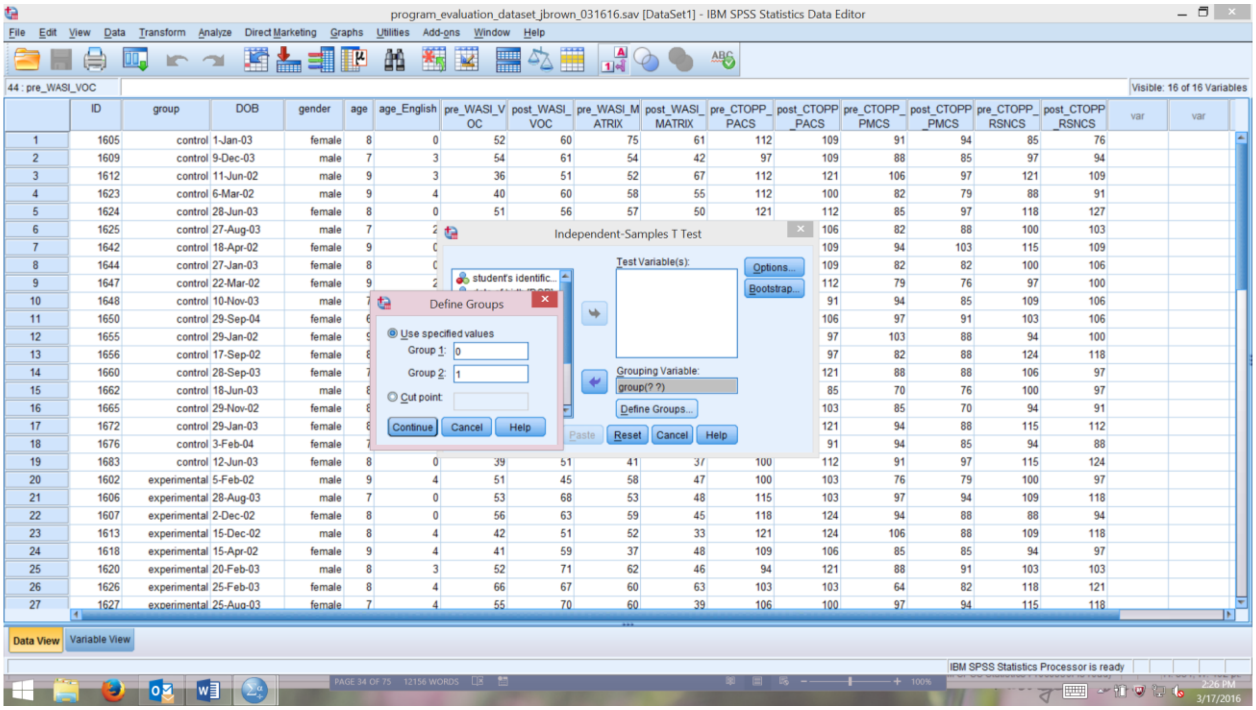
1. Analyze → Compare Means → Independent Samples T Test



2. Select the *group* variable. Use the arrow to move it to the **grouping variable** box.

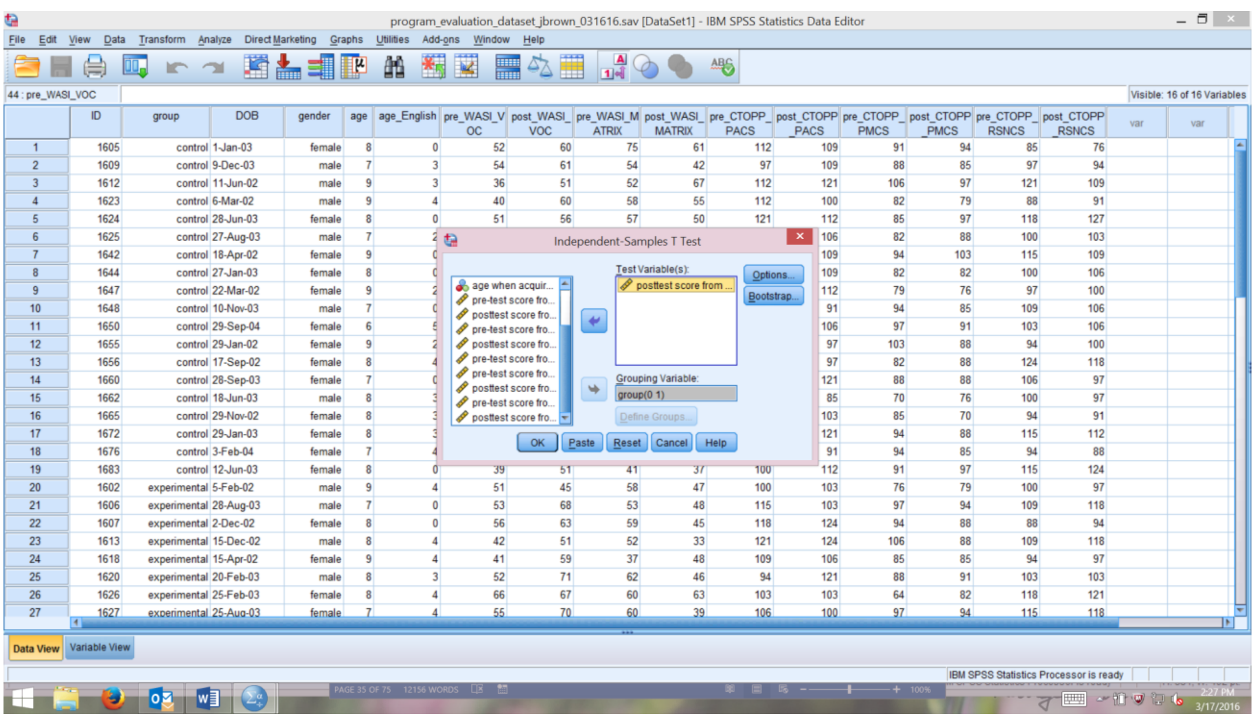


3. Select **Define Groups**. In the **Group 1** box, type “0”. In the **Group 2** box, type “1”.



4. Select **Continue**.

5. Select the *post_CTOPP_PACS* variable on the left side and use the arrow to move it to the **test variable(s)** box.



6. Select **OK**. (The analysis will appear on the output screen.)

SPSS Output

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
posttest score from CTOPP: Phonological Awareness Composite Score (Elision, Blending Words, and Phoneme Isolation)	control	19	105.84	10.319	2.367
	experimental	23	107.30	12.430	2.592

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
posttest score from CTOPP: Phonological Awareness Composite Score (Elision, Blending Words, and Phoneme Isolation)	Equal variances assumed	.455	.504	-.409	40	.685	-1.462	3.574	-8.685	5.761
	Equal variances not assumed			-.417	39.996	.679	-1.462	3.510	-8.557	5.632

There was not a statistically significant difference between the groups for the posttest score of the CTOPP Phonological Awareness subtest, *t*(40) = -0.41; *p* = .69.

After conducting a *t*-test, you need to examine the Levene's Test for Equality of Variances results. (See the purple oval.) The *F* value was 0.46 and the *p* value, or significance, value was .50, meaning there was not a significant difference, which is what you want to see. For this example, you should use "equal variances assumed" to report the results, which is the first row. The independent *t*-test value is highlighted with the green rectangle. Notice, the results include the *t*-test value, degrees of freedom, and the significant value. The *p*, or significance, value was above the alpha level of .05; therefore, the difference between the groups was not statistically significant. SPSS also gives you the descriptives for each group.

PRACTICE

You try with the *post_WASI_MATRIX* variable. Compare your output the following output.

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
posttest score from WASI: Matrix Reasoning Subtest	control	19	48.00	11.035	2.532
	experimental	23	49.09	8.393	1.750

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
posttest score from WASI: Matrix Reasoning Subtest	Equal variances assumed	5.695	.022	-.363	40	.719	-1.087	2.998	-7.147	4.973
	Equal variances not assumed			-.353	33.125	.726	-1.087	3.078	-7.348	5.174

There was not a statistically significant difference between the groups for the posttest score of the Wechsler's Matrix Reasoning Subtest, *t*(33.125) = -0.35; *p* = .73.

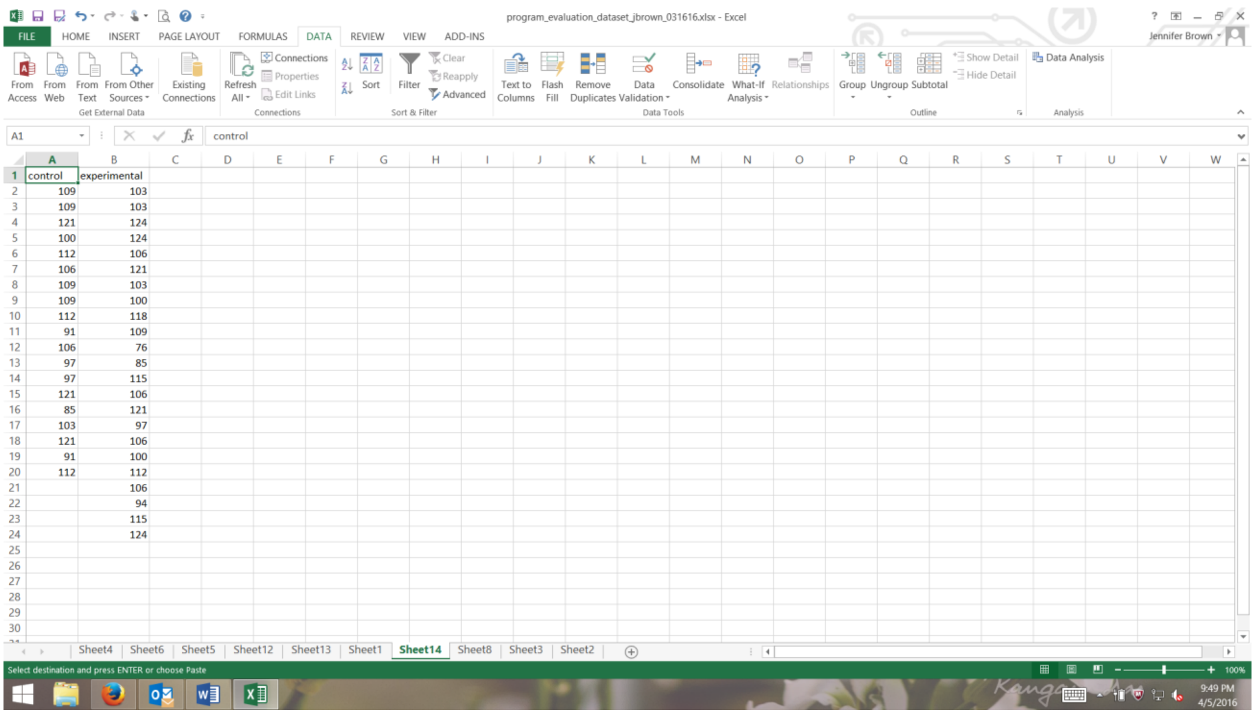
With this variable, you cannot assume equal variance. (See the Levene's *F* value. Also, notice the difference in the two standard deviations, which is where the problem lies.) Therefore, you should use the bottom row for "Equal variances not assumed". (*Note*: An italicized lowercase *t* is the symbol for a *t*-test.) The *p*, or significance, value was above the alpha level of .05; therefore, the difference between the groups was not statistically significant.

How to Conduct an One-Way ANOVA in Excel*

*You need to make sure the data are sorted by group before analyzing.

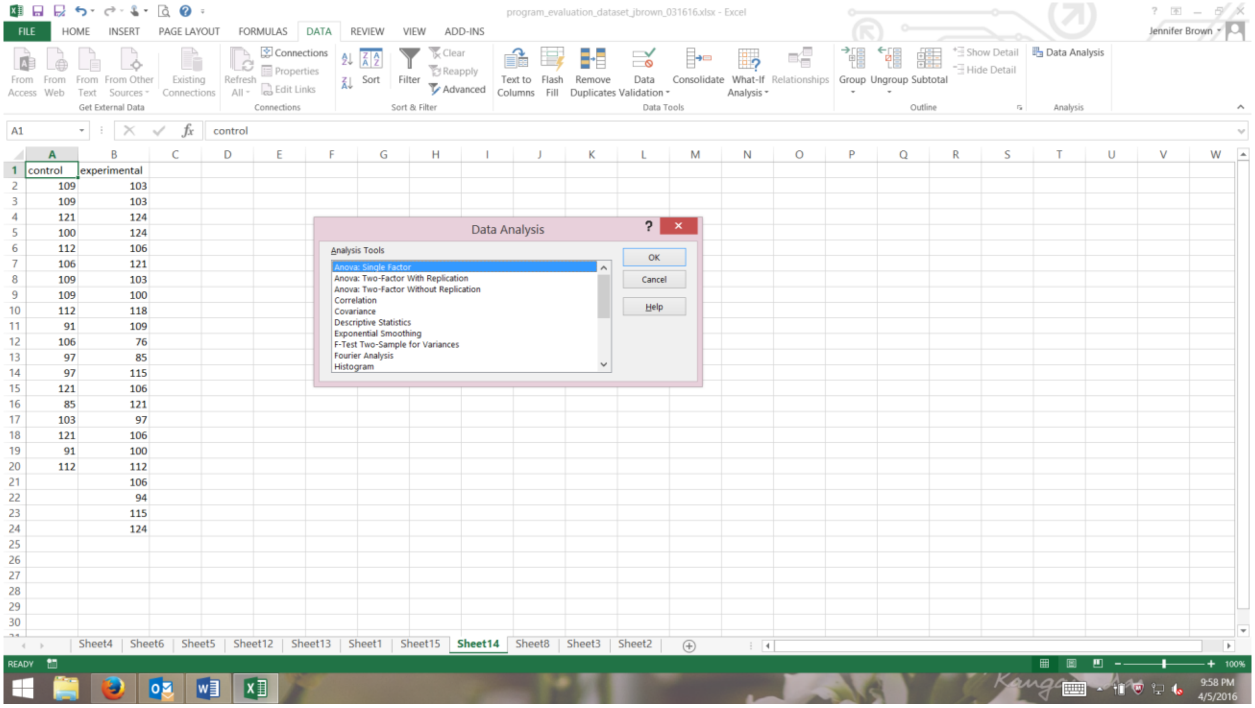
1. Open a new worksheet by selecting the **+** in the lower left corner.
2. Copy all of the data in *post_CTOPP_PACS* column for the control group. (For this example, you should highlight L2 through L20.)
3. Type the variable label in cell A1 of the new worksheet. (For this example, type "control".)

4. Paste the data into column A beginning in cell A2.
5. Copy all of the data in *post_CTOPP_PACS* column for the control group. (For this example, you should highlight L21 through L43.)
6. Type the variable label in cell B1 of the new worksheet. (For this example, type “experimental”.)
7. Paste the data into column B beginning in cell B2.



control	experimental
109	103
109	103
121	124
100	124
112	106
106	121
109	103
109	100
112	118
91	109
106	76
97	85
97	115
121	106
85	121
103	97
121	106
91	100
112	112
	106
	94
	115
	124

8. Select the **Data** tab.
9. Select “Data Analysis”.
10. In the pop-up window, select “Anova: Single Factor”.



11. Select **OK**.

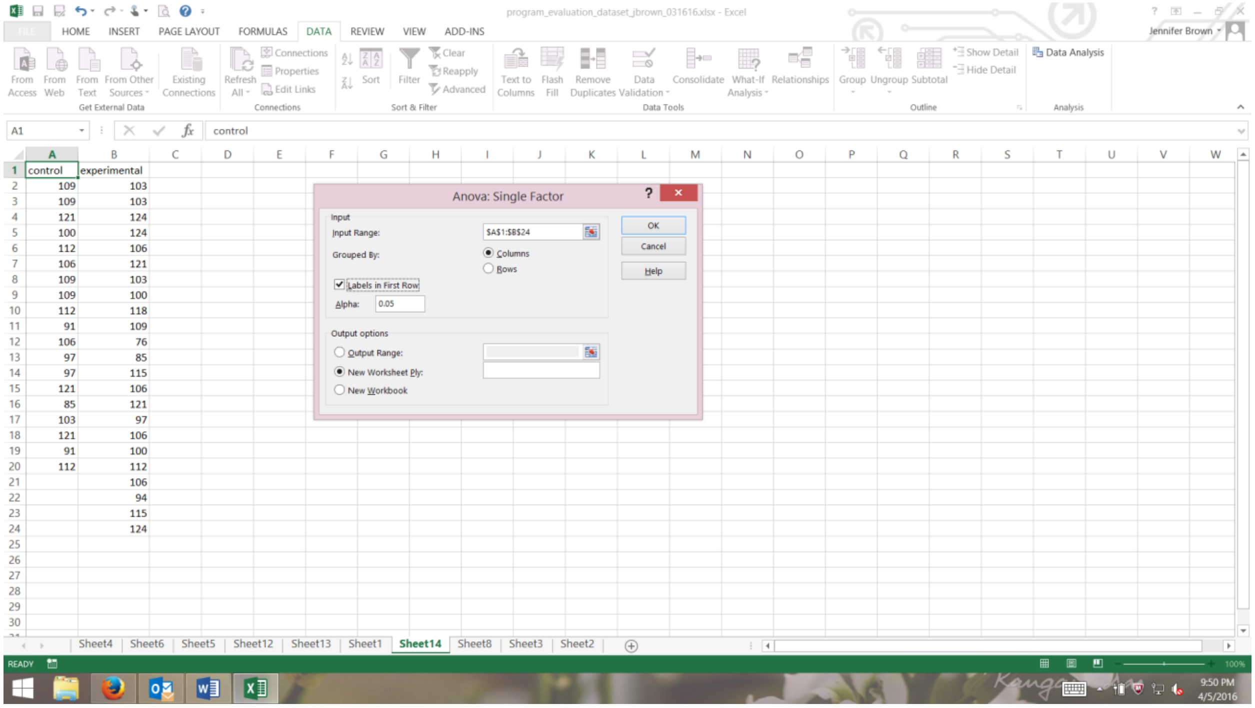
12. Click inside the “Input Range” box.

13. Highlight all of the data in the “control” column and in the “experimental” column. (For this example, you should highlight from A1 to B24.)

14. Check the box beside “Labels in First Row”.

15. Make sure the “alpha” level is 0.05.

16. Make sure the radial beside “New Worksheet Ply” is checked under “Output”.



17. Select **OK**. (A new worksheet will open with the output.)

Excel Output

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
control	19	2011	105.8421	106.4737
experimental	23	2468	107.3043	154.4941

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	22.24698	1	22.24698	0.167415	0.684602	4.084746
Within Groups	5315.396	40	132.8849			
Total	5337.643	41				

From this output, the same results were found as the independent samples *t*-test. (See the green rectangle.) If you square the *t*-test value, which was -0.409, you will get 0.167, which is the same as the *F* value. (*Note:* An italicized uppercase *F* is the symbol for an ANOVA.) The notation with the between groups and within groups' degrees of freedom would look like the following: $F(1, 40) = 0.17$; $p = .69$. The *p*, or significance, value was above the alpha level of .05; therefore, the difference between the groups was not statistically significant.

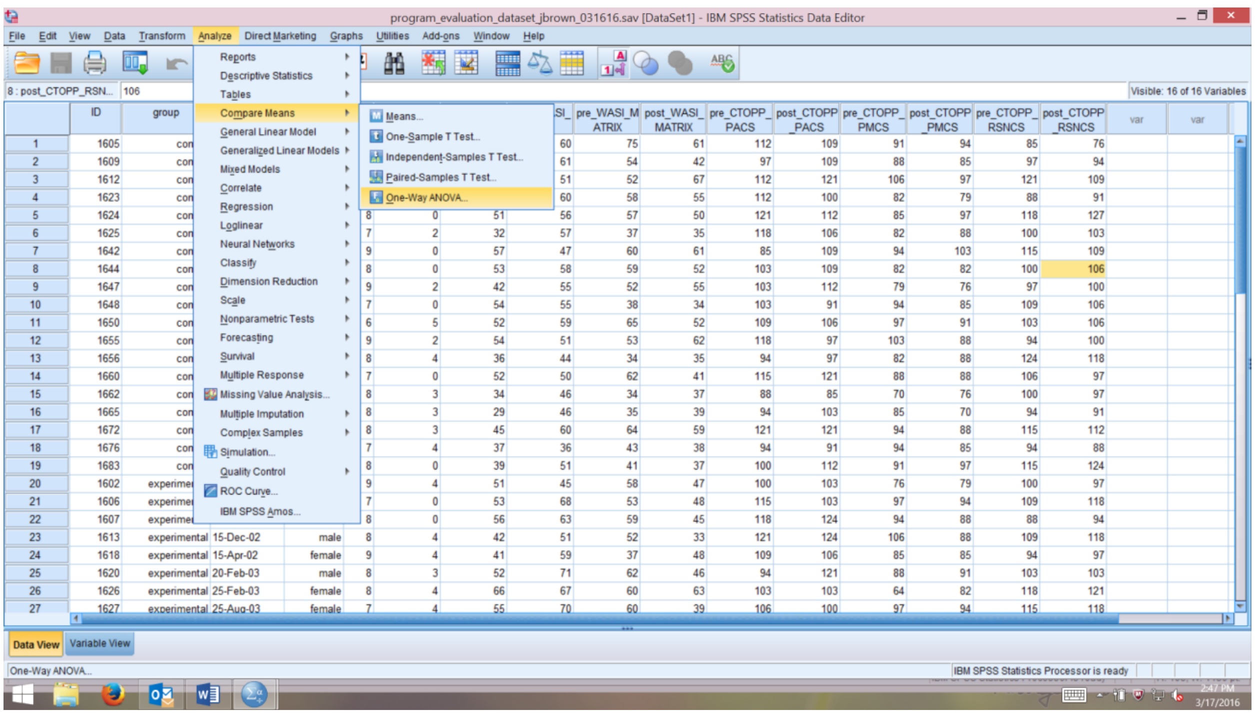
Why?

Why did we select a One-Way ANOVA to analyze these data?

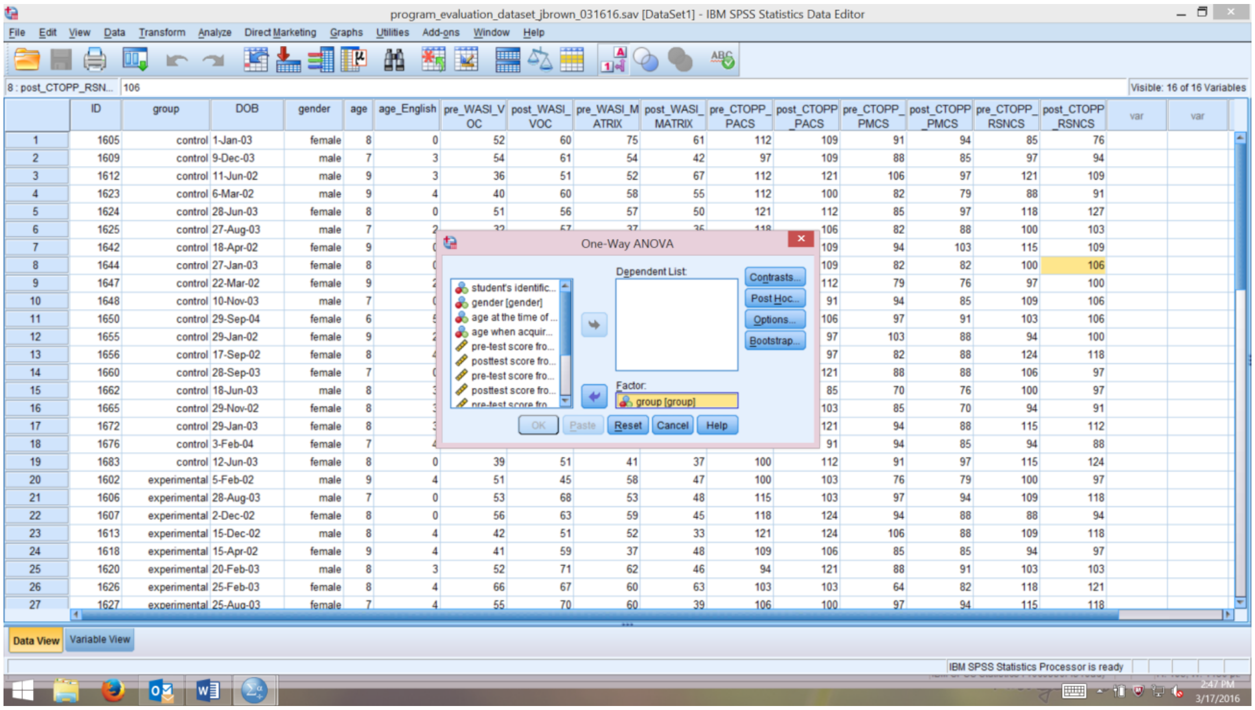
An ANOVA requires a categorical IV and continuous DV. From the previous example, you can see that you get the same results from the *t*-test and ANOVA statistics. The decision between the two statistics should be based on sample size and the number of levels or groups within the IV. If the sample has less than 20 participants, then a *t*-test would be appropriate. If there was not equal variance between the groups, you can utilize the corrected *t*-test. If the number of levels or groups within the IV exceeds two, then you must use an ANOVA. Based on my experience as a program evaluator, I tend to use the *t*-test more often.

How to Conduct a One-Way ANOVA in SPSS

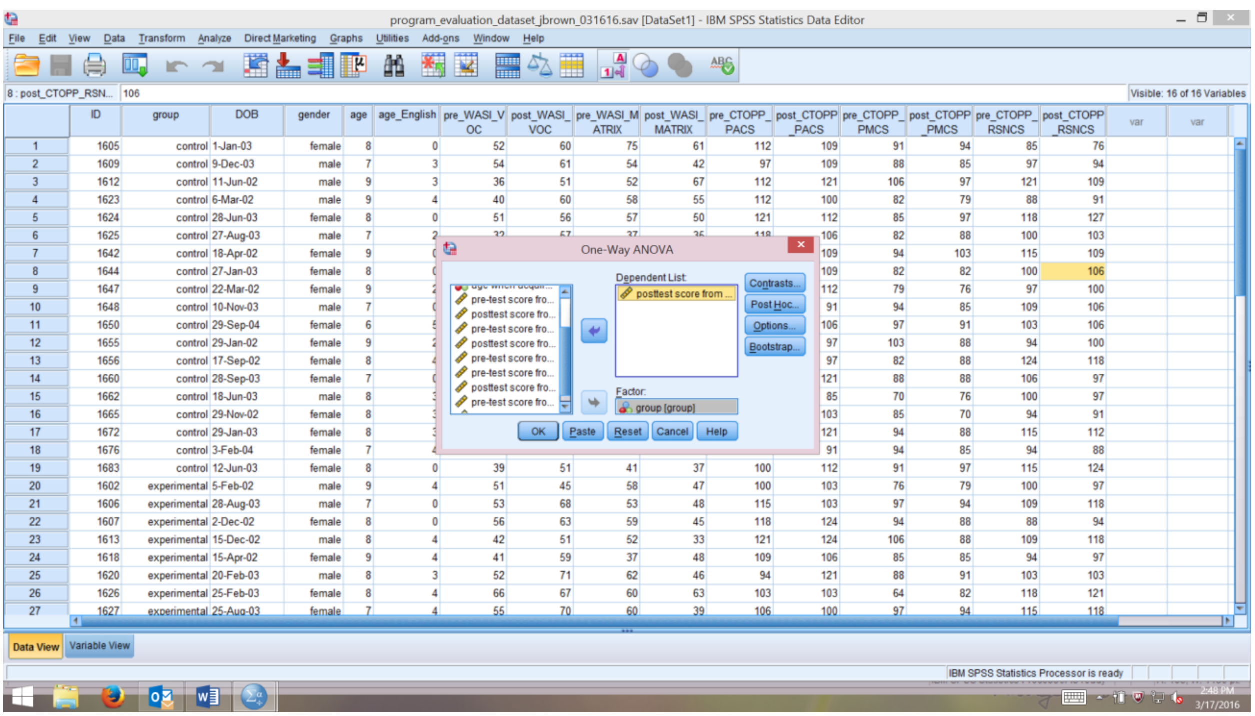
1. Analyze → Compare Means → One-Way ANOVA



2. Select the "group" variable. Use the arrow to move it to the **factor** box.



3. Select the "post_CTOPP_PACS" variable on the left side and use the arrow to move it to the **Dependent List** box.



4. Select **OK**. (The analysis will appear on the output screen.)

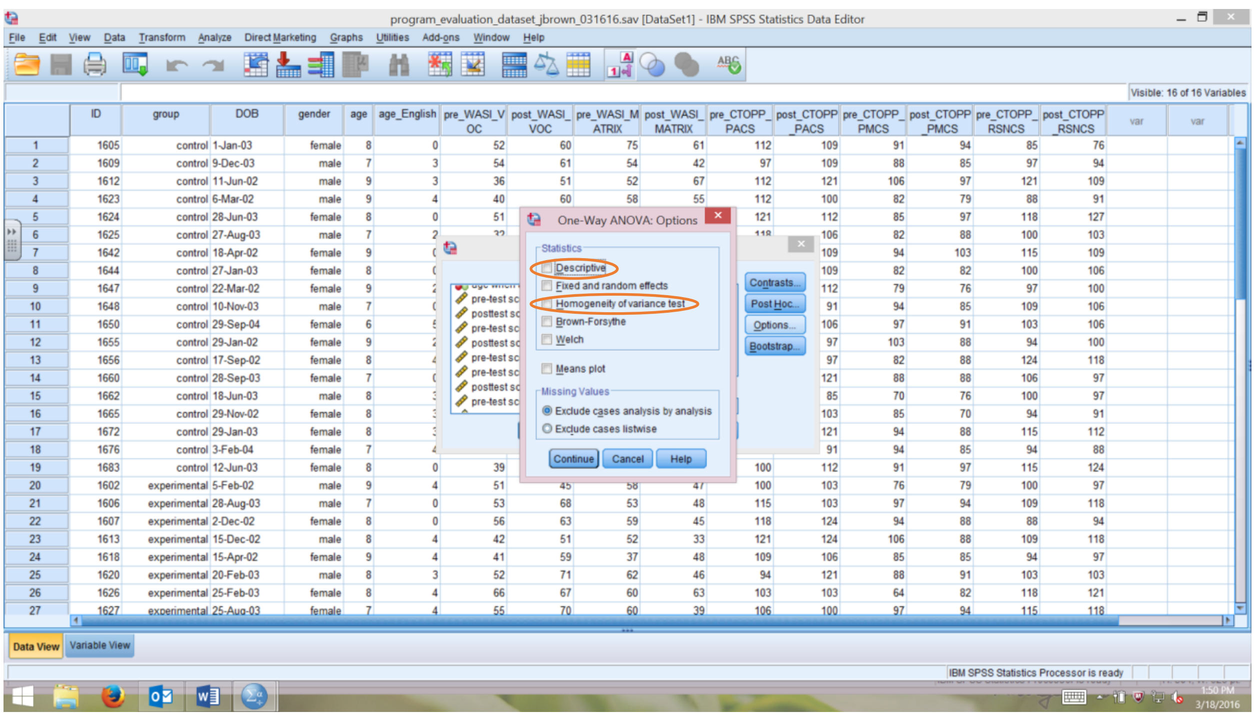
SPSS Output

ANOVA

posttest score from CTOPP: Phonological Awareness Composite Score (Elision, Blending

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	22.247	1	22.247	.167	.685
Within Groups	5315.396	40	132.885		
Total	5337.643	41			

As you can see from the output, the same results were found as the independent samples *t*-test. (See the green rectangle.) Notice, the significance values are the exact same. Unfortunately, with an ANOVA, you would need to select the “Options” button to obtain the Levene’s Test results and descriptives. (See the orange ovals.)



PRACTICE

You try with the *post_CTOPP_PMCS* variable. Compare your output the following output.

Descriptives

posttest score from CTOPP: Phonological Memory Composite Score (Memory for Digits and Nonword Repetition)

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
control	19	87.21	8.357	1.917	83.18	91.24	70	103
experimental	23	89.70	8.668	1.807	85.95	93.44	73	106
Total	42	88.57	8.517	1.314	85.92	91.23	70	106

Test of Homogeneity of Variances

posttest score from CTOPP: Phonological Memory Cc

Levene Statistic	df1	df2	Sig.
.199	1	40	.658

ANOVA

posttest score from CTOPP: Phonological Memory Composite Score (Memory for Digits an

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	64.258	1	64.258	.883	.353
Within Groups	2910.027	40	72.751		
Total	2974.286	41			

A One-Way Analysis of Variance (ANOVA) was conducted to determine if a difference existed between the control and experimental group on the posttest of CTOPP: Phonological Memory subtest. For the control group, the mean posttest score was 87.21 with a standard deviation of 8.36. For the experimental group, the mean posttest score was 8.67 with a standard deviation of 1.81. There was not statistically significant difference between groups, $F(1, 40) = 0.88$; $p = .35$.

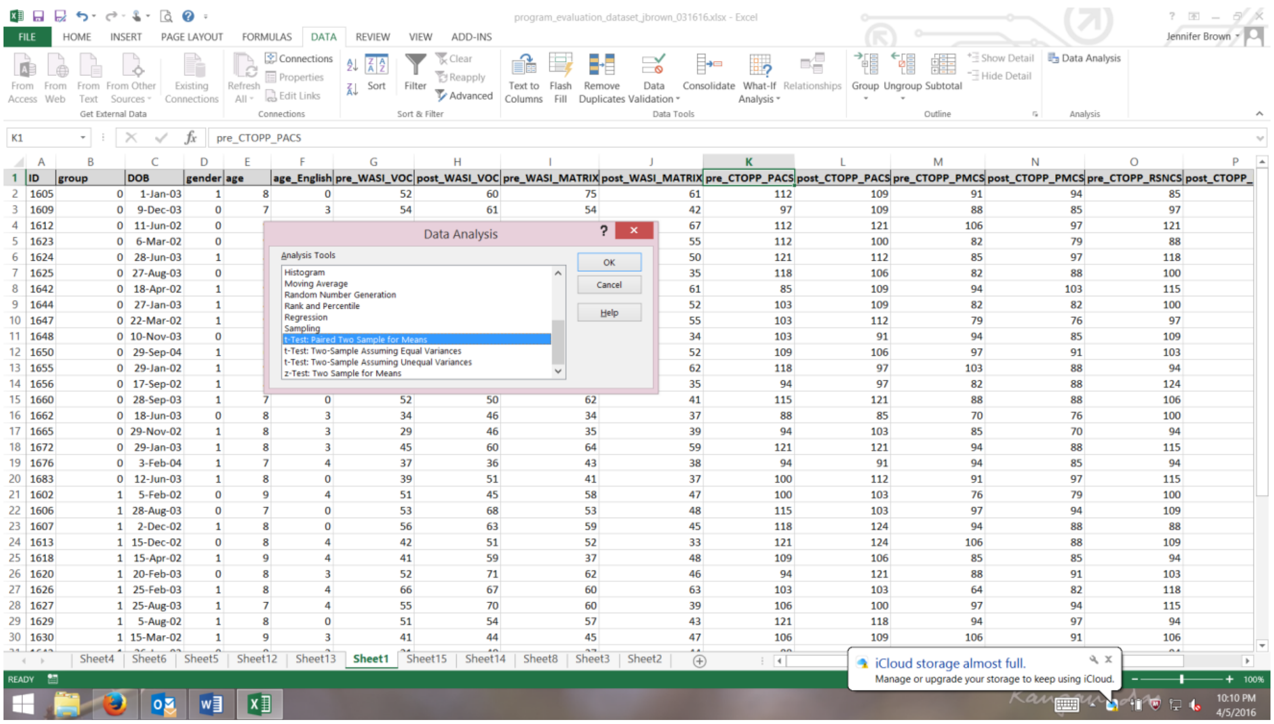
How to conduct a paired-samples *t*-test in Excel*

*This statistic is for dependent samples, meaning the data were collected from same unit (e.g., participants). In this example, you are analyzing pre-test and posttest data.

1. Select the **Data** tab.

2. Select "Data Analysis".

3. In the pop-up window, select "t-Test: Paired Two Sample for Means".



4. Select **OK**.

5. Click inside the "Variable 1 Range" box under "Input".

6. Highlight all of the data in *pre_CTOPP_PACS* including the heading. (For this example, you should highlight K1 through K43.)

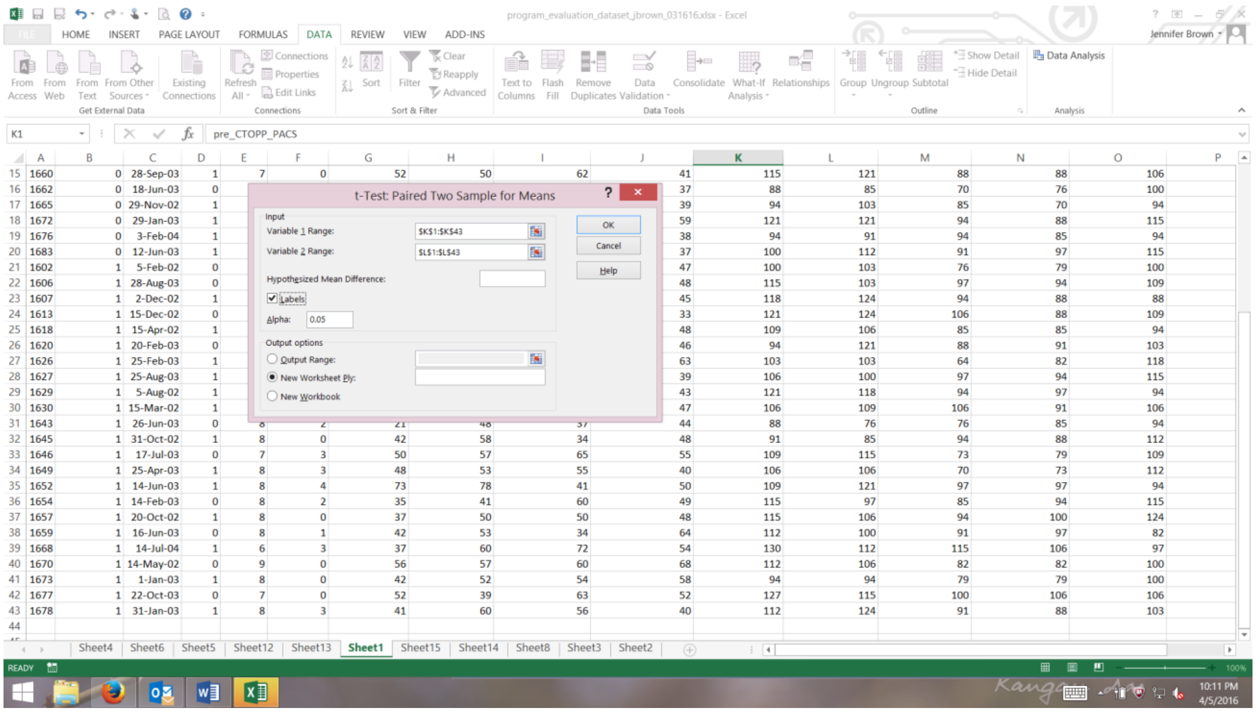
7. Click inside the "Variable 2 Range" box under "Input".

8. Highlight all of the data in *post_CTOPP_PACS* including the heading. (For this example, you should highlight L1 through L43.)

9. Check the box beside "Labels".

10. Make sure the “alpha” level is 0.05.

11. Make sure the radial beside “New Worksheet Ply” is checked under “Output”.



12. Select **OK**. (A new worksheet will open with the output.)

Excel Output

t-Test: Paired Two Sample for Means

	pre_CTOPP_PACS	*post_CTOPP_PACS*
Mean	107.4285714	106.6428571
Variance	123.4703833	130.1864111
Observations	42	42
Pearson Correlation	0.548355344	
Hypothesized Mean Difference	0	
df	41	
t Stat	0.475636527	
P(T<=t) one-tail	0.318428933	
t Critical one-tail	1.682878002	
P(T<=t) two-tail	0.636857866	
t Critical two-tail	2.01954097	

There was not a statistically significant change from pre-test to posttest with the CTOPP Phonological Awareness subtest, $t(41) = 0.48$; $p = .64$. (See the green rectangle in the output.) If you examine the means for each group, you will see that the means were fairly similar. (See the purple rectangle.) The *p*, or significance, value was above the alpha level of .05; therefore, the difference between the groups was not statistically significant. (*Note*: An italicized lowercase *t* is the symbol for a *t*-test.)

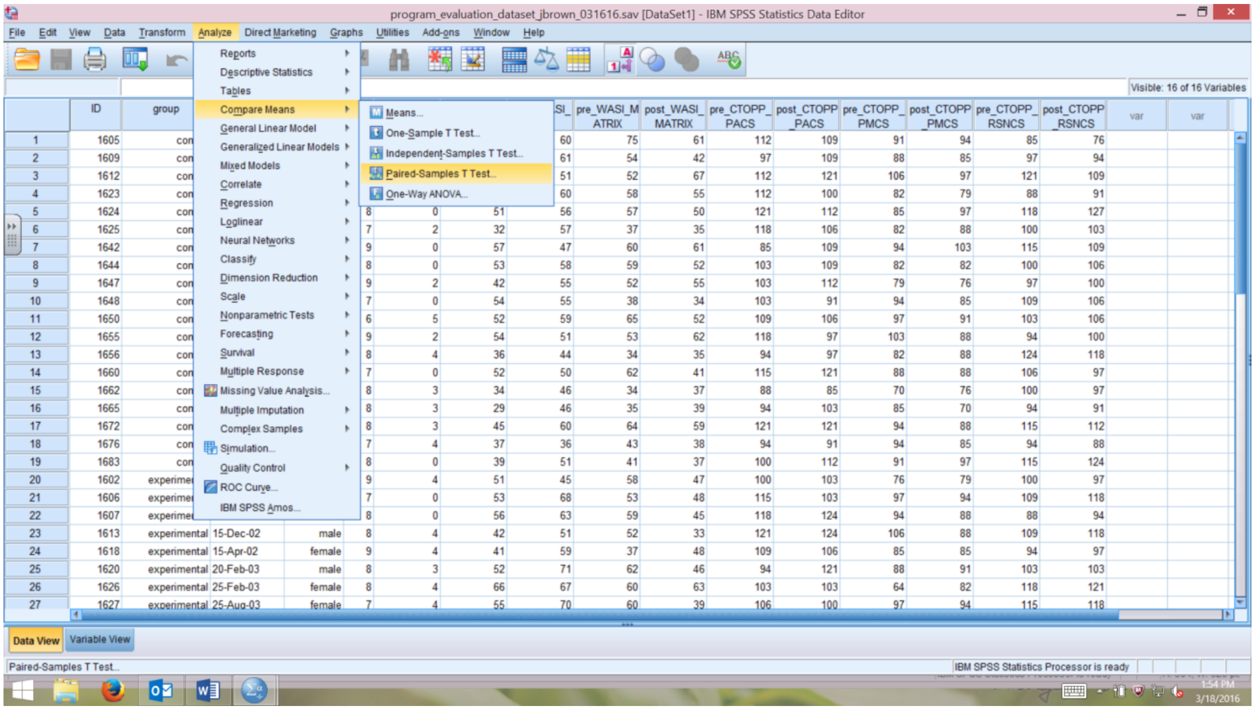
Why?

Why did we select a paired-samples *t*-test to analyze these data?

An assumption for an independent *t*-test and ANOVA is independence among the cases. Sometimes, we have the same participant contribute multiple data points across time. If you want to determine if a change occurred among these data points, you must the paired-samples *t*-test. With this previous example, the same participant took the pre-test and posttest. This statistic still requires a categorical IV (time) and a continuous DV (phonological awareness as measured by the CTOPP: Phonological Awareness subtest). See Program Evaluation Report Example #2 in Appendix E for application.

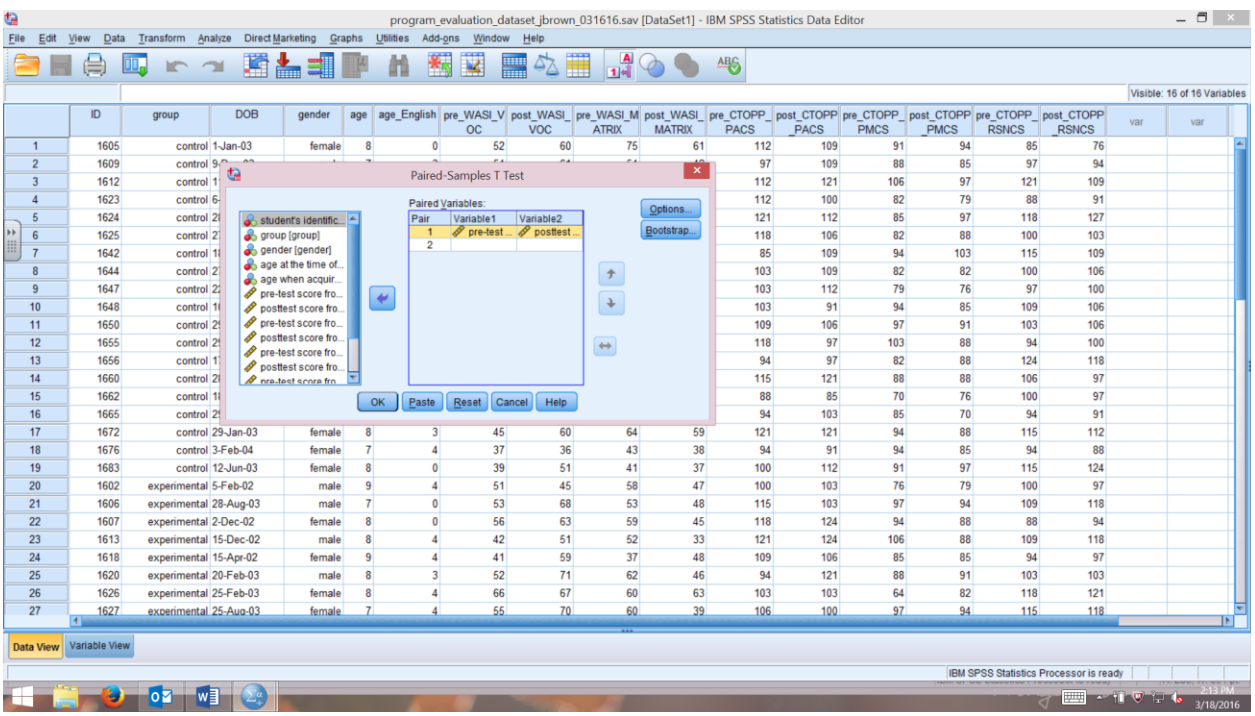
How to Conduct a Paired-Samples *t*-test in SPSS

1. Analyze → Compare Means → Paired-Samples T Test



2. Select the *pre_CTOPP_PACS* variable. Use the arrow to move it to the **Variable 1** box.

3. Select the *post_CTOPP_PACS* variable. Use the arrow to move it to the **Variable 2** box.



4. Select **OK**. (The analysis will appear on the output screen.)

SPSS Output

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	pre-test score from CTOPP: Phonological Awareness Composite Score (Elision, Blending Words, and Phoneme Isolation)	107.43	42	11.112	1.715
	posttest score from CTOPP: Phonological Awareness Composite Score (Elision, Blending Words, and Phoneme Isolation)	106.64	42	11.410	1.761

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	pre-test score from CTOPP: Phonological Awareness Composite Score (Elision, Blending Words, and Phoneme Isolation) - posttest score from CTOPP: Phonological Awareness Composite Score (Elision, Blending Words, and Phoneme Isolation)	.786	10.706	1.652	-2.550	4.122	.476	41	.637

There was not a statistically significant change from pre-test to posttest with the CTOPP Phonological Awareness subtest, $t(41) = 0.48$; $p = .64$. (See the green rectangle.) If you examine the descriptives, you will see that the means are fairly similar. (See the purple rectangle.) The *p*, or significance, value was above the alpha level of .05; therefore, the difference between the groups was not statistically significant.

PRACTICE

You try with the *pre_WASI_VOC and post_WASI_VOC* variables. Compare your output the following output.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	pre-test score from WASI: Vocabulary Subtest	46.02	42	10.067	1.553
	posttest score from WASI: Vocabulary Subtest	54.79	42	8.794	1.357

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	pre-test score from WASI: Vocabulary Subtest - posttest score from WASI: Vocabulary Subtest	-8.762	8.941	1.380	-11.548	-5.976	-6.351	41	.000

There was a statistically significant change from pre-test to posttest on the Wechsler's Vocabulary Subtest, $t(41) = -6.35$; $p = .00$). The *p*, or significance, value was below the alpha level of .05; therefore, the difference between the groups was statistically significant. (*Note:* The negative *t* value indicates whether the difference was positive or negative. The *t*-test value will be interpreted the same way. If you enter the posttest score variable first, the *t*-test value will be positive.)

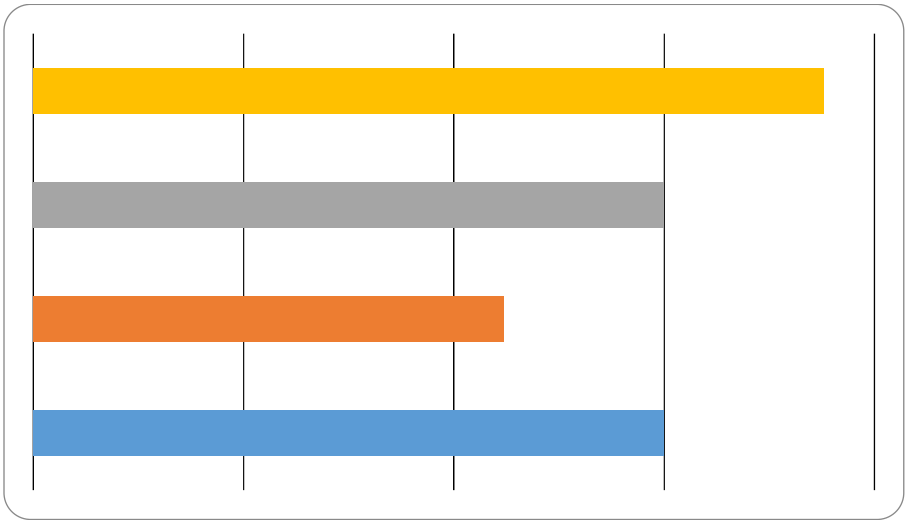
Chi Square

If you have a categorical independent variable and a categorical dependent variable, then you will need to conduct a **non-parametric statistic**. (See the assumptions of a *t*-test and ANOVA.) One non-parametric statistic is the chi-square. A **chi square** tells you if there is a difference from what is observed in the data and what is expected in the data. For example, see the charts below. Our sample included 112 males and 188; however, based on population, you should see 150 males and 150 females.

We want to determine if there is a statistically significant difference from what we observed and what we expected.

Respondent's Sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	112	37.3	37.3	37.3
	Female	188	62.7	62.7	100.0
	Total	300	100.0	100.0	



Use the following formula to determine to the observed frequency count differs from the expected frequency count. E represents expected, and O represents observed. Chi square is symbolized with the small Greek letter chi (χ^2).

$$\chi^2 = \Sigma (E - O)^2/E$$

$$\chi^2 = [(\square - \square)^2/\square] + [(\square - \square)^2/\square]$$

$$\chi^2 = [(150 - 112)^2/150] + [(150 - 188)^2/150]$$

$$\chi^2 = [(38)^2/150] + [(-38)^2/150]$$

$$\chi^2 = [1444/150] + [1444/150]$$

$$\chi^2 = [9.63] + [9.63]$$

$$\chi^2 \approx 19.25$$

$\chi^2_{CV} = 3.84$ ($df = 1$) (See the chi-square distribution chart on the next page.)

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

There was a statistically significant difference between the observed and expected data for gender, $\chi^2 = 19.25$. (*Note:* The chi square value of 19.25 exceeds the critical value of 3.84; therefore, there was a statistically significant difference.)

How to Conduct a Chi-Square in Excel*

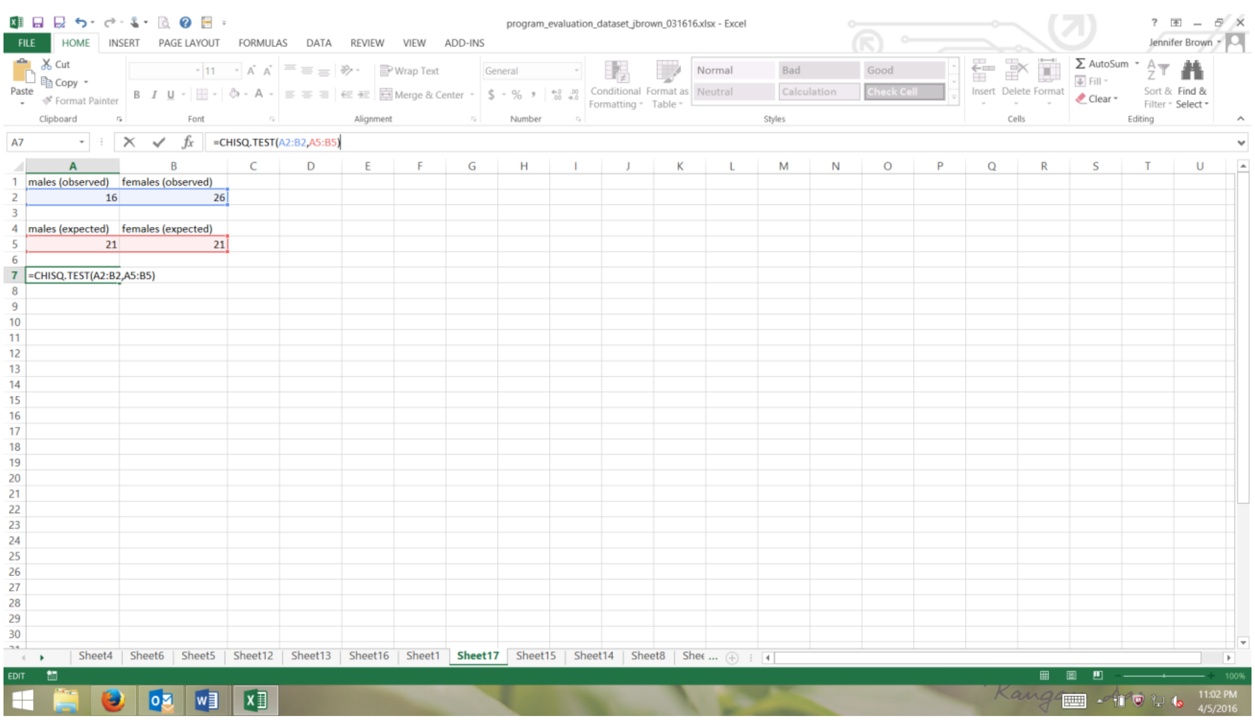
*You cannot conduct a chi-square using the Analysis ToolPak. You will need to enter the formula.

1. Conduct a frequency analysis to obtain counts for *gender*.
2. Open a new worksheet by selecting the **+** in the lower left corner.
3. Type the label for the first observed group in cell A1 of the new worksheet. (For this example, type “males (observed)”.)

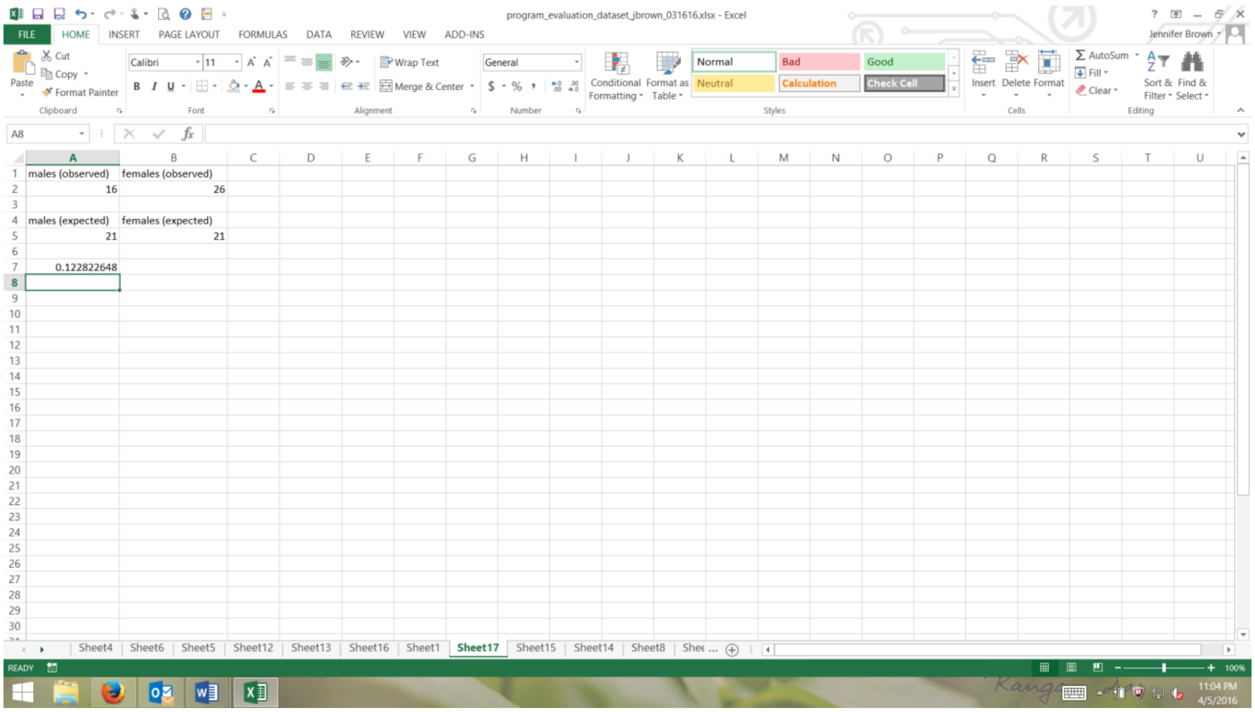
4. Type the label for the second observed group in cell B1 of the new worksheet. (For this example, type "females (observed)".)
5. Type the label for the first observed group in cell A4 of the new worksheet. (For this example, type "males (expected)".)
6. Type the label for the second observed group in cell B4 of the new worksheet. (For this example, type "females (expected)".)
7. Enter the data from the frequency analysis for each group in row 2. (For this example, type "16" in cell A2 and "26" in cell B2.)
8. Enter the expected values for each group in row 5. (For this example, type "21" in cell A5 and "21" in cell B5 because you would expected equal numbers in each group, 42/2 = 21.)
9. Select the cell A7.
10. To conduct the chi square, you will use the **CHISQ.TEST** function. In A7, type the following:

    =**CHISQ.TEST**(A2:B2,A5:B5)

    A2:B2 is the observed data, and A5:B5 is the expected data.



11. Select **Enter**. (*Note:* It is important that you select the **Enter** key after entering each formula.)



There was not a statistically significant difference between the expected 50/50 and the observed gender data, $p = .12$.

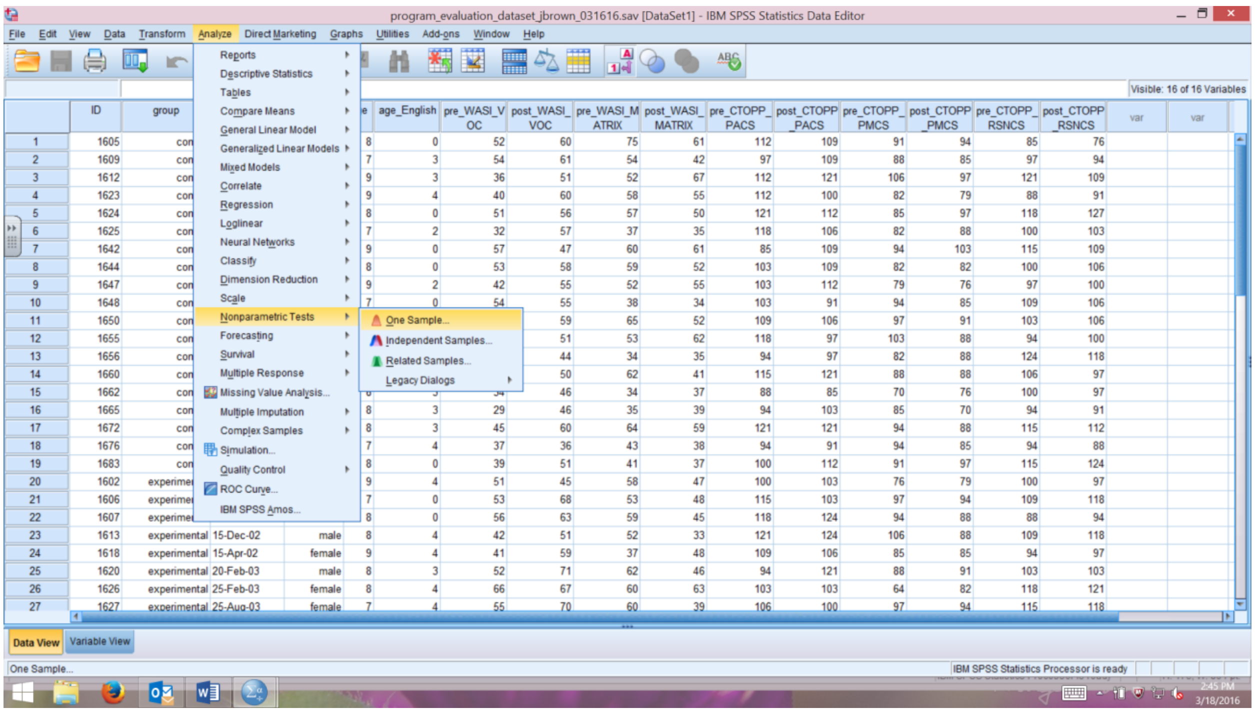
Why?

Why did we select a chi square to analyze these data?

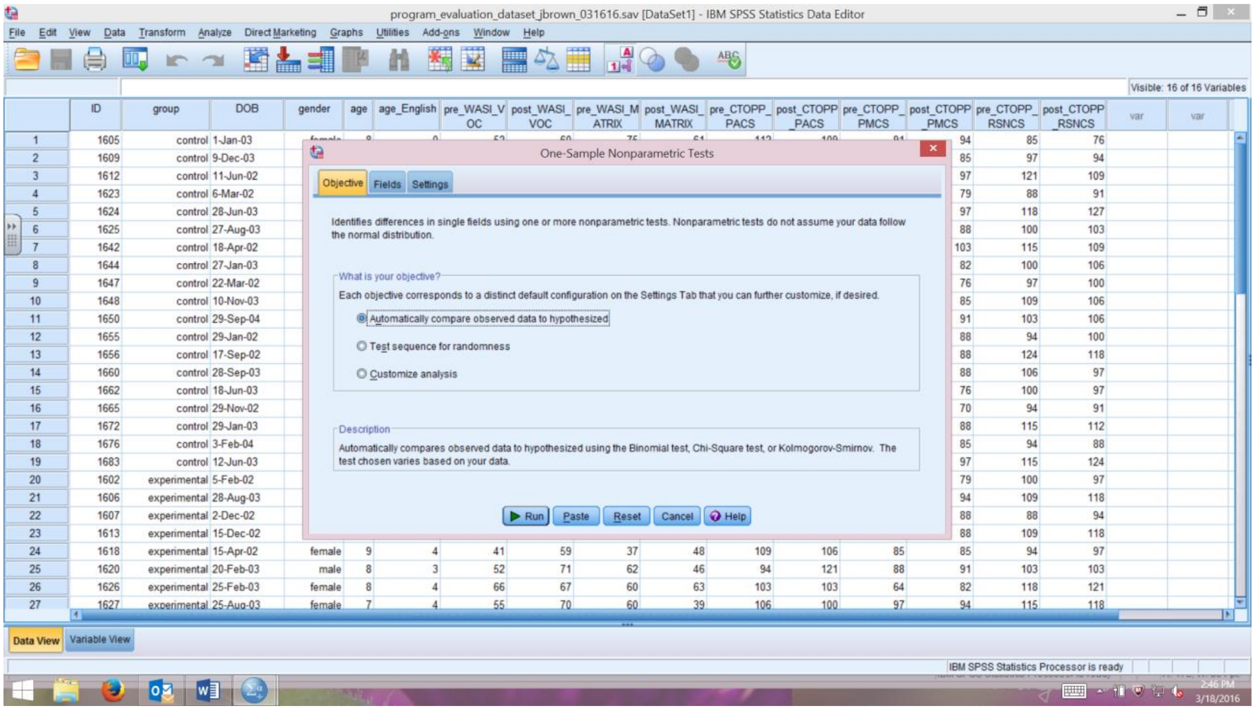
With this previous example, we wanted to determine if there were a difference between the observed, or collected, data and the expected data for the gender variable. If a student was to walk into your classroom, there would be a 50/50 chance the student's gender would be male. Sometimes, there are big differences between the observed and expected data. After we determine if a statistically significant difference exists, then you can explain that difference. For example, in the Program Evaluation Report Example #4 in Appendix G, you can see that approximately 80% of the freshman cohort were female. An explanation for this difference could be education and nursing tend to be career fields for females.

How to Conduct a Chi-Square in SPSS

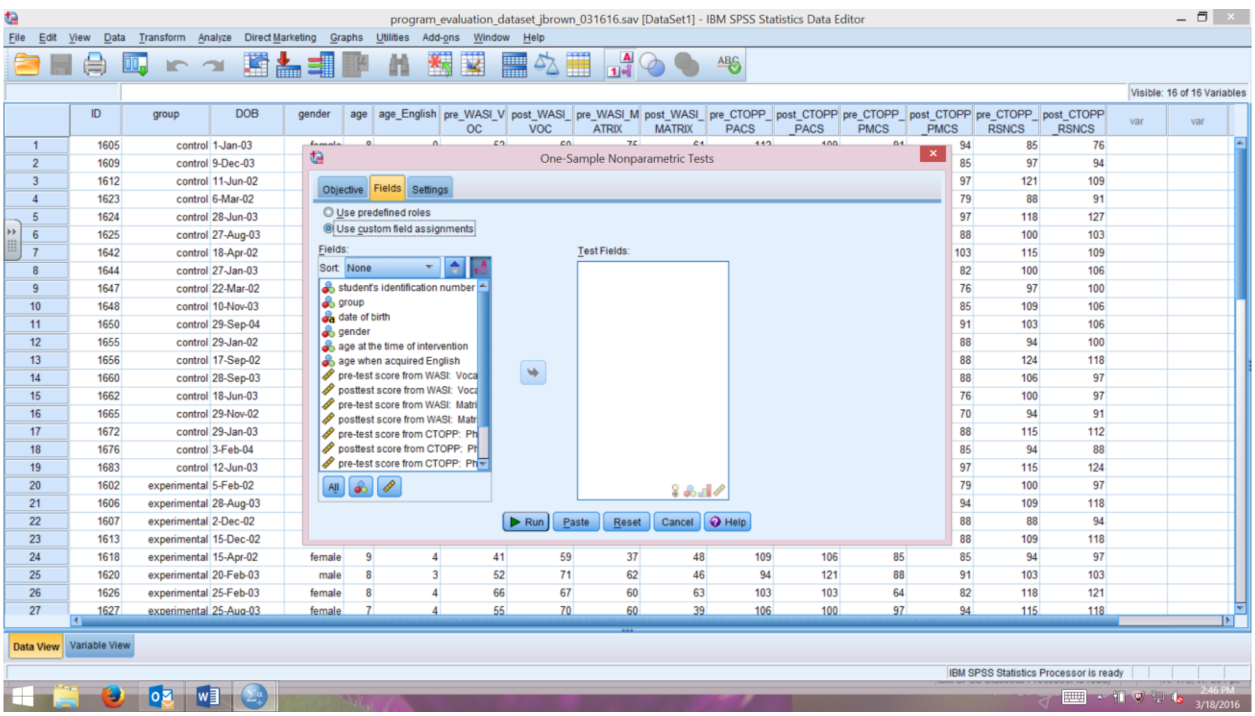
1. Analyze → Nonparametric Tests → One Sample



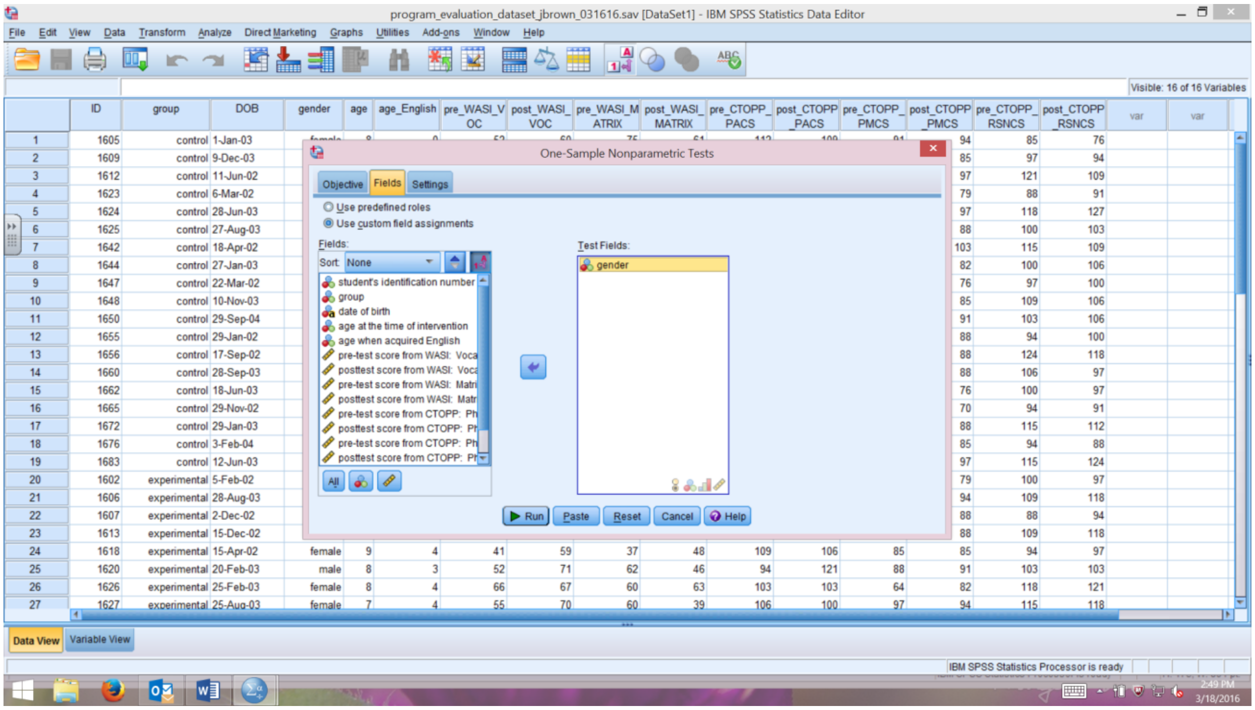
2. Under the "Objective" tab, make sure "Automatically compare observed data to hypothesized" is selected.



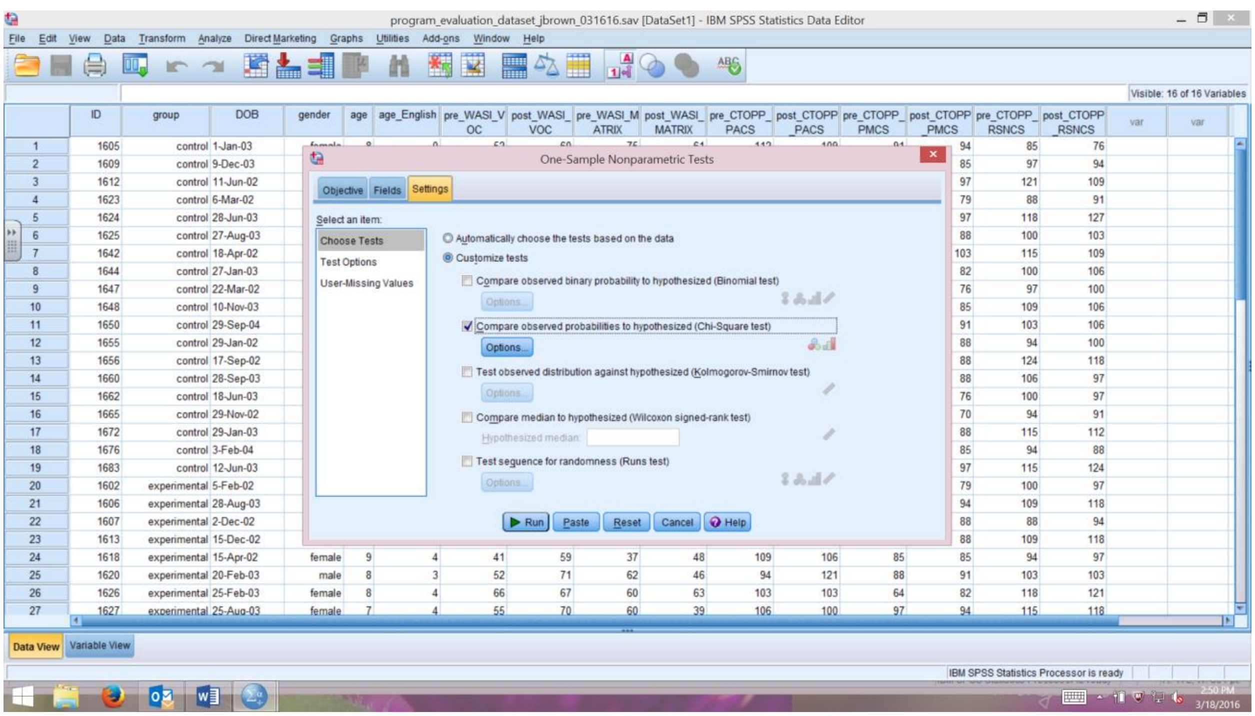
3. Under the “Fields” tab, select “Use custom field assignments”.



4. Select *gender* from the left box, and move it to the “Test Fields” box.



5. Under the “Settings” tab, select “Customize tests” and “Compare observed probabilities to hypothesized (Chi-Square test)”.



6. Select **Run**. (The analysis will appear on the output screen.)

SPSS Output

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories of gender occur with equal probabilities.	One-Sample Chi-Square Test	.123	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

There was not a statistically significant difference between the expected 50/50 and the observed gender data, p = .12.

If you conduct a frequency analysis, you can see that there were 38% males and 62% females. (*Note:* A chi square statistic is affected by sample size. See the formula. If you have a larger sample, then you are more likely to find statistical significance.)

gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	16	38.1	38.1	38.1
	female	26	61.9	61.9	100.0
	Total	42	100.0	100.0	

PRACTICE

You try with the *group* variable. Compare your output the following output.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The categories of group occur with equal probabilities.	One-Sample Chi-Square Test	.537	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

There was not a statistically significant difference between the observed and expected data for group, *p* = .54.

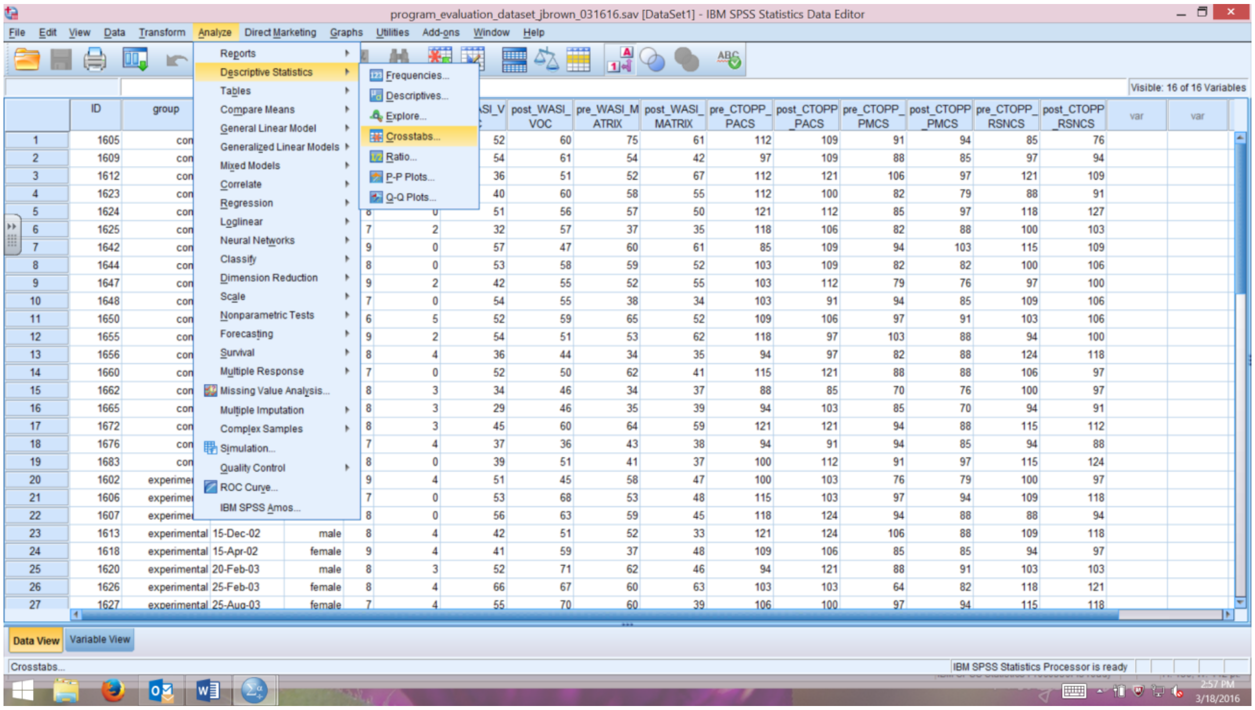
group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	control	19	45.2	45.2	45.2
	experimental	23	54.8	54.8	100.0
	Total	42	100.0	100.0	

For the frequency output above, you can see the percentage of participants in the control group was 45% and the percentage of participants in the experimental group was 55%. You would expect group membership to be 50/50.

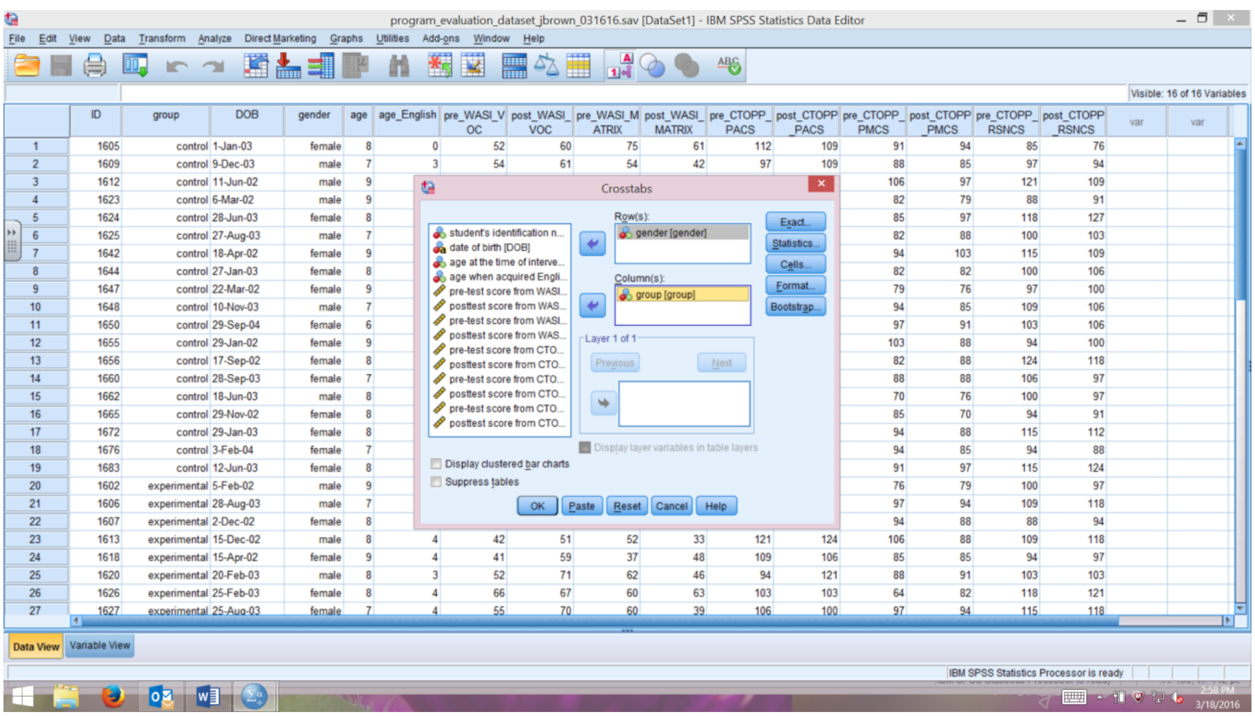
Using our dataset, we would like to know if the observed data differ from the expected within the groups, which involves the analysis of two variables. To analyze this purpose, you will need to conduct a chi square using the Crosstabs procedure in SPSS.

1. Analyze → Descriptives → Crosstabs



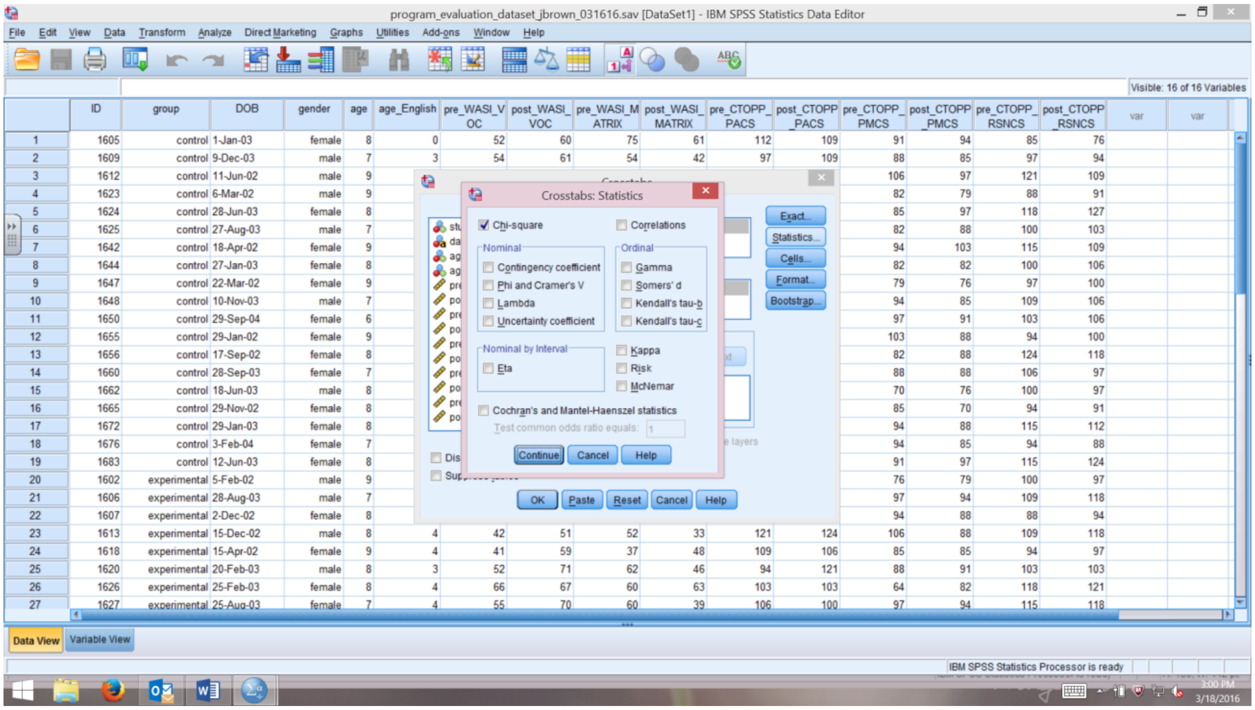
2. Select the *gender* variable, and move it to the “Row(s)” box.

3. Select the *group* variable, and move it to the “Row(s)” box.



4. Select **Statistics**.

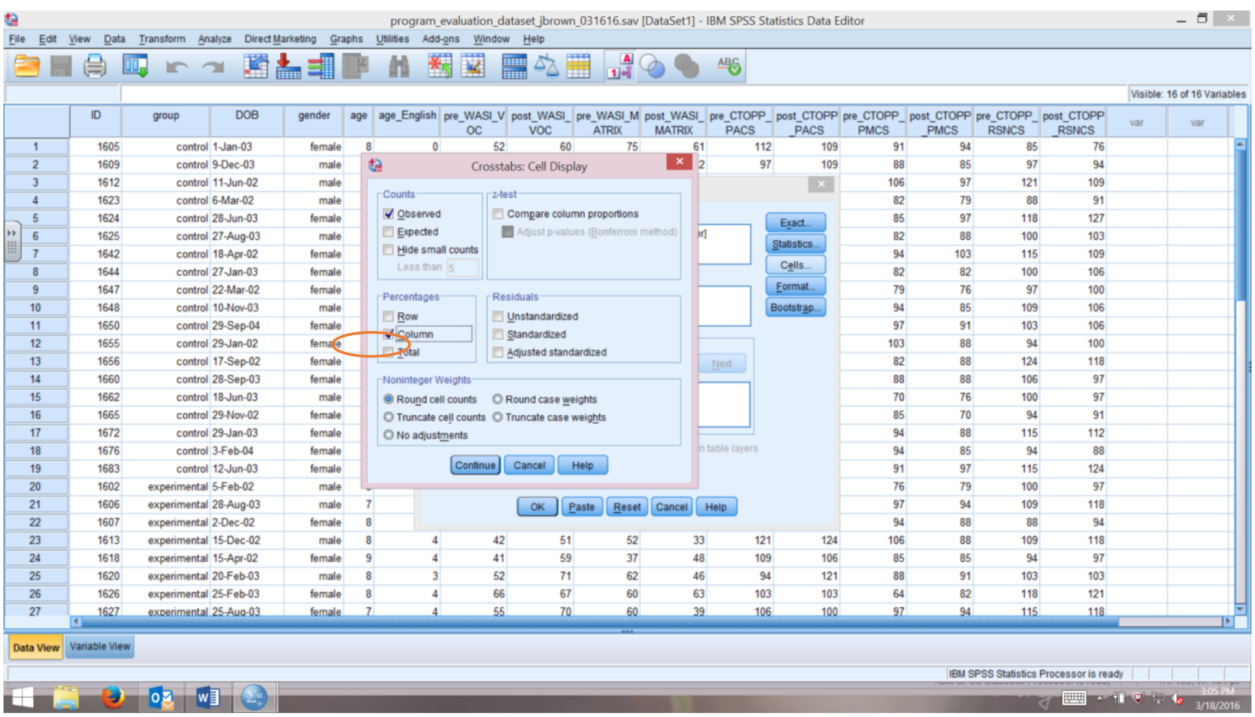
5. Select “Chi-square” in the pop-up window.



6. Select **Continue**.

7. Select **Cells**.

8. Select "Column" in the *Percentages* box.



9. Select **Continue**.

10. Select **OK**. (The analysis will appear on the output screen.)

SPSS Output

gender * group Crosstabulation

			group		Total
			control	experimental	
gender	male	Count	6	10	16
		% within group	31.6%	43.5%	38.1%
	female	Count	13	13	26
		% within group	68.4%	56.5%	61.9%
Total		Count	19	23	42
		% within group	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.625[a]	1	.429		
Continuity Correction[b]	.222	1	.638		
Likelihood Ratio	.629	1	.428		
Fisher's Exact Test				.530	.320
Linear-by-Linear Association	.610	1	.435		
N of Valid Cases	42				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.24.

b. Computed only for a 2x2 table

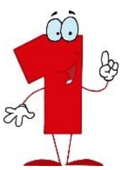
There was not a statistically significant difference between observed and expected data within each group, $\chi^2 = 0.63$; $p = .43$.

Why?

Why did we select a chi square to analyze these data?

In the previous example, both the IV and DV were categorical (i.e., gender and group). Thus, we cannot utilize the *t*-test or ANOVA statistics based on the assumptions. With the gender and group example, each variable contained two options so you would expected 25% of the sample to fall within cell (e.g., the number of males within the control group).

Additional Guided Practice



Here are two more quanitaitve datasets for you to practice the statistical analysis techniques. First, the FNO dataset is a portion of the original dataset from my FNO program evaluation (Brown, 2012b). The Excel and SPSS datasets can be downloaded from http://www.bugforteachers.com/prog_eval.html. See Appendix B for the background information, variable names, and labels. As you analyze the dataset, think about the rationale for selecting that statistic. I included the SPSS output so you can check your findings. Do not forget to examine the Levene's Test results if applicable. In addition, you should practice writing interpretations for the output. You can utilize my practice examples or excerpts from the program evaluation report examples in Appendices D through G. For independent practice, analyze the remaining variables within the dataset.

A. **Conduct a frequency count for number of students who attended review sessions (*Attended_Review_Session*).**

Did the student attend a review session for that assessment?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	212	69.1	69.1	69.1
	yes	95	30.9	30.9	100.0
	Total	307	100.0	100.0	

B. **Conduct a descriptive analysis for the difference between the students' retake test score and original test score (*Difference*).**

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Difference between retake test score and original test score	307	-59	58	3.77	18.801
Valid N (listwise)	307				

C. **Conduct a Pearson *r* to determine if a relationship exists between the students' original test score (*Original_Score*).and the retake test score (*Retake_Score*).**

Correlations

		Original test score before the retake	Test score for the retake assessment
Original test score before the retake	Pearson Correlation	1	.505**
	Sig. (2-tailed)		.000
	N	307	307
Test score for the retake assessment	Pearson Correlation	.505**	1
	Sig. (2-tailed)	.000	
	N	307	307

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

D. Conduct a paired-samples *t*-test to determine if there is a difference between the original test score (*Original_Score*) and the retake test score (*Retake_Score*).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Test score for the retake assessment	58.16	307	21.349	1.218
	Original test score before the retake	54.39	307	14.542	.830

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Test score for the retake assessment - Original test score before the retake	3.765	18.801	1.073	1.654	5.877	3.509	306	.001

E. Conduct a chi square to determine if the observed gender (*Gender*) data differs from the expected gender data within each class period (*Class_Period*).

class period * gender Crosstabulation

			gender		Total
			male	female	
class period	2	Count	32	24	56
		% within class period	57.1%	42.9%	100.0%
	3	Count	26	18	44
		% within class period	59.1%	40.9%	100.0%
	4	Count	40	45	85
		% within class period	47.1%	52.9%	100.0%
	5	Count	17	28	45
		% within class period	37.8%	62.2%	100.0%
	6	Count	48	29	77
		% within class period	62.3%	37.7%	100.0%
Total		Count	163	144	307
		% within class period	53.1%	46.9%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.128[a]	4	.058
Likelihood Ratio	9.180	4	.057
Linear-by-Linear Association	.000	1	.989
N of Valid Cases	307		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 20.64.

Second, the Math 2 dataset is a fictionous study created for educational purposes only. The purpose of the study was to determine the effects of rearranging the traditional order of unit instruction for a Math 2 course. The data sources were unit assessments. For this study, there were control and treatment groups. See Appendix C for the background information, variable names, and labels. Again, I included the SPSS output so you can check your findings. Again, I encourage you to examine the Levene's Test results and descriptives, if applicable, practice writing the interpretations, and analyze the remaining variables within the dataset for independent practice. As with any skill, increasing the amount of practice with analysis and intrepretation, the more comfortable and proficient you will become.

A. Conduct a frequency count for racial classification (*race*).

race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White	37	66.1	66.1	66.1
	Black	15	26.8	26.8	92.9
	Hispanic	4	7.1	7.1	100.0
	Total	56	100.0	100.0	

B. Conduct a descriptive analysis for average of student's unit tests (*average_test*).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
average of student's unit tests	56	65	91	81.85	5.838
Valid N (listwise)	56				

Statistics

average of student's unit tests

N	Valid	56
	Missing	0
Median		82.72

C. Conduct a Pearson *r* to determine if a relationship exists between the student grades on the Mid-Unit 1 Test (*Mid_Unit_1_Test*).and the Mid-Unit 5 Test (*Mid_Unit_5_Test*).

Correlations

		Grade from student's Mid-Unit 1 Test	Grade from student's Mid-Unit 4 Test
Grade from student's Mid-Unit 1 Test	Pearson Correlation	1	.328*
	Sig. (2-tailed)		.014
	N	56	56
Grade from student's Mid-Unit 4 Test	Pearson Correlation	.328*	1
	Sig. (2-tailed)	.014	
	N	56	56

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

D. Conduct an independent *t*-test to determine if there is a difference between the groups on the Unit 4 Test (*Unit_4_Test*).

Group Statistics

	group	N	Mean	Std. Deviation	Std. Error Mean
Grade from student's Unit 4 Test	control	28	86.07	5.270	.996
	treatment	28	91.18	9.718	1.837

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Grade from student's Unit 4 Test	Equal variances assumed	39.525	.000	-2.444	54	.018	-5.107	2.089	-9.296	-.918
	Equal variances not assumed			-2.444	41.615	.019	-5.107	2.089	-9.325	-.890

E. Conduct a One-Way ANOVA to determine if there was a difference between the groups on the Unit 5 Test (*Unit_5_Test*).

ANOVA

Grade from student's Unit 5 Test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5520.286	1	5520.286	33.815	.000
Within Groups	8815.429	54	163.249		
Total	14335.714	55			

F. Conduct a chi square to determine if the observed gender (*gender*) data differs from the expected gender data within each group (*group*).

gender * group Crosstabulation

			group		Total
			control	treatment	
gender	male	Count	7	8	15
		% within gender	46.7%	53.3%	100.0%
	female	Count	21	20	41
		% within gender	51.2%	48.8%	100.0%
Total		Count	28	28	56
		% within gender	50.0%	50.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.091[a]	1	.763		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.091	1	.763		
Fisher's Exact Test				1.000	.500
Linear-by-Linear Association	.089	1	.765		
N of Valid Cases	56				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.50.

b. Computed only for a 2x2 table

Summarizing the Quantiative Data Analysis

The following flowchart (see Figure 8) summarizes the application of each statistic discussed based on whether the variable is categorical or continuous and the purpose of the statistic.

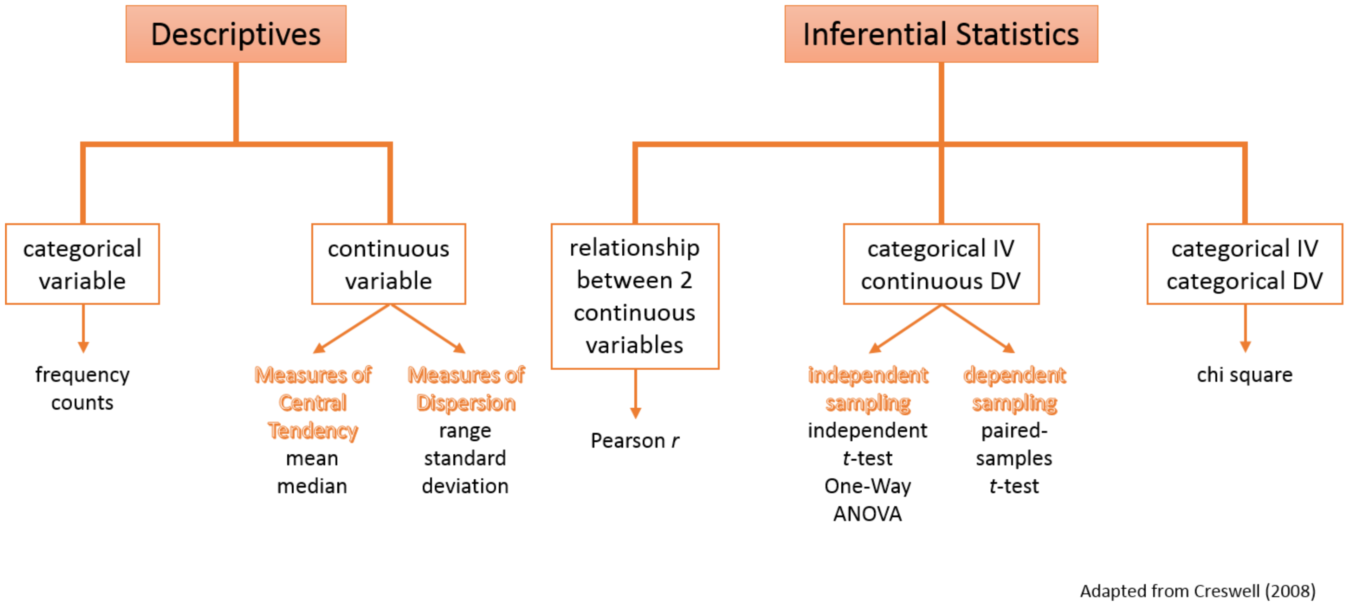


Figure 8. Summarizing quantiative data analysis techiques.

Analyzing and Interpreting Qualitative Data

According to Miles and Huberman (1994), qualitative data analysis occurs in three phases using an induction approach: Data reduction, data display, and conclusion drawing. These phases do not have to be sequential, and the process is interactive and cyclical, which begins during data collection. During the **data reduction** phase, the analyzer will reduce the qualitative data, which can include transcripts and field notes. This reduction process will result in writing summaries of the data and coding. **Codes** are defined as labels for assigning units of meaning to data. For the next phase, **data display** is the process of organizing the reduced data. These displays can be a Word document, Post-It notes on the wall, or highlighted transcripts. In this age of technology, I prefer highlighting the text within a Word document then using cut/paste to rearrange it. From your display, you should see themes, or chunks, begin to appear. These themes may be general words or phrases, such as "math class", or more specific, such as "Miss Smith's 10th grade math class. (*Note*: **Emerging themes** refers the themes, or topics, that emerge from the data during the preliminary stages of qualitative data analysis. Typically, qualitative data analysis takes about three times longer than quantiative data analysis.) How do these themes emerge? While there are many ways to examine the data for emerging themes, I prefer the following four methods.

1. repetition in the data, or coding that reoccurs
2. transitions, such as pauses, changes in voice tone, and transitional phrasing
3. similarities and differences within the coding or data
4. linguistic connectors, such as "if…then", "because", and "since", which imply causal relationships (Ryan & Bernard, 2003).

Lastly, **conclusion drawing** is where the analyzer determines the patterns and/or explanations based on the data reduction and data display. An example of a pattern would be "Most of the 9th grade teachers felt the newly implemented secondary mathematics curriculum had more real-world application for the students."

See the Program Evaluation Report Example #5 in Appendix H for an example of qualitative data analysis. Within this example, under "Student Perceptions of Academic Programs", the paragraph lists four phrases that were coded from the focus groups' responses (i.e., class size, campus resources as support, academic factors, and satisfaction). These four coded items were grouped into the "Student Perceptions of Academic Programs" theme, which was referred to as an emerging theme.

Using the secondary mathematics curriculum example, here is my data analysis plan. After the initial descriptives are conducted with the scores from the graduation examination, benchmark examinations, and the Advanced Placement Calculus Examination, a series of paired-samples *t*-test will be conducted to determine if level of mathematical proficiency has changed across implementation years, across grade level, and between groups. One of the components for evaluating the implementation activities will be teacher interviews that are conducted before, during, and after the implementation year. These data will be analyzed for emerging themes.

"Data don't make any sense,
we will have to resort to statistics."

Chapter 6

Writing the Evaluation Report

Step 6: Writing the Evaluation Report

The purpose of the **evaluation report** is to disseminate the findings of your program evaluation. This dissemination can occur in several formats (e.g., formal presentation using PowerPoint, small group meeting, or fact sheet). Whichever format is selected, typically, a written report will be included with all of those dissemination formats. This written report is often referred to as a **deliverable**. The written report includes the following information (Frechtling, 2002): Again, remember to avoid the use of statistical jargon. You want a user-friendly deliverable for your given audience.

- Background (typically presented in an executive summary or introduction and includes the purpose of the evaluation)
- Research questions
- Methods (includes all components of data collection: sample/participants, measures, interventions, and procedures)
- Data analysis and Results
- Conclusions and Recommendations

The results of the evaluation plan for the secondary mathematics curriculum example will be reported to the school faculty each semester as a formative report and during the pre-service faculty meeting as a summative report. Once a semester, the evaluation team will meet with the Superintendent individually and with the local school board during a caucus meeting. Afterwards, an annual summative report will be presented at a public school board meeting. The expected findings include increased mathematical proficiency as measured by graduation exit examination scores in mathematics and Advanced Placement Calculus Examination scores. In addition, the team would expect successful curriculum implementation from the staff members' point of view.

In the appendices D through H, there are five sample program evaluations that I have conducted during recent years. The original documents have been edited for educational purposes. (*Note*: Considering the length of Program Evaluation Example Report #4, I only provided small excerpts.)

GLOSSARY

Activities – Part of the logic model that outlines any services, materials, and/or events associated with the program's implementation.

ANOVA – *see Analysis of Variance*

Analysis of Variance – Parametric statistics that compare means to determine if there is a difference between two or more groups (e.g., One-Way, Repeated Measures, and Factorial).

Bivariate – Two variables.

Case Study – A research design occurs when the program evaluator wants an extensive study of a group of individuals.

Categorical Data – Data that can be counted (e.g., gender).

Causal Comparative Research – A research design where pre-existing groups will be compared.

Chi Square – A non-parametric statistic for determining if there is a difference between the observed data and expected data.

Codes - Labels for assigning units of meaning to data.

Conclusion Drawing – The process where the analyzer determines the patterns and/or explanations based on the data reduction and data display.

Continuous Data – A range of numbers on a continuum (e.g., test scores).

Control Group – The group that did not receive the intervention.

Correlation – The relationship between two variables.

Correlational Research – A research design will answer the question, "What is the relationship between two or more variables?"

Critical Value – The value used to determine statistically significance based on the predetermined alpha level.

Data Display – A process of organizing the reduced data.

Data Reduction Phase – A process where the analyzer will reduce the qualitative data, which can include transcripts and field notes.

Deliverable – A written report that contains the findings of a program evaluation.

Dependent Variable – A variable that is dependent upon another observation.

Descriptive Research – A research design will answer the question, "How much exists?"

Descriptives - The numerical summary of a given dataset.

Dosage – Amount of program activities received.

Emerging Themes – Themes, or topics, that emerge, or appear, from the data during the preliminary stages of qualitative data analysis.

Evaluation Plan – The systematic plan that will be used to answer your research questions.

Evaluation Purpose – The reason for conducting a program evaluation.

Evaluation Report – *See deliverable*.

Experimental Group - The group that did receive the intervention.

Experimental Research – A research design where a stakeholder manipulates the conditions and randomly assigns students to the groups.

Fidelity – Extent to which program activities were implemented based on standardized procedures.

Formative Evaluation – An evaluation used to determine the quality or effectiveness of a program and to indicate strengths or weaknesses, which provides the program staff with feedback.

Frequencies – Counting values or labels within a variable.

Homogeneity of Variance – Equal variance among groups.

Independent *t*-test – A parametric statistic that compares means to determine if there is a difference between two independent groups.

Independent Variable – A variable that is not dependent on other observations.

Inferential Statistics – These statistics are used to analyze sample data, then the findings are generalized back the targeted population.

Inputs - Any funding sources and/or resources provided to support the program.

Interval Data – Level of measurement where there is equal and meaningful distance between the scores (e.g., test scores).

Intervention – The program activities that were implemented.

Levene's Test for Equality of Variance – A statistical test to determine if the groups have equal variance among the scores.

Logic Model – A flowchart that serves as a blueprint for the program, including the inputs, activities, short-term objectives, and long-term objectives.

Long-term Objectives – Part of the logic model that outlines the enduring impacts of the program.

Mean – The average of a given dataset.

Measures of Central Tendency – Measures that describe the center or middle of a given dataset (e.g., mean and median).

Measures of Dispersion – Measures that describe the spread or variability of a given dataset (e.g., range and standard deviation).

Median – The middle value of a sequentially ordered dataset.

Negative Correlation – A bivariate relationship where one value decreases and the other value increases.

Nominal Data – A level of measurement where the values are predetermined labels or names (e.g., gender and racial classification).

Non-Parametric Statistics – The analysis techniques utilized for categorical data.

One-Way ANOVA – *See Analysis of Variance*

Ordinal Data – Level of measurement where the scores are ranked (e.g., 5-point rating scale).

Outlier – The value that is significantly outside the range of the other values in the dataset.

Paired-samples *t*-test - A parametric statistic that compares means to determine if there was a change from one data point to another using the same participants.

Parametric Statistics - The analysis techniques utilized for continuous data.

Pearson Product Moment Correlation – A parametric statistics used to determine if a relationship exists between two variables.

Pearson *r* – *see Pearson Product Moment Correlation*

Positive Correlation - A bivariate relationship where one value increases and the other value increases.

Program Evaluation - Systematic collection of data about the activities and outcomes of a program.

Purposeful Sampling – A sampling technique where persons will be selected based on the context of a qualitative evaluation.

Qualitative Data – Data that describes a characteristic or observation.

Quantitative Data – Data that measures a characteristic or observation.

Quasi-experimental Research – A research design where the conditions of the target sample are manipulated.

Random Sampling – A sampling technique where each person has an equal chance of being selected.

Range – The difference between the minimum value and the maximum value.

Ratio Data – Level of measurement where there is an absolute zero (e.g., temperature).

Reach - Extent to which the targeted population received the scheduled intervention dosages.

Research Design – A strategy for conducting the program evaluation.

Sample – A representative subset of a targeted population.

Short-term Objectives – Part of the logic model that outlines the immediate impact of the implementation activities.

Simple Random Sampling – A sampling technique where every person is thrown into the pot then will be selected.

Standard Deviation – The typical difference between the value and mean.

Stakeholder - Any individual or group that has a "stake" or interest in the outcome of the program evaluation.

Stratified Random Sampling – A sampling technique where the persons will be selected based on a given characteristics (e.g., gender or racial classification).

Summative Evaluation – An evaluation used to determine program quality based on outcomes after the program has ended.

Systematic Random Sampling – A sampling technique where every nth person will be selected from a list (e.g., alphabetize list of 10th-grade students with a high school).

Targeted Population – The entire group of observations from which a sample can be drawn.

Treatment Group – *see Experimental Group*

Variable – A characteristic or observation where values are given.

Variance – A value given to indicate the spread of individual data.

Volunteer Sampling – A sampling technique where each person will be selected by convenience and self-selected.

References

Ary, D., Jacobs, L. C., Razavieh, A., & Sorensen, C. (2006). *Introduction to research in education* (7th ed.). Belmont, CA: Thomson Wadsworth.

Bell, J. L. (2008). *An examination of cognitive and non-cognitive factors and academic success in the pre-engineering curriculum at a four-year southeastern university* (Order No. 3333110). Available from ProQuest Dissertations & Theses Full Text. (304688037).

Bell, J. L., Halpin, G., & Halpin, G. (2007). *An evaluation training*. Roundtable Discussion presented at the Annual Meeting of the National Staff Development Council, Dallas, Texas.

Brown, J. L. (2012a). *An evaluation of the fall 2012 semester freshman learning communities for EDUC 2120 and EDUC 2130.* Unpublished manuscript, Department of Teacher Education, Columbus State University, Columbus, Georgia.

Brown, J. L. (2012b). The impact of the failure is not an option policy on student grades. *Perspectives in Learning, 13*(1), 22-28.

Brown, J. L., & Andrews, A. (2015). *College of education and health professions longitudinal retention study of freshman cohorts entering 1999 through 2014.* Unpublished Manuscript, Department of Teacher Education, Columbus State University, Columbus, Georgia.

Brown, J. L., & Bentley, E. (2013). Do other people "gape" at your writing? *National Teacher Education Journal, 6*(3), 33–36.

Brown, J. L., & Robinson-McDonald, D. (2014). An exploratory study of instructional strategies, academic integration, and subsequent institutional commitment. *Journal of Research in Education, 24*(2), 160-172.

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

Creswell, J. W. (2008). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (3rd ed.). Boston, MA: Pearson.

Frechtling, J. (Ed.) (2002). *The 2002 user friendly handbook for program evaluation* (NSF 02-057). Washington, DC: The National Science Foundation, Directorate for Education & Human Resources, Division of Research, Evaluation, and Communication.

Fitzpatrick, J. L., Sanders, J., R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). Boston, MA: Pearson.

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

Killion, J. (2002). *Assessing Impact: Evaluating staff development.* Oxford, OH: National Staff Development Council.

Klingner, J. K., Ahwee, S., Pilonieta, P., & Menendez, R. (2003). Barriers and facilitators in scaling up research-based practices. *Exceptional Children, 69,* 411-429.

Mathison, S. (2008). What is the difference between evaluation and research – and why do we care? In N. L. Smith, & P. R. Brandon (Eds.), *Fundamental Issues in Education* (pp. 183 – 196). New York, NY: The Guilford Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage Publications.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 et seq. (2001).

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Ryan, G. W., & Bernard, H. R. (2003). Techniques to identify themes. *Field Methods, 15*(1), 85-109.

Slater, J., Strait, D. L., Show, E., O'Connell, S., Thompson, E., & Kraus, N. (2014). Longitudinal effects of group music instruction on literacy skills in low-income children. *PLOS ONE, 9*(11), 1-9.

Suvedi, M., & Morford, S. (2003). *Conducting program and project evaluation: A primer for natural resource program managers in British Columba*. Kamloops, BC: Forest Research Extension Partnerships.

Wall, J. E. (n.d.). *Program evaluation model: 9-step process*. Retrieved from http://region11s4.lacoe.edu/attachments/article/34/%287%29%209%20Step%20Evaluation%20Model%20Paper.pdf.

(*Note*: Graphics were retrieved from bing.com image search and google.com images.)

A. Background for Musical Training Quantitative Dataset

Participants: The 42 bilingual (Spanish/English) students were pseudo-randomly assigned to groups because of the need to keep the groups equal in terms of sex, age, dominant hand, IQ, age of exposure to the English language, English reading ability, and maternal education level.

Setting: The research project was implemented in schools where there are at least 90% of students on free or reduced lunch from Los Angeles, California.

Intervention: The participants in the treatment group underwent musical training for 1 hour, two times a week, for 3 to 10 months (*M* = 5 months). The musical training included rhythm, pitch, performance, improvisation, composition, musical vocabulary, and orchestral instrumentation. None of the participants had previous musical training.

Original dataset retrieved from:

Slater, J., Strait, D. L., Show, E., O'Connell, S., Thompson, E., & Kraus, N. (2014). Longitudinal effects of group music instruction on literacy skills in low-income children. *PLOS ONE, 9*(11), 1-9.

Measures

Wechsler Abbreviated Scale of Intelligence

Ages: 6-0 through 89-11
Testing Time: Two-subtest form, 15 minutes
Administration: Individual
Publisher: Pearson

Uses

Psychologists, clinicians, and researchers can get a fast and reliable measure of intelligence when screening for mental retardation, giftedness, or for other purposes. In addition, the WASI is useful for reassessing individuals who have had a comprehensive evaluation and need reevaluation. Other applications include:

- Estimating IQ scores for large samples when administration of a full battery is not feasible or necessary
- Screening to determine need for an in-depth evaluation
- Obtaining estimates of current cognitive functioning for individuals referred for psychiatric evaluations
- Estimating IQ scores for vocational or rehabilitation purposes
- Estimating IQ scores for research purposes

Subtests

- Vocabulary subtest for measuring word knowledge, verbal concept formation, and fund of knowledge
- Matrix Reasoning for measuring visual information processing and abstract reasoning skills
- An estimate of general intellectual ability can be obtained from the two subtests, which can be given in about 15 minutes.

Information retrieved directly from Pearson (http://www.pearsonclinical.com/education/products/100000593/wechsler-abbreviated-scale-of-intelligence-wasi.html)

CTOPP: Comprehensive Test of Phonological Processing

Ages: 4-0 through 24-11
Testing Time: 40 minutes
Administration: Individual
Publisher: pro-ed

Uses

The CTOPP has four principal uses: (1) to identify individuals who are significantly below their peers in important phonological abilities, (2) to determine strengths and weaknesses among developed phonological processes, (3) to document individuals' progress in phonological processing as a consequence of special intervention programs, and (4) to serve as a measurement device in research studies investigating phonological processing.

Composite Scores and Subtests

Phonological Awareness Composite Score (PACS) comprises the standard scores of three subtests- Elision, Blending Words, and Phoneme Isolation for 7 through 24 year olds. The PACS represents the examinee’s awareness of and access to the phonological structure of oral language.

- Elision measures the ability to remove phonological segments from spoken words to form other words.
- Blending Words measures the ability to synthesize sounds to form words.
- Phoneme Isolation measures the ability to isolate individual sounds within words.

Phonological Memory Composite Score (PMCS) comprises the standard scores of two subtests - Memory for Digits and Nonword Repetition-for all individuals. The PMCS represents the examinee's ability to code information phonologically for temporary storage in working or short-term memory.

- Memory for Digits measures the ability to repeat numbers accurately.
- Nonword Repetition measures the ability to repeat nonwords accurately.

The Rapid Symbolic Naming Composite Score (RSNCS) comprises the standard scores of two subtests-Rapid Digit Naming and Rapid Letter Naming –for all individuals. The RSNCS measures the examinee’s ability to include efficient retrieval of phonological information from long-term or permanent memory and execute a sequence of operations quickly and repeatedly.

- Rapid Digit Naming measures the ability to rapidly name numbers.
- Rapid Letter Naming measures the ability to rapidly name letters.

Information retrieved directly from pro-ed (http://www.proedinc.com/customer/productView.aspx?ID=5187)

Variable Name	Label	Coding
ID	student's identification number	
group	Group	0 = control group 1 = experimental group
DOB	date of birth	
gender	Gender	0 = male 1 = female
age	age at the time of intervention	
age_English	age when acquired English	
pre_WASI_VOC	pre-test score from WASI: Vocabulary Subtest	
post_WASI_VOC	posttest score from WASI: Vocabulary Subtest	
pre_WASI_MATRIX	pre-test score from WASI: Matrix Reasoning Subtest	
post_WASI_MATRIX	posttest score from WASI: Matrix Reasoning Subtest	
pre_CTOPP_PACS	pre-test score from CTOPP: Phonological Awareness Composite Score (Elision, Blending Words, and Phoneme Isolation)	
post_CTOPP_PACS	posttest score from CTOPP: Phonological Awareness Composite Score (Elision, Blending Words, and Phoneme Isolation)	
pre_CTOPP_PMCS	pre-test score from CTOPP: Phonological Memory Composite Score (Memory for Digits and Nonword Repetition)	
post_CTOPP_PMCS	posttest score from CTOPP: Phonological Memory Composite Score (Memory for Digits and Nonword Repetition)	
pre_CTOPP_RSNCS	pre-test score from CTOPP: Rapid Symbolic Naming Composite Score (Rapid Digit Naming and Rapid Letter Naming)	
post_CTOPP_RSNCS	posttest score from CTOPP: Rapid Symbolic Naming Composite Score (Rapid Digit Naming and Rapid Letter Naming)	

B. Background for FNO Quantitative Dataset

Participants: All ninth-grade students at Brownville High School who took Math I during the 2008-2009 school year.

Intervention: The high school implemented the FNO Policy at the beginning of the 2008-2009 school year. The policy stated that any student who scored less than 70% on a major assessment was required to retake the assessment at least once. The only exception to this policy was the assessments administered in Advanced Placement courses. Within each department, a retake administrator coordinated the retake sessions unless directed by the teacher of record. Retake sessions were scheduled for Tuesdays and Thursdays afterschool. Students were encouraged, but not required, to participate in review sessions prior to retaking the assessments. Mondays and Wednesdays were designated as review session days. On these days, the student could work with his or her teacher or with a teacher who supervised the tutoring sessions within each department.

Measures

- Unit assessments developed by the school's math department

Variable Name	Label	Coding
Student ID	student identification number	
Gender	gender	1 = male 2 = female
Race	racial classification	1 = white 2 = black 3 = Hispanic
Special Needs	Does the student receive services for special needs (e.g., SST, 504, special education, or ESOL)?	0 = no 1 = yes
Class Period	class period	
Original Score	Original test score before the retake	
Attended Review Session	Did the student attend a review session for that assessment?	0 = no 1 = yes
Retake Score	Test score for the retake assessment	
Difference	Difference between retake test score and original test score	

C. Background for Math 2 Quantitative Dataset

Participants: The control group included 28 Math II students from the first block, and the treatment group included 28 Math II students from the second block. All participants were 10th-grade students at Brownville High School during the 2009-2010 school year.

Intervention: The control group's unit instruction followed the traditional order: Unit 1, Unit 2, Unit 3, Unit 4, Unit 5, and Unit 6. The treatment group's unit instruction followed an alternative order, which begin with less difficult content: Unit 4, Unit 6, Unit 1, Unit 5, Unit 2, and Unit 3. Both classes were taught by the same teacher using the same instructional materials and assessments. The concepts for each unit are presented below:

- Unit 1 – Quadratic Functions (Concepts include graphing and solving quadratic functions and inequalities and arithmetic series.)
- Unit 2 – Right Triangle Trigonometry (Concepts include similar and special right triangles, trigonometric ratios, and solving right triangles.)
- Unit 3 – Circles and Spheres (Concepts include properties of circles, arcs, chord, angle relationships, segment lengths, and spheres.)
- Unit 4 – Statistics: Data Analysis (Concepts include exploring and collecting data, mean and standard deviations, and comparing data sets.)
- Unit 5 – Piecewise, Exponential, and Inverses (Concepts include piecewise functions, exponential functions, geometric sequences, composition of functions, and inverse functions.)
- Unit 6 – Statistics: Finding the Best Model (Concepts include examining relationships, linear models, and quadratic models.)

Measures

- Unit assessments developed by the school's math department

Variable Name	Label	Coding
student ID	student identification number	
gender	gender	1 = male 2 = female
race	racial classification	1 = white 2 = black 3 = Hispanic
group	group	0 = control group 1 = treatment group
Mid-Unit 1 Test	Grade from student's Mid-Unit 1 Test	
Unit 1 Test	Grade from student's Unit 1 Test	
Unit 2 Test	Grade from student's Unit 2 Test	
Mid-Unit 3 Test	Grade from student's Mid-Unit 3 Test	
Unit 3 Test	Grade from student's Unit 3 Test	
Unit 4 Test	Grade from student's Unit 4 Test	
Mid-Unit 5 Test	Grade from student's Mid-Unit 5 Test	
Unit 5 Test	Grade from student's Unit 5 Test	
Unit 6 Test	Grade from student's Unit 6 Test	
average test	average of student's unit tests	

The Impact of the Failure is not an Option Policy on Student Grades

The purpose of this program evaluation was to determine the impact of the Failure is Not an Option (FNO) Policy at LaGrange High School on student test grades. To evaluate this policy, the following research questions were used: (1) Did the students who retook their assessments improve their assessment scores?; (2) Did the change in assessment scores differ by department?; and (3) Was there a difference with the change in assessment scores between the students who participated in a review session and the students who did not participate in a review session?

Methods

Participants

The selected participants were enrolled at LaGrange High School, which is part of a school district that contains 3 high schools, 3 middle schools, and 14 elementary schools. The high school, with Grades 9 through 12, has a total enrollment of 1,355. The gender classification is 48% male and 52% female. The racial makeup of the school is 53% White, 43% Black, and 4% who classify themselves as belonging to other racial groups. Six and a half percent of the students receive special education services. Forty-eight percent of the students are eligible for free or reduced meals. In 2008, the graduation rate was 70.6%, which exceeds the district graduation rate of 68.9% but falls below the state graduation rate of 75.4% (The Governor's Office of Student Achievement, 2008).

Intervention Activities

Description. The high school piloted the FNO Policy for the school system. The policy stated that any student who scored less than 70% on a major assessment was required to retake the assessment at least once. The only exception to this policy was the assessments administered in Advanced Placement courses. Within each department, a retake administrator coordinated the retake sessions unless directed by the teacher of record. Retake sessions were scheduled for Tuesdays and Thursdays afterschool. Students were encouraged, but not required, to participate in review sessions prior to retaking the assessments. Mondays and Wednesdays were designated as review session days. On these days, the student could work with his or her teacher or with a teacher who supervised the tutoring sessions within each department.

When the assessment was returned to the student, he or she completed a simple contract with the teacher of record and selected four possible retake dates. A copy of the contract was given to the student, teacher of record, and retake administrator. From the date that the assessment was returned to the student, the student had 2 weeks to retake the assessment. If the student did not retake the assessment within the allotted time, he or she was referred to the appropriate administrator, who assigned the student to an in-school suspension retake session. If a student scored 70% or greater on the original assessment, then he or she could opt to retake the assessment using the same procedures. In addition, if a student chose, he or she could continue to retake the assessment as many times as needed to improve his or her score to the desired level within the same semester.

Procedure: Data Collection. At the beginning of each semester, the principal sent a blank spreadsheet with column headings to each certified staff member via email. The column headings included student's name, teacher's name, class period, course title, assessment type, assessment title, original score, date of original assessment, retake score, date of retake assessment, exceptionality, and participation in a review session. At the end of each semester, the certified staff members were instructed to submit the spreadsheet that contained the itemized information for each retake to the main office via email. An administrative assistant for the school compiled the data into a master spreadsheet. The researcher requested and received the master spreadsheet for each semester via email from the principal.

Data Analysis and Results

Research Question 1

A series of frequency and descriptive analyses were conducted to determine the difference between original and retake scores. Across eight departments, a total of 2,163 retakes were administered during the first semester and 3,580 retakes during the second semester. Thus, the average student at the high school retook approximately two assessments during the first semester and approximately three assessments during the second semester. The school had an increase of 65.51% in the number of retakes from first to second semester. This difference could be contributed to more students participating in the program and/or consistency in record keeping procedures.

For first semester, mean difference for the school was 18.03 points. Using the school's grading policy, the average student could improve his or her final course grade as much as 7.35 points by retaking assessments in a given course. For second semester, the mean difference for the school was 16.82 points. The average student could improve his or her final grade by 6.73 points. The improvements in assessment scores were similar between the two semesters. Hence, a student could increase his or her final grade in a given course as much as one letter grade. Table 1 displays the original and retakes scores by semester and department.

Table 1

Original and Retake Scores by Semester and Department

	First Semester				Second Semester			
Department	*n*	Original	Retake	Difference	*n*	Original	Retake	Difference
English	311	54.50	73.19	18.69	483	49.16	72.40	23.24
Math	572	55.23	57.01	1.78	930	54.67	59.97	5.30
Science	765	52.34	60.68	8.34	1045	50.16	57.57	7.41
Social Studies	317	53.13	70.16	17.03	790	63.97	60.91	-3.06
CTAE	18	43.28	75.78	32.50	60	56.97	69.40	12.43
PE	56	54.36	75	20.64	8	50.38	58.14	7.76
Foreign Language	108	52.45	74.94	22.48	140	50.16	69.16	19.00
Fine Arts	16	47.63	70.40	22.77	124	0	62.45	62.45
Total	2163	51.26	69.63	18.03	3580	46.93	63.75	16.82

Research Question 2

A series of frequency and descriptive analyses were conducted to determine the differences among the eight departments (See Table 1). The number of retakes within a department ranged from 16 to 765 for first semester and from 8 to 1,045 for second semester. The greatest number of retakes was administered in the science department for the first and second semesters. The least number of retakes was administered in the fine arts department for the first semester and in the physical education department for the second semester. This variation in the number of retakes could be contributed to the content and assessment within each department.

The difference between the original and retake scores ranged from 1.78 to 32.50 for the first semester and from -3.06 to 62.45 for the second semester. The largest difference occurred in the CTAE department for the first semester and in the fine arts department for the second semester. The smallest difference for first semester occurred in the math department and for second semester in the social studies department. These differences could be related to the course content and/or the consistency of the record keeping procedures within each department. The FNO policy had a great impact on the students' final course grades in at least 60% of the departments.

Research Question 3

A chi-square analysis was conducted to determine the statistical difference between participation in a review session and change in assessment scores after retaking the assessment for each semester. As a follow-up, individual chi-square analyses were conducted with the frequencies of increased, decreased, and unchanged scores and with participation in a review session. A criterion of .05 for the *p*-value was established as statistically significant. A criterion of .10 for phi coefficient (φ) was established as meaningful.

For the first semester, with 2,057 cases, there was a statistically significant and meaningful difference between participation in a review session and change in the assessment scores ($\chi^2 = 34.01$; $\varphi = .13$; $p < .001$). There was a statistically significant difference between participation in a review session and the number of unchanged assessment scores ($\chi^2 = 7.84$; $p = .01$). There was a statistically significant difference for the number of increased assessment scores ($\chi^2 = 26.80$; $p < .001$) and for the number of decreased assessment scores ($\chi^2 = 96.63$; $p < .001$).

Second semester analyses, with 3,081 cases, yielded similar results ($\chi^2 = 119.21$; $\varphi = .20$; $p < .001$). There was a statistically significant difference between participation in a review session and the number of unchanged assessment scores ($\chi^2 = 52.56$; $p < .001$). There was a statistically significant difference for the number of increased assessment scores ($\chi^2 = 62.76$; $p <$ .001) and for the number of decreased assessment scores ($\chi^2 = 286.07$; $p < .001$).

If the student retook an assessment, then that student was more likely to increase his or her assessment score. By participating in a review session, for the first semester, 76% of the students improved their scores an average of 15.83 points compared to 64% of the students who did not participate in a review session and who improved their scores an average of 7.72 points. For the second semester, 79% of the students improved their scores an average of 15.99 points by participating in a review session compared to 64% of the students who did not participate in a review session and who improved 8.97 points. The majority of the students did not participate in a review session before retaking an assessment for either semester; however, for the first semester, 68.98% of the students who retook assessments increased their scores, and, for the second semester, 66.28% of the students increased their assessment scores. Thus, two-thirds of the students who retook assessments increased their scores regardless of participation in a review session. Table 2 displays the frequencies for the change in assessment scores and for the participation in a review session by semester.

Table 2

Frequencies for Change in Assessment Scores and Participation in Review Session by Semester

	First Semester				**Second Semester**			
	Review		Without Review		Review		Without Review	
	n	%	n	%	n	%	n	%
Increased	612	76.21	807	64.35	842	79.06	1200	59.52
Decreased	155	19.30	383	30.54	182	17.09	678	33.63
Unchanged	36	4.48	64	5.10	41	3.85	138	6.85
Total	803	100.00	1254	100.00	1065	100.00	2016	100.00

Note: Frequencies may vary depending on available data.

Conclusions and Recommendations

The results of this program evaluation support the continued implementation of the FNO Policy at the high school and the implementation of Bloom's process of mastery learning in a traditional classroom. Nearly the entire student body participated in the policy at least twice during the school year. On average, the students increased their assessment scores from 16 to 18 points. This increase could potentially increase the students' final course grade as much as one letter grade if they scored less than 70% on the original assessment. By participating in a review session, the students were more likely to increase their assessment scores than those students who did not participate in a review session if they scored less than 70% on the original assessment. Change in assessment scores varied by department, but these differences could be contributed to varying content and assessment procedures and/or consistency of record keeping procedures.

The following recommendations are intended to improve the data collection procedures. There were inconsistent recordkeeping procedures along with incomplete data in numerous cases across departments. To improve record keeping procedures, (a) determine how to code review sessions conducted in class and those review sessions conducted after school, (b) determine how to gather complete data from all teachers, (c) determine whether to include those students who missed the major assessment due to absence or disciplinary suspension, (d) determine a procedure for recording scores for those students who retook assessments in the in-school suspension retake sessions, and (e) determine a procedure for those students who retook an assessment in class and whether that retake should be included in the spreadsheet.

Source:

Brown, J. L. (2012b). The impact of the failure is not an option policy on student grades. *Perspectives in Learning, 13*(1), 22-28.

An Evaluation of the GAPE Mini-Lesson Intervention

To address current preservice teachers' writing deficiencies and to better prepare them as future writing teachers, the purpose of this program evaluation was to determine the effects of a mini-lesson unit on knowledge of common grammar, audience, and punctuation errors among students enrolled in a teacher preparation program.

Methods

Participants

The participants were part of an introduction of teaching course, which was a requirement for admission to Teacher Education, at Columbus State University. The purpose of this course was to analyze the historical and philosophical influences that impact education in the United States, examine the legal and ethical requirements of the teaching profession, apply the various learning theories to classroom practice, and analyze effective instructional design, delivery, and assessment within the classroom setting. The pre-test and posttest assessments were matched for 12 students. Of the 12 students, there were 10 (83.3%) females and 2 (16.7%) males. In terms of racial classification, 9 (75%) students were white, and 3 (25%) students were black. The majority of the students had a declared major in early childhood education (n = 8) followed by special education (n = 2), secondary education: history (n = 1), and physical education (n = 1).

Data Collection

Procedures. On the first day of class, the students were administered a 10-item pretest contained four commas errors, two pronoun errors, and four audience errors (e.g., use of contractions), which were the most common errors within student writing. On the last day of class, the students were administered a posttest with the same 10 items. The pretest and posttest scores were compared to determine the effectiveness of the GAPE mini-lessons. As a follow-up, the students were asked to reflect on the GAPE mini-lessons on the end-of-the-semester course evaluation.

Intervention. The GAPE (Grammar, Audience, and Punctuation Errors) mini-unit was developed by the course instructor to improve grammar, audience, and punctuation errors within an introduction to teaching course. Written assignments for the course are expected to be without grammar and punctuation errors and presented with a formal writing tone based on APA (6th edition) Style Guidelines; however, student writing assignments collected over the past three semesters indicated that many students are submitting written work still containing numerous mechanical errors. As a result, the instructor developed an ongoing pedagogical strategy (GAPE) to address the most commonly occurring mechanical errors. At the beginning of each subsequent class, the students were given two sentences as a bellringer. These sentences were a representative sample of typical writing submitted for the introduction to teaching course. The students were directed to locate and correct the grammar, audience, and/or punctuation errors. If the sentence was correct, they were to write "correct". The sentences were presented on the Promethean Board and within their daily class handouts. Then, the instructor reviewed each sentence by asking one of the students to come to the Promethean Board and correct the error. Afterwards, the instructor offered other variations to correct the similar errors (e.g., a run-on sentence can be correct with a period, comma and conjunction, or a semi-colon). The errors include similar issues placed on the pre/posttest.

- Ambiguous pronouns ($n = 3$)
- Coordinating conjunctions and comma usage ($n = 4$)

- Run-on sentences with two or more independent clauses (*n* = 5)
- Direct quotes within the text (*n* = 2)
- Use of contractions in formal writing (*n* = 3)
- Repetitive word structure (*n* = 1)
- Comma usage with introductory dependent clauses (*n* = 2)
- Noun/pronoun agreement (*n* = 2)
- Use of colloquial expressions (*n* = 2)
- Comma usage with series of three or more items (*n* = 3)

In addition, the bellringers addressed the following issues:

- Appropriate word usage (*n* = 4), such as effect/affect
- Essential and non-essential clauses along with comma usage (*n* = 5)
- Comma usage with compound predicates (*n* = 2)
- Beginning a sentence with a conjunction (*n* = 2)

Data Analysis and Results

A series of descriptive analyses were conducted to summarize the pre-test and posttest scores. The number of correct items on the pretest ranged from 1 to 7 with a mean of 3.7 and a standard deviation of 1.6. Considering the wide range of dispersion, the median was 3.5. On the posttest, the number of correct items ranged from 2 to 9, with a mean of 7.2 and a standard deviation of 2.3. The median was 8. (*Note*: Two students did not complete the backside of the posttest.) On average, the students increased their recognition of grammar, audience, and punctuation errors by 94.6%. A paired samples *t*-test was conducted to determine if a significant change in knowledge occurred between the pretest and posttest. There was a significant increase between the two assessments, $t(11) = 5.66$; $p < .001$. Two of the reoccurring comma issues within the posttest were using a comma with two parts of a compound predicate and using commas with a series of three of more items. In addition, some of the students did not recognize formal writing avoids the use of colloquial expressions and onomonopia.

When asked to comment on the GAPE bellringers on the course evaluations, the majority of students had favorable reviews of the mini-lessons. One student responded, "It helped me think before I write." Another student said, "…they helped me remember things I'd forgotten and taught me things I should've [known] already." The results indicate that the mini-lessons improved the recognition of common grammar, audience, and punctuation errors. It is hopeful that the quality of writing will improve as the students generalize the recognition into practice within their written assignments.

Conclusions and Recommendations

We recommend education faculty continue to provide students support as they develop and refine their writing skills by employing best practices for teaching writing within these entry-level classes. Such best practices may include modeling the writer's workshop, guiding students in peer review workshops, providing students with one-on-one writing assistance, and embedding writing mini-lessons within instructional time. As evident from this study, students need practice with transferring and generalizing the skills into other settings. We recognize that it is not feasible (or desirable) for education faculty to modify their course learning outcomes or content to the extent that these courses become "writing courses" per se. In order to help students receive the amount of writing support needed for them to improve their skills (and pedagogical practices), we recommend improving collaboration among English composition and education faculty members.

Source:

Brown, J. L., & Bentley, E. (2013). Do other people "gape" at your writing? *National Teacher Education Journal, 6*(3), 33–36.

F. Program Evaluation Report Example #3

An Evaluation of the Fall 2012 Semester Freshman Learning Community for EDUC 2130

The fall semester of the 2012-2013 academic year was the first time that a freshman learning community was offered for the EDUC 2130 (Exploring Learning and Teaching) course at Columbus State University. The instructor for EDUC 2130 was Dr. Evelyn Blalock, and the course was paired with a section of ENGL 1101 (English Composition 1), which was taught by Mrs. Sundi Rose-Holt. The purpose of this evaluation was to determine the effectiveness of the mentor program and Freshman Learning Community format on students who enrolled in EDUC 2130.

Methods

At the end of the semester, the students in EDUC 2130 received surveys to evaluate the freshman learning community experience and specific course content and components. The evaluation items varied depending on the specific activities that occurred within the course. This evaluation report presents the findings of these surveys and offers conclusions and possible implications for future freshman learning communities of this type.

Participants

The majority of the students in EDUC 2130 were traditional-aged students. There were a few transfer and/or non-traditional students. Of these students, 5 (20.8%) were males, and 19 (79.2%) were females. The officially declared majors among these students varied. Nine students (37.5%) were Early Childhood Education majors. The remaining students were Fine Arts (n = 2), Middle Grades Education (n = 2), Physical Education (n = 1), Secondary Education (n = 1), and Special Education (n = 1) majors. In addition to the education majors, the declared majors included Criminal Justice (*n* = 5), Computer Science (*n* = 1), Biology (*n* = 1), and Undeclared (*n* = 1).

Data Collection Procedures

Mentor Program. The eight students who participated in the mentoring program for EDUC 2130 along with the two mentors were emailed a link on Tuesday, November 13, 2012, for a web-based survey in Qualtrics. A reminder email was sent on Tuesday, November 20, 2012. The survey contained 13 items to evaluate the mentor program and experience. The respondents were not given an incentive for completing the survey.

All Students. The students in EDUC 2130 were administered a paper-pencil survey on Thursday, November 15, 2012, at the beginning of the regularly scheduled class meeting by a faculty member who was not the teacher of record. The survey contained 15 items about demographics, field experience, lesson planning and implementation, and evaluation items for the freshman experience.

Data Analysis and Results

Mentor Program

As a pilot program, eight mentees were divided into two groups; each group was assigned to one of two mentors who have served as University Supervisors through the SAFE Office. The mentor met with each mentee at the respective field placement throughout the semester. These mentees were enrolled in the EDUC 2130 course with a declared major in Early Childhood Education. Of the eight mentee students, five students completed the web-based survey. A series of descriptive and frequency statistics were conducted to analyze the survey responses. For the five open-ended items, a content analysis was conducted to analyze the data. When asked if the mentor responsibilities were clearly defined, one mentor responded *Strongly*

Agree, and the other mentor responded *Disagree*. When asked to rate the overall mentor program, one of the mentors responded *Fair*, and the other mentor responded *Good*. For the mentees, the responses ranged from 2 (*Fair*) to 4 (*Excellent*) with a mean of 3.20 with a standard deviation of 0.84. The responses given by the mentors when asked to describe the relationship with the mentees ranged from 2 (*Fair*) to 3 (*Good*), but the responses given by the mentees ranged from 1 (*Poor*) to 4 (*Excellent*) with a mean of 2.80 and a standard deviation of 1.10. The variation may have resulted from the sample size.

One of the mentors met with her mentees on a weekly basis, and the other mentor met with her mentees on a monthly basis. One of the mentees responded that there were "no set meeting times". Both mentors felt that the time spent with the mentees was not sufficient; however, both of the mentors felt the time spent was helpful. On the other hand, all of the mentees felt the amount of time was sufficient and helpful. These mentees did not desire more time with their respective mentor. The mentees responded that their mentor gave them constructive feedback and answered any questions. According to the mentors, the mentees were "eager" to learn and improve. Both mentors primarily discussed lesson planning and implementation with their mentees. These topics were reiterated by the mentees. One mentee stated, "I learned to pay more attention to some of the things that I was doing…", things "that could have been done differently…", and things "that I didn't realize on my own". The mentees liked the additional resource within the classroom to offer "much insight". The mentees suggested that all of the students who were enrolled in EDUC 2130 should be assigned a mentor in the future.

All Students

In this section of EDUC 2130, all students were required to write a lesson plan, demonstrate it with their college classroom peers, and implement it within their field placement. The course instructor modeled several mini-lessons across multiple class meetings to prepare these novice students for this instructional activity. Of the 24 students, only 13 students completed the in-class paper-pencil survey. A series of descriptive and frequency statistics were conducted to analyze the survey responses. For the nine open-ended items, a content analysis was conducted to analyze the data. When asked to rate the field experience component for EDUC 2130, the responses ranged from 2 (*Fair*) to 4 (*Excellent*) with a mean of 3.23 and a standard deviation of 0.83. Some of the students commented that the field experience was "amazing", "excited", and "wonderful". Many of the student responded that they liked "getting [the] hands on experience" and "interacting with my students"; however, some students stated, "My cooperating teacher was not very good… She told me … she was ready to retire," and "the teacher assigned was not helpful." The overwhelming majority of students felt the lesson planning process was a good experience and appropriate training and support regarding best practices was provided. When asked the origination of the instructional lesson's idea, four students stated the sources as the cooperating teacher, one student stated the EDUC 2130 professor, one student stated peers, and the remaining seven students stated "I came up with it on my own". When asked if they would prefer to observe multiple classroom settings instead of one classroom placement, the responses ranged from 1 (*Strongly Disagree*) to 4 (*Strongly Agree*) with a mean of 2.53 and a standard deviation of 0.88. A few students felt the number of required hours for the field experience (i.e., 30) was difficult to complete. The suggestions for improvement included better communication of classroom expectations with the cooperating teachers and clarification of the background check process.

When asked about the overall first-semester experience, the students responses ranged from 2 (*Fair*) to 4 (*Excellent*) with a mean of 3.46 and a standard deviation of 0.66. When asked if the students planned to change their major within the next 6 months, 84.6% responded *No*. One of the two students who responded *Yes* changed from Early Childhood Education to Middle Grades Education, and the other student changed from Undeclared to Early Childhood Education. All of the responding students planned to remain at the University. The rationales included location, affordability, and specific degree programs, primarily education and theatre. One student stated the University has a "friendly environment".

Conclusions and Recommendations

Based on the data analysis, the following conclusions and possible implications were offered. First, the students indicated their first-semester experience was good. Nearly 90% of students planned to stay at the University and pursue the same declared major. Responses for remaining at this University included location, reasonable costs, faculty and peer relationships, and specific degree programs. From other data sources, unfortunately, the College will lose one out of every three students between now and next fall semester. With such a positive first-semester experience, further research is needed to determine the effect of the second semester on their intentions to stay. In addition, because this cohort was the first group of freshman students to participate in the EDUC 2130 Freshman Learning Community, further research is needed to determine the effect of these specific cohort classes on long-term retention and graduation rates, especially considering the faculty and peer relationships formed during the first semester. One note of difficulty was the large number of non-education majors enrolled in the EDUC 2130

Freshman Learning Community. Those students were excluded from the in-class paper-pencil survey; therefore, it was difficult to assess the impact of the EDUC 2130 course on these non-education majors, which included one-third of the total student enrollment.

The students indicated the hands on and interactive experiences were beneficial learning experiences in the classroom and in the field placements. Moving forward, it is necessary to engage these students in more kinesthetic and applicative activities to motivate their continued success, such as the lesson planning and implementation activity. Particularly in the EDUC 2130, the students indicated the desire to view multiple classroom settings. By offering more of a "fish bowl" approach during a lengthened class meeting, the students could see multiple teaching philosophies, observe various exemplar teachers, and reduce the additional field experience hours needed outside of the classroom. Furthermore, this approach could provide other sources for lesson plan origination.

In addition, the mentor program should be expanded to include more education students, and the specific responsibilities of the mentor should be outlined, including expectations for meeting with the mentees. Lastly, a process for better communication with the cooperating teacher is needed; such communication could be an email or written letter from the course instructor. The findings of this evaluation revealed the success of the Fall 2012 Semester Freshman Learning Community for EDUC 2130 and achieved the primary goal of increasing the freshman students' enthusiasm about their future profession.

Source:

Brown, J. L. (2012a). *An evaluation of the fall 2012 semester freshman learning communities for EDUC 2120 and EDUC 2130.* Unpublished manuscript, Department of Teacher Education, Columbus State University, Columbus, Georgia.

G. Program Evaluation Report Example #4

College of Education and Health Professions Longitudinal Retention Study of Freshman Cohorts Entering 1999 through 2014

Executive Summary

During the last 3 years, the program evaluation team worked to "clean up" the longitudinal database, particularly students who enrolled at Columbus State University (CSU), left the university, and returned as either undergraduate or graduate students. These student cases caused a possible skewness in the data. This database was recreated in 2012 using pre-existing data requested from Institutional Research, which caused some errors with previous cohort data that had to be cleaned up manually. Of the 140 identified student cases, 32 of them had graduated with their bachelor's degree and returned to the university for either post-baccalaureate work or graduate studies. Another 32 students left the university and transferred to other institutions before returning to the university. The remaining 76 students "stopped out" then returned to the university. This manual search and correct process was time-consuming; however, it allowed for a more accurate depiction of retention, progression, and graduation within the College of Education and Health Professions.

Based on the recent data analysis, the number of students who declared a major within the College decreased over the past 4 years, from 298 in 2011 to 218 in 2014. The cohort demographics and pre-college aptitude characteristics (i.e., high school grade point average [GPA] and standardized test scores) remained relatively stagnant among students who declared an initial major within the College. Notably, there was an increased percentage of continuing-status students within the last four cohorts. In addition, the first-semester and first-year GPAs remained relatively unchanged along with their relationship with the final CSU GPAs. A strong

relationship existed between first-semester GPA and final GPA ($r = .84$; $p < .001$) and between first-year GPA and final GPA ($r = .893$; $p < .001$).

Retention rates appeared to be increasing, but this trend cannot be confirmed until the pattern continues over multiple years. Similar trends have occurred since 1999. The overall retention rate followed an exponential decay model with 3 out of every 4 students returning each year. Graduation rates appeared to be relatively unchanged. On average, 22.4% of the cohort students graduated with their initially declared major. Another 8.5% changed their majors but remained in the College, and an additional 9.8% graduated with a degree from another college. The cumulative graduate rate from the university was 40.7%.

As part of this project and another research project, the program evaluation team examined some individual programs within the College of Education and Health Professions, particularly nursing, who has a low retention rate after the second year (from 57.5% to 49.7%). This time period is the nursing admission milestone. If the student was not accepted into the program, he or she tended to change majors or leave the university. This further examination could be beneficial for other programs within the College to determine possible reasons for attrition. The length of time between initial enrollment and graduation was 4.64 years for all cohort students. Nursing majors had the shortest length of time between enrollment and graduation ($M = 4.58$). One reason could be the prescribed pre-nursing curriculum and nursing curriculum, which does not exist with some of the other programs within the College.

Another major task completed within the last 3 years was the utilization of the National Student Clearinghouse data, which accounted for students who were denoted as "dropping out" in the previous 2012 report. These data were collected as part of a data sharing agreement between Dr. Brown and the Board of Regents. After the data were obtained and the database was revised, most of the “drop out” students were categorized as transfer students. Nearly 35% of each cohort will transfer to another institution, typically during the first 2 years of study. Of this group, over one-third will transfer to a technical college (e.g., Columbus Technical College), and over 37% will transfer to other 4-year institutions in Georgia, primarily in the Atlanta area (e.g., Georgia Perimeter, Georgia State University, and Kennesaw State). Students who graduated outside the College (*M* = 3.08) and students who transferred had lower GPAs (*M* = 2.23) compared to students who graduated with their initially declared major (*M* = 3.38). Academic reasons may contribute to their decision to change majors or transfer to another institution. Similar trends were found by cohort, gender, racial classification, majors, and parents’ level of education with retention rates and graduation rates. The freshman year experience continues to have the greatest influence on retention, progression, and graduation rates within the College as evident from the strong relationship between the first-year and final CSU GPAs. While pre-college aptitude characteristics (e.g., high school GPA and SAT scores) contribute moderately to academic success, the connections made with fellow students, staff members, and faculty tends to have a greater impact on student retention as evident by the high retention rates among the Fine Arts majors.

Methods

Participants

The purpose of this program evaluation was to examine longitudinal trends that affect retention, progression, and graduation rates within College of Education and Health Professions at Columbus State University. The inclusion criteria for the sample were incoming freshman students who enrolled in CSU during fall semester 1999 through fall semester 2014 and declared a major within the College, which resulted in 3,357 students within the database.

Data Collection Procedures

An eQuest was submitted to the Office of Institutional Research at CSU to obtain the demographic, pre-college aptitude characteristic, retention, graduation, and GPA data. The data regarding transfer institutions were obtained through the Office of Research and Policy Analysis for the University System of Georgia (USG). The name and date of birth of students who left CSU without graduating were consolidated into one Excel file and submitted to the USG. The Office of Research and Policy Analysis denoted whether the students enrolled in another institution during the two fall semesters following their last semester completed at CSU (e.g., if the student completed the spring 2008 semester, his or her transfer status was tracked during fall semester 2008 and fall 2009 semester) using the National Student Clearinghouse database. If the student transferred to more than one institution during that time period, then the first transfer institution was used. The same data collection procedures were repeated during the summer semester after each academic year to obtain new fall cohort data and update student cases that were categorized as "still enrolled".

Using the collected data, a longitudinal case was created for each student who enrolled as a first-time freshman and declared a major within the College, which tracked his or her retention, progression, and graduation while enrolled continuously at CSU. If the student appeared in the database more than once (e.g., students with double majors), one data entry was eliminated based on the graduation status.

To "clean up" the database that was recreated in 2012, the program evaluation team selected all students who were enrolled 5.5 years or more within the database. A total of 140 student cases were examined during this procedure. For each of these students, their 909 number was entered into ISIS, and the undergraduate transcripts were examined for any break in continuous enrollment during the fall-spring academic year. If there was a break, the last semester completed was changed within the database along with the final CSU GPA.

Data Analysis

Once the database was created and updated, a series of frequency and descriptive statistics were conducted to examine trends by cohort, gender, racial classification, initially declared major, and parents' level of education. A series of Pearson Product Moment Correlations were conducted to determine the strength of bivariate relationships between first-semester, first-year, and final CSU grade point averages.

Results

Cohort Demographics

Table 1

Frequency and Percentages of Gender and Racial Classification by Cohort

Cohort	Female	Male	White	Black	Other	Cohort Total
1999	88 (80.0%)	22 (20.0%)	73 (66.4%)	29 (26.4%)	8 (7.3%)	110 (100.0%)
2000	106 (76.3%)	33 (23.7%)	101 (72.7%)	30 (21.6%)	8 (5.8%)	139 (100.0%)
2001	137 (78.7%)	37 (21.3%)	121 (69.5%)	38 (21.8%)	15 (8.6%)	174 (100.0%)
2002	136 (80.0%)	34 (20.0%)	116 (68.2%)	44 (25.9%)	10 (5.9%)	170 (100.0%)
2003	157 (73.7%)	56 (26.3%)	143 (67.1%)	56 (26.3%)	14 (6.6%)	213 (100.0%)
2004	175 (81.0%)	41 (19.0%)	140 (64.8%)	58 (26.9%)	18 (8.3%)	216 (100.0%)
2005	185 (80.1%)	46 (19.9%)	152 (65.8%)	52 (22.5%)	27 (11.7%)	231 (100.0%)
2006	157 (78.9%)	42 (21.1%)	119 (59.8%)	58 (29.1%)	22 (11.1%)	199 (100.0%)
2007	122 (74.8%)	41 (25.2%)	110 (67.5%)	38 (23.3%)	15 (9.2%)	163 (100.0%)
2008	150 (76.9%)	45 (23.1%)	131 (67.2%)	50 (25.6%)	14 (7.2%)	195 (100.0%)
2009	188 (80.7%)	45 (19.3%)	147 (63.1%)	65 (27.9%)	21 (9.0%)	233 (100.0%)
2010	198 (82.8%)	41 (17.2%)	137 (57.3%)	86 (36.0%)	16 (6.7%)	239 (100.0%)
2011	233 (78.2%)	65 (21.8%)	144 (48.3%)	128 (43.0%)	26 (8.7%)	298 (100.0%)
2012	225 (76.5%)	69 (23.5%)	147 (50.0%)	115 (39.1%)	32 (10.9%)	294 (100.0%)
2013	208 (78.5%)	57 (21.5%)	139 (52.5%)	94 (35.5%)	32 (12.1%)	265 (100.0%)
2014	158 (72.5%)	60 (27.5%)	130 (59.6%)	69 (31.7%)	19 (8.7%)	218 (100.0%)
Total	2,623 (78.1%)	734 (21.9%)	2,050 (61.1%)	1,010 (30.1%)	297 (88%)	3,357 (100.0%)

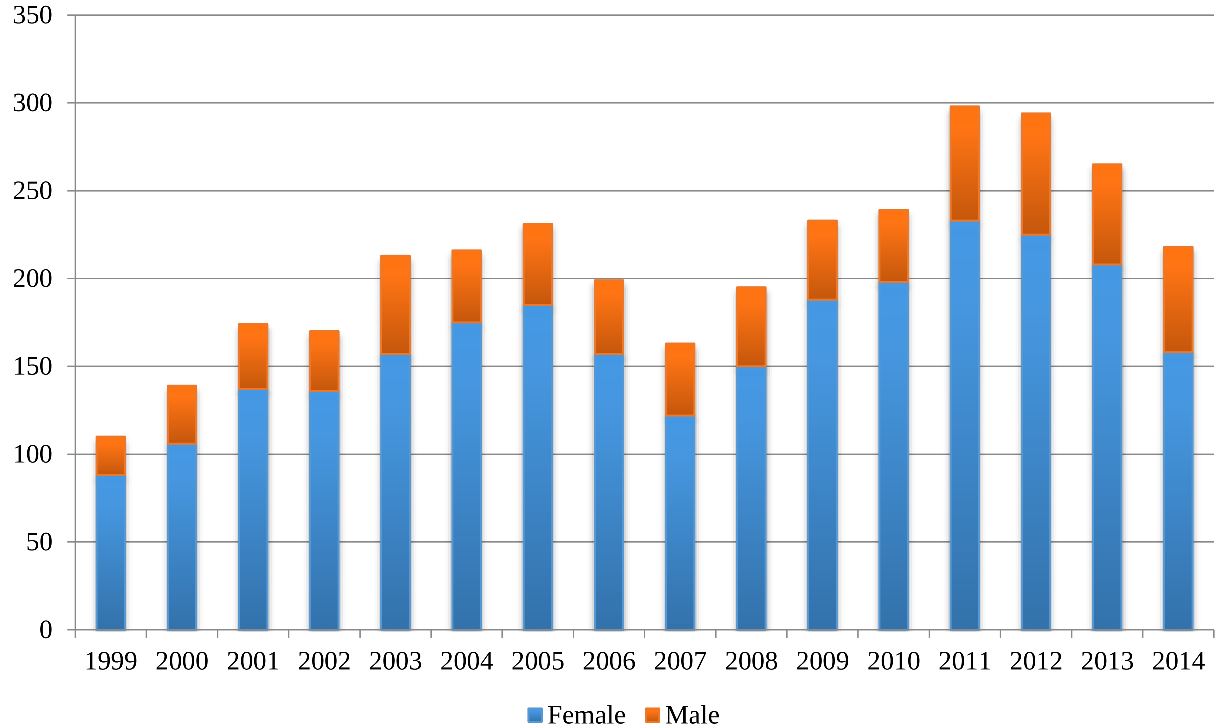


Figure 1. Frequency of Gender by Cohort.

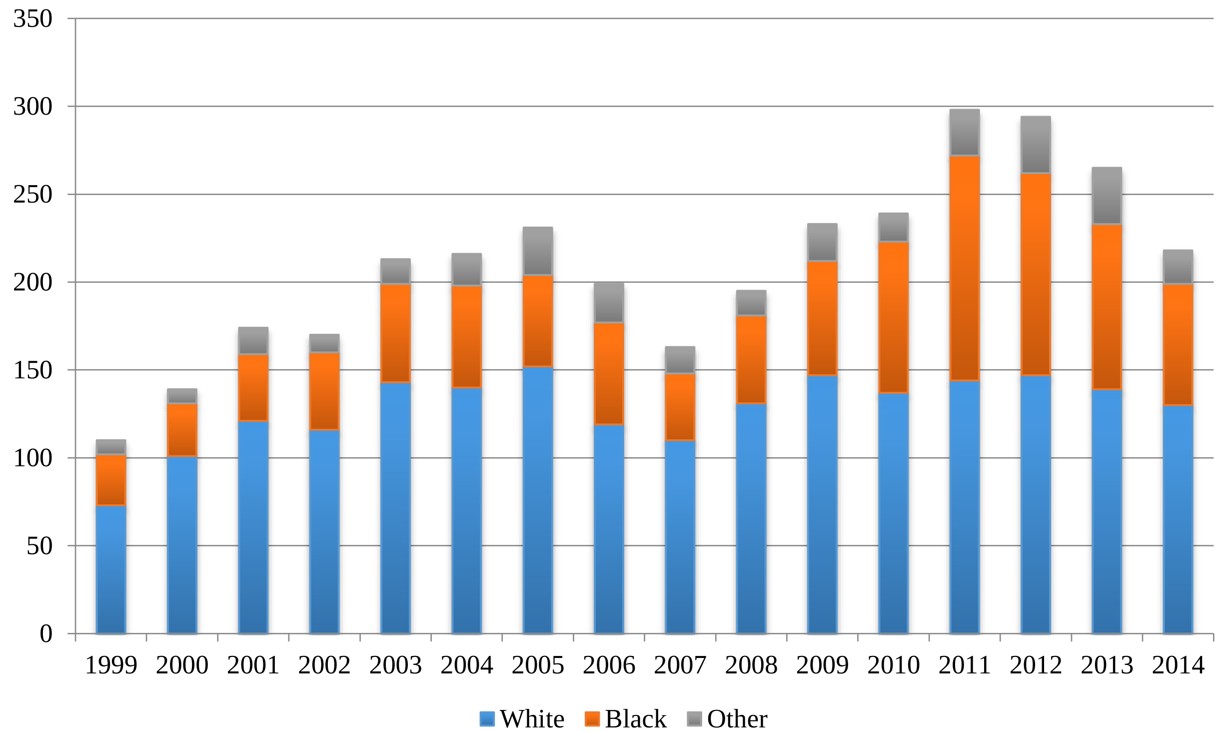


Figure 2. Frequency of Racial Classification by Cohort.

Pre-College Aptitude Characteristics

Table 4

Means and Standard Deviations for High School GPA and Standardized Test Scores by Cohort

Cohort	High School GPA		SAT Math		SAT Verbal/ Critical Reading		ACT Composite	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1999	3.02	0.49	476.08	68.38	497.35	59.54	20.45	3.53
2000	3.16	0.49	477.44	71.17	492.78	66.63	18.90	2.89
2001	3.16	0.55	485.13	69.50	502.34	67.08	19.43	2.78
2002	3.11	0.45	480.31	63.04	490.82	61.00	19.46	2.59
2003	3.17	0.46	490.10	71.06	498.60	70.23	20.28	3.65
2004	3.16	0.51	493.14	63.62	505.39	68.73	20.28	3.47
2005	3.16	0.45	503.53	69.67	510.35	64.22	19.86	2.81
2006	3.19	0.50	503.65	65.07	512.92	66.38	20.84	2.92
2007	3.11	0.49	497.46	64.90	511.90	72.38	20.89	3.47
2008	3.14	0.43	502.24	69.01	501.30	68.04	20.67	2.95
2009	3.18	0.43	499.50	69.37	509.11	69.71	20.53	2.89
2010	3.14	0.43	487.07	69.93	493.04	63.21	19.96	3.00
2011	3.15	0.43	482.08	73.03	493.81	73.47	19.77	2.99
2012	3.19	0.44	479.91	75.41	490.76	80.15	20.34	3.11
2013	3.19	0.57	485.44	79.33	494.61	76.64	20.24	3.60
2014	3.18	0.46	482.20	74.72	493.40	61.31	20.67	3.48
Total	3.16	0.47	489.46	70.68	499.84	69.18	20.22	3.18

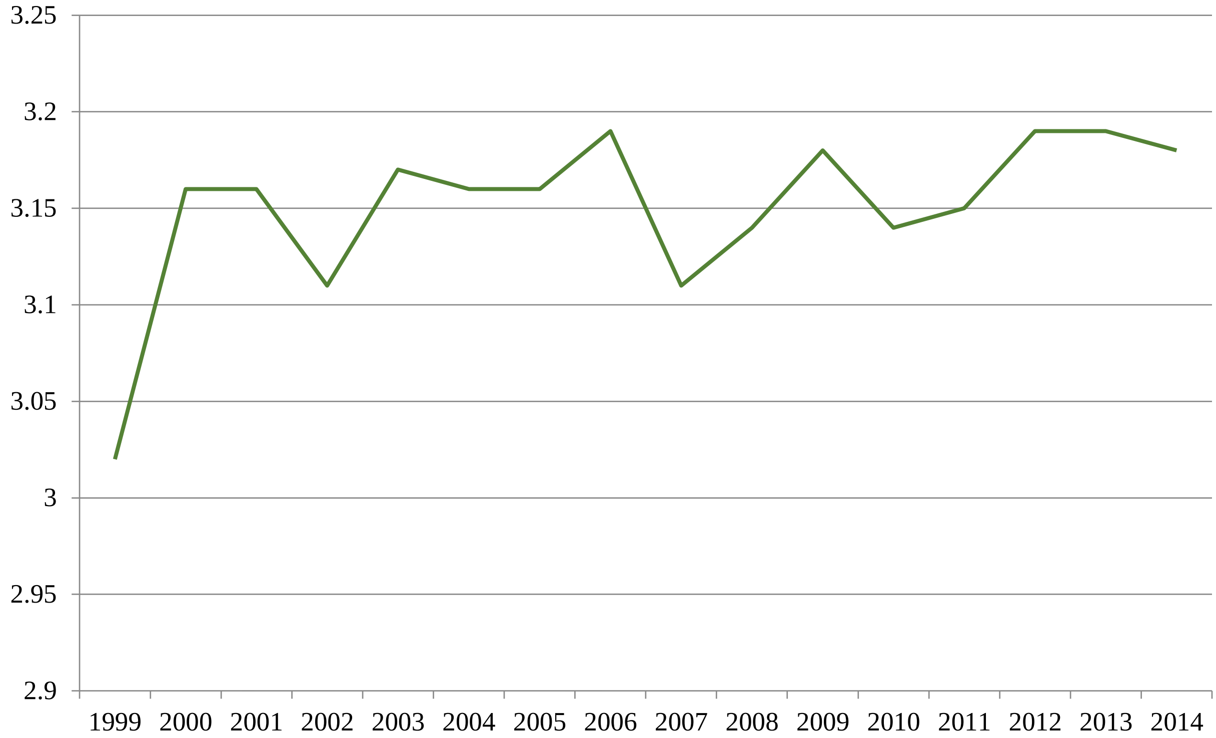


Figure 5. Means for High School Grade Point Average by Cohort.

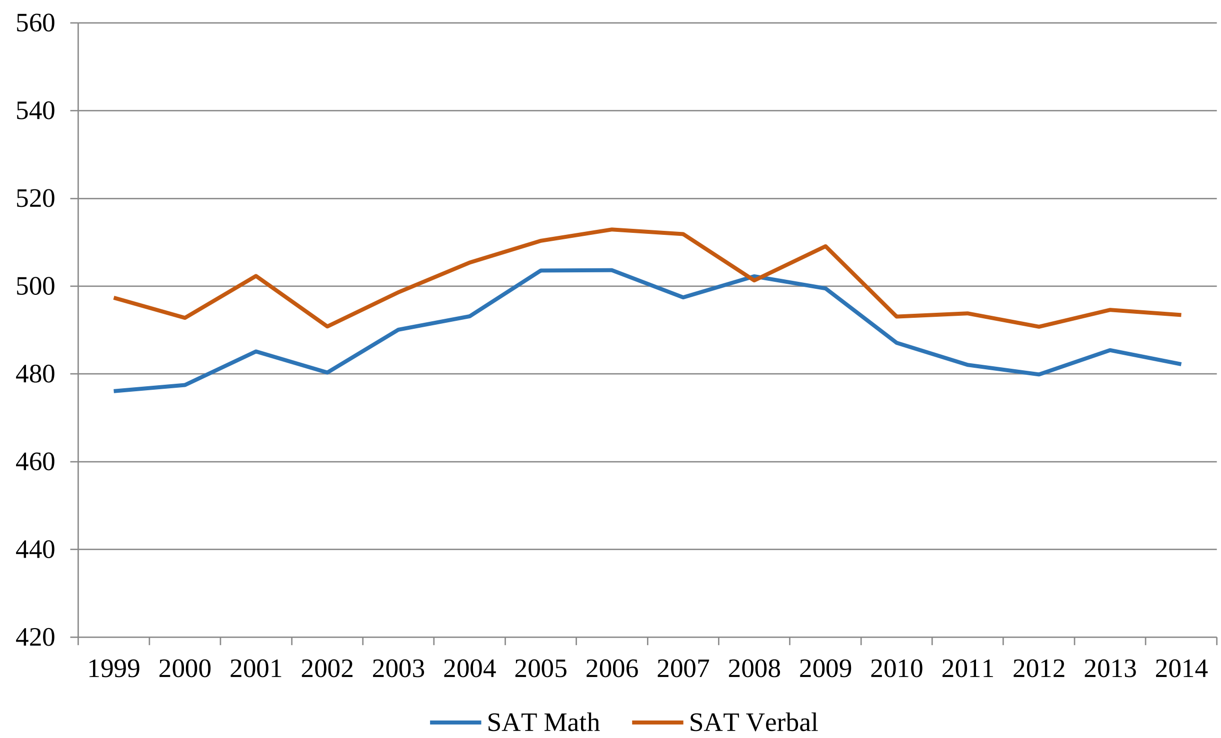


Figure 6. Means for SAT Math and Verbal/Critical Reading by Cohort.

Retention Rates

Table 15

Frequency and Percentages of Annual Retention Rates by Cohort

Cohort	Second Year	Third Year	Fourth Year	Fifth Year
1999	89 (80.9%)	63 (57.3%)	57 (51.8%)	34 (30.9%)
2000	102 (73.4%)	84 (60.4%)	75 (54.0%)	48 (34.5%)
2001	138 (79.3%)	112 (64.4%)	93 (53.4%)	59 (33.9%)
2002	19 (81.8%)	109 (64.1%)	93 (54.7%)	64 (37.6%)
2003	168 (78.9%)	136 (63.8%)	113 (53.1%)	72 (33.8%)
2004	182 (84.3%)	143 (66.2%)	125 (57.9%)	80 (37.0%)
2005	181 (78.4%)	144 (62.3%)	121 (52.4%)	71 (30.7%)
2006	165 (82.9%)	131 (65.8%)	111 (55.8%)	53 (26.6%)
2007	133 (81.6%)	108 (66.3%)	95 (58.3%)	44 (27.0%)
2008	144 (73.8%)	125 (64.1%)	112 (57.4%)	67 (34.4%)
2009	184 (79.0%)	147 (63.1%)	127 (54.5%)	73 (31.3%)
2010	174 (72.8%)	120 (50.2%)	106 (44.4%)	68 (28.5%)
2011	217 (72.8%)	166 (55.7%)	146 (49.0%)	--
2012	226 (76.9%)	184 (62.6%)	--	--
Total	2,242 (78.0%)	1,772 (61.7%)	1,374 (53.3%)	733 (32.1%)

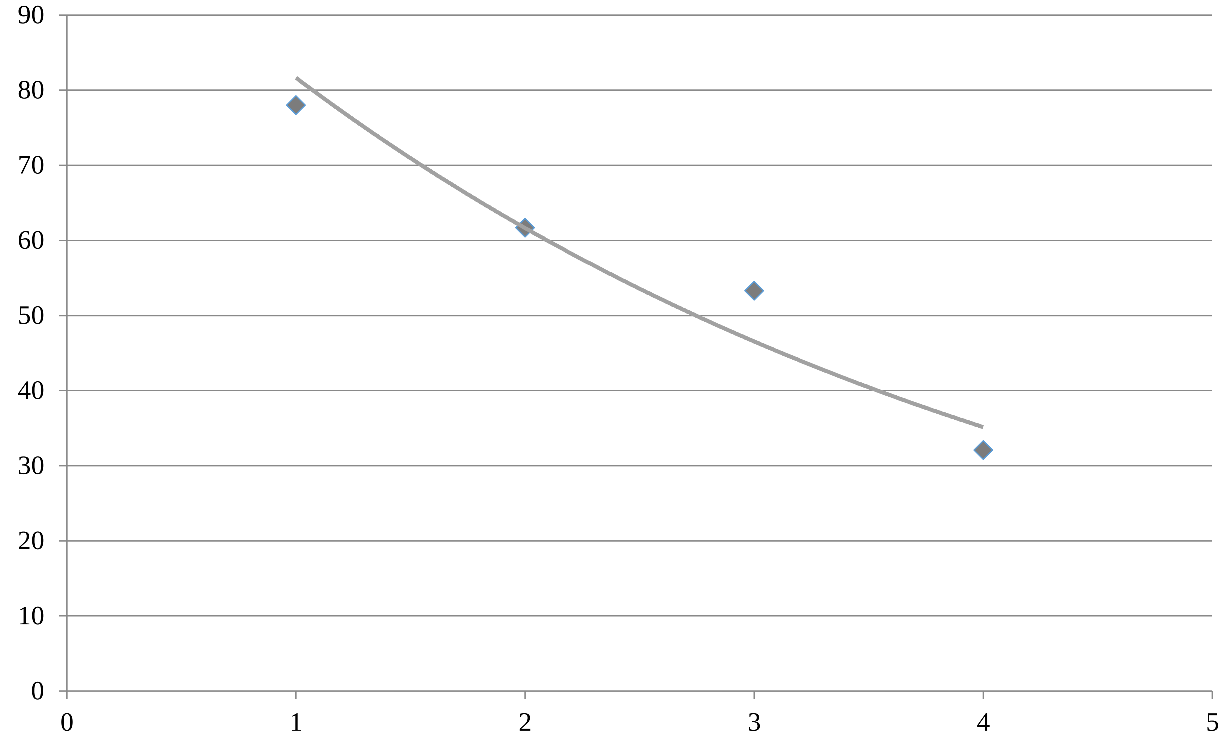


Figure 25. Exponential Decay Trendline for Cumulative Annual Retention Rates.

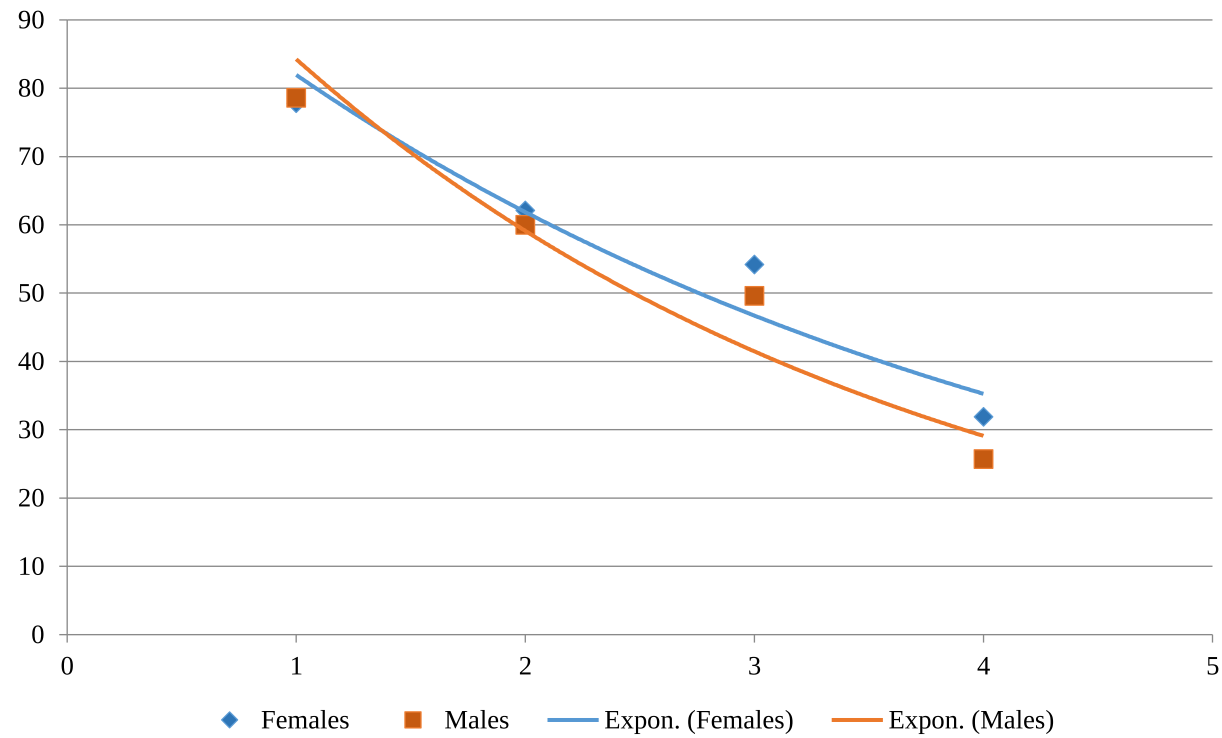


Figure 26. Exponential Decay Trendline for Cumulative Annual Retention Rates by Gender.

Graduation Rates

Table 20

Frequency and Percentages of Graduation Rates by Cohort

Cohort	Graduated with Initially Declared Major	Graduated within the College	Graduated outside the College	Did not Graduate (GPA < 2.0)	Did not Graduate (2.0 ≥ GPA > 3.0)	Did not Graduate (GPA ≥ 3.0)	Transferred to Another Institution	Cohort Total
1999	22 (20.0%)	11 (10.0%)	11 (10.0%)	14 (12.7%)	16 (14.5%)	2 (1.8%)	34 (30.9%)	110 (100.0%)
2000	24 (17.3%)	12 (8.6%)	23 (16.5%)	13 (9.4%)	19 (13.7%)	3 (2.2%)	45 (32.4%)	139 (100.0%)
2001	38 (21.8%)	12 (6.9%)	25 (14.4%)	15 (8.6%)	16 (9.2%)	5 (2.9%)	63 (36.2%)	174 (100.0%)
2002	40 (23.5%)	19 (11.2%)	15 (8.8%)	12 (7.1%)	11 (6.5%)	3 (1.8%)	70 (41.2%)	170 (100.0%)
2003	50 (23.5%)	20 (9.4%)	22 (10.3%)	11 (5.2%)	19 (8.9%)	5 (2.3%)	86 (40.4%)	213 (100.0%)
2004	61 (28.2%)	21 (9.7%)	20 (9.3%)	9 (4.2%)	15 (6.9%)	5 (2.3%)	85 (39.4%)	216 (100.0%)
2005	47 (20.3%)	18 (7.8%)	29 (12.6%)	19 (8.2%)	19 (8.2%)	5 (2.2%)	94 (40.7%)	231 (100.0%)
2006	57 (28.6%)	9 (4.5%)	22 (11.1%)	12 (6.0%)	13 (6.5%)	5 (2.5%)	81 (40.7%)	199 (100.0%)
2007	48 (29.4%)	12 (7.4%)	13 (8.0%)	8 (4.9%)	20 (12.3%)	6 (3.7%)	56 (34.4%)	163 (100.0%)
2008	52 (26.7%)	22 (11.3%)	14 (7.2%)	16 (8.2%)	18 (9.2%)	3 (1.5%)	63 (32.3%)	195 (100.0%)
2009	46 (19.7%)	24 (10.3%)	17 (7.3%)	22 (9.4%)	33 (14.2%)	7 (3.0%)	65 (27.9%)	233 (100.0%)
2010	27 11.3%	14 (5.9%)	12 (5.0%)	41 (17.2%)	31 (13.0%)	13 (5.4%)	51 (21.3%)	239 (100.0%)
Total	512 (22.4%)	194 (8.5%)	223 (9.8%)	192 (8.4%)	230 (10.1%)	62 (2.7%)	793 (34.8%)	2,282 (100.0%)

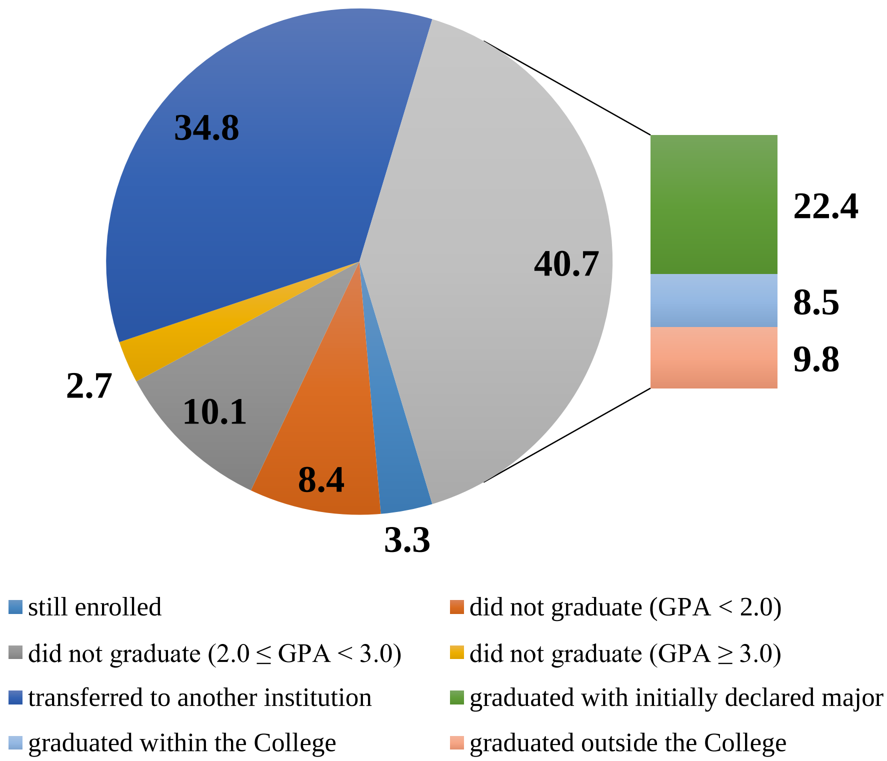


Figure 30. Percentage of Cumulative Graduation Rates for All Students.

Source:

Brown, J. L., & Andrews, A. (2015). *College of education and health professions longitudinal retention study of freshman cohorts entering 1999 through 2014.* Unpublished Manuscript, Department of Teacher Education, Columbus State University, Columbus, Georgia.

H. Program Evaluation Report Example #5

An Exploratory Study of the *Seven Principles for Good Practice in Undergraduate Education,* Academic Integration, and Subsequent Institutional Commitment

Multifaceted and complex problems, such as student persistence at commuter institutions, require more than one single solution. More attention should be focused on the events that occur inside the classroom, and the relationship between in-class and out-of-class experiences as they relate to academic integration and student persistence (Braxton, Bray, & Berger, 2000). The Seven Principles for Good Practice in Undergraduate Education is broad enough to be applicable across disciplines, teaching methods, learning styles, and institutional context yet they are grounded in research and practice (Sorcinelli, 1991). The purpose of this program evaluation was to examine the relationship between the Seven Principles for Good Practice in Undergraduate Education (Chickering & Gamson, 1987), academic integration, and subsequent institutional commitment for students who completed an education-based freshman learning community.

Methods

Participants

The sample consisted of first-time freshman students who were enrolled at Columbus State University during the Fall of 2012, who declared their major within a specific college, and participated in a freshman orientation and freshman learning community. Pseudonyms were assigned to participants to enhance anonymity. Participants included one traditional-aged White female (Michelle), one traditional-aged African American female (Vanessa), and one non-traditional aged White female (Sarah), who was married with three children. One participant, Michelle, lived on campus, and the other two participants lived at home in surrounding areas.

Data Collection Procedures

Focus groups were scheduled in the Spring of 2013 to gather additional information about the experiences of first-year student who completed the web based survey. A research proposal was submitted and approved by the Institutional Review Board at a southeastern university. At the end of the web-based survey, there was a question that asked the students would be interested in participating in an interview to gather additional information about the experiences of first-year students. If the respondent indicated *Yes*, then the researchers contacted the participants via email to schedule the interviews. The sessions were conducted in a meeting room within the College and lasted approximately 45 minutes. Initially, three focused groups were scheduled. One participant was interviewed during each scheduled session as several potential participants did not attend the focused groups that were scheduled during mutually agreed upon times. Handwritten notes were taken by both researchers during the interviews and were reviewed after interview sessions.

Data Analysis and Results

The research team analyzed the data that were collected and built a consensus on emerging primary themes and subthemes. Grounded theory (Glaser & Strauss, 1967) was utilized to guide the methodology.

Academic Integration

Academic integration consisted of the how students perceived the academic programs at the institution as well as their experiences with specific instructional methods that either enhanced or were deterrents to learning. As participants were asked to describe the culture or climate of the University, what they liked most and least about the University, and about the courses that they were enrolled in during the Fall and Spring semesters, they shared their perception of the academic programs at the University and their level of satisfaction with instructional methods. Academic integration appeared to be linked to the primary themes of student perceptions of academic programs and student satisfaction was connected to instructional methods.

Student Perceptions of Academic Programs

There was evidence to suggest that students' perceptions of the academic programs were linked to 1) class size; 2) campus resources as support; 3) academic factors related to the specific college environment; and 4) satisfaction that was connected to instructional methods.

Students' perceptions of the academic programs were linked to class size. Vanessa reported that what she liked most about the University was that the classes were small. She described this as, "the best part of the University." She reported that she enjoyed classes that ideally included 30 students.

Campus resources also emerged as a subject of students' perceptions of the academic programs. The campus resources appeared to be linked to services provided to assist students who need additional academic support. Sarah reported that the campus writing center provided her with academic support. Michelle identified math tutoring as a campus resource that she found helpful.

Another subject that emerged from students' perceptions of the academic programs was academic factors related to the specific college environment. These factors included the program of study and support provided through the Freshman Learning Communities (FLCs). Vanessa reported that she became aware of the teaching program at the institution from her eighth grade teacher. One of the reasons that Vanessa plans to continue at the University and within the College was based on the program's reputation. Sarah suggested that the FLCs assisted students in learning study strategies to be academically successful. In addition, Sarah felt the FLCs provided consistency for the students.

Participants described their satisfaction with the academic programs as being connected to instructional methods. Participants described satisfaction in courses in which instructors were "energized and animated," encouraged interaction, utilized active group discussions versus lectures, stopped to make sure that everyone understood the information before continuing, provided feedback, set clear expectations, were available for questions, asked open-ended questions, and explained concepts in different ways. Participants tended to be less satisfied with courses in which instructors were not focused on the topic of the course, there was limited interaction, instructors did not explain concepts, and lecture material was not included on the tests.

Student-Faculty Contact

Student interactions with faculty and staff was one Principle that emerged from the interview data. Participants described support from faculty and staff and willingness to seek support as factors that contributed to their interaction with faculty and staff. All three participants reported that overall they felt as if they received support from faculty and staff at the University. Comments made by participants suggested that perceived support may have been associated with faculty and staff making efforts to reach out to students, showing genuine concern for students, and being able to assist students when needed. One participant, Sarah, stated, "People are always thinking about you even though you have no idea they are there sometimes…I feel like I am being looked after and I feel like they are doing that. I have enjoyed the learning I am getting." Another participant, Vanessa stated, "I feel like my professors really reached out…my professors have been a big support for me." Sarah and Vanessa suggested that willingness to seek support is tied to academic success. Sarah stated that it is important that students are not afraid to ask for help. Vanessa stated, "They [instructors] are good at engaging and encourage us to ask questions, but if you are scared it can be a barrier...So many people don't want to ask questions…" Participants also suggested that it is important that students get to know the professors.

Collaboration among Students

Collaboration among Students was another Principle that emerged from the interview data. Participants suggested that the FLCs provided an opportunity for students to interact. Sarah reported that, as a non-traditional student, she believed that the FLCs were helpful for her, as well as for students who were just coming from high school. She stated that the FLCs helped to create an environment in which, "you don't feel like you're on your own….FLCs help with social interactions without even working at it…you don't realize they will be your support… it helps." Vanessa reported that she was able to meet two new friends as a result of the FLCs.

Conclusions and Recommendations

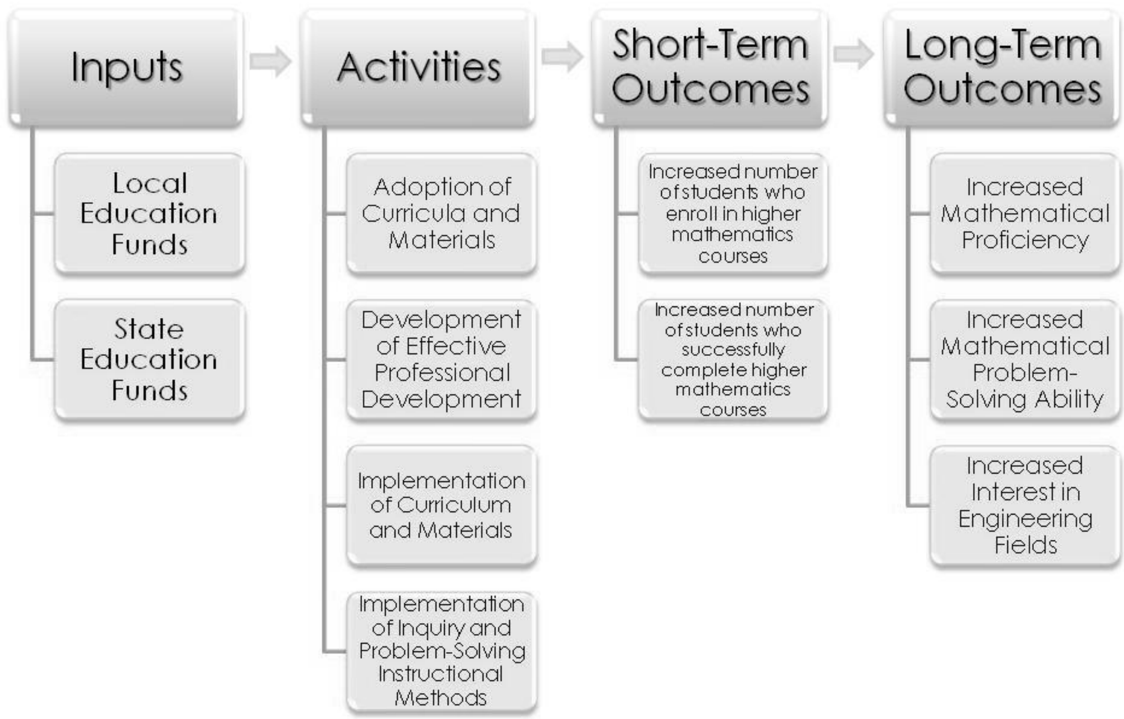
Although there have been numerous studies, which provide significant information on persistence of undergraduate students, this evaluation provided information specific to students enrolled in a commuter university and identified some possible factors that may be attributed to student persistence. Qualitative data suggested that academic integration included factors, such as students' perceptions of academic programs, class size, campus resources, academic factors related to the specific college environment, along with instructional methods.

This evaluation provides implications to educators and commuter institutions. The study suggests that factors that are connected to academic integration can possibly serve as a buffer to students who are enrolled in commuter institutions and thus impact student persistence. It also suggested that freshman learning communities can serve as a source of academic and social support for students. Students described experiences in which they learned specific strategies and were able to be connected with their peers as a result of being enrolled in freshman learning communities. There was also evidence to suggest that the Seven Principles of Good Practice in Undergraduate Education was connected with students' perceptions of their programs. Institutions could provide professional development to faculty regarding the implementation of the Seven Principles within the classroom. Their use requires little or no expenditure of money by an institution, and the faculty can learn and incorporate the Seven Principles into the classroom easily, especially if they participate in faculty development programs.

Source:

Brown, J. L., & Robinson-McDonald, D. (2014). An exploratory study of instructional strategies, academic integration, and subsequent institutional commitment. *Journal of Research in Education, 24*(2), 160-172.

I. Program Evaluation Proposal Example



Goal #1

To increase the mathematical proficiency of secondary students.

Objectives (Outcome)

1. To increase mathematical proficiency levels across implementation years and across mathematics courses based on benchmark examinations.
2. To increase advanced placement calculus scores across implementation years.
3. To increase graduation exit examination mathematics subtest scores across implementation years.

Goal #2

To increase the mathematical problem-solving ability of secondary students.

Objectives (Outcome)

1. To increase mathematical problem-solving abilities across implementation years and across grade levels.

Goal #3

To increase the interest in engineering fields.

Objectives (Outcome)

1. To increase the number of students who intend to major in engineering fields as they enter post-secondary institutions.
2. To increase the number of students who are admitted to a school of engineering.
3. To increase the number of students who graduate with a bachelor's degree in engineering.

Methods

Participants

The mathematics curriculum will be phased in over a 4-year period. The selected participants will be all high school students within the school district over the implementation period. The school district, with a total enrollment of 12,000, includes three high schools (grades 9 through 12) with an approximate enrollment of 3,490. The number of students increases an average of 2% each academic year. The gender classification is 48% male and 52% female. The racial make-up of the district is 54% White, 41% Black, and 5% who classify themselves as belonging to other racial groups. Eight percent of the students receive special education services. Fifty-nine percent of the students are eligible for free or reduced meals.

Intervention Activities

Description. The secondary mathematics curriculum will have an engineering focus. Each unit across all four courses (i.e., geometry, algebra II, pre-calculus/trigonometry, and advanced placement calculus AB) will have NCTM Standards-based expectations, one of more engineering connections (e.g., chemical, civil, electrical, or mechanical engineering), mathematical concepts involved with the unit topic, instructional goal(s), key terms, any required equipment needed for the unit, and a performance assessment. The performance assessment at the end of each unit will be a cumulating activity for the students to apply the mathematical concepts to the engineering field. For example, in the Pipeline Design unit for algebra II, the following task will serve as the culminating performance assessment: Given a specific terrain, design a pipeline to transport a golf ball. Include a scale drawing, calculations for intended transported material, estimated construction costs, and three-dimensional model. Develop an evaluation plan to test and assess your pipeline.

During the year prior to implementation, the evaluator and teachers will use curriculum units to develop instructional lessons and incorporate applicable lessons from their previous lesson materials. The geometry curriculum consists of six units: (a) land and water navigation, (b) horticulture/landscape design, (c) bridge building, (d) adaptive devices, (e) simple and complex machines, and (f) friction. The navigation unit covers the geometric concepts related to triangles and parallel lines. The horticulture unit covers the properties and theorems associated with circles. In the bridge building unit, the content includes three-dimensional shapes. The adaptive devices unit covers symmetry and transformations. For the simple and complex machines unit, the content includes Euclid's axioms and postulates. The friction unit focuses on the geometric concept of surface area. The algebra II curriculum consists of five units: (a) thermodynamics, (b) viral diseases, (c) HVAC systems, (d) cellular respiration, and (e) pipeline design. The thermodynamics unit covers addition of functions, inequalities, and transformation of functions. The viral diseases unit covers linear functions, systems of equations, and tree diagrams. The HVAC systems unit includes area and volume. For the cellular respiration unit, the content includes additive growth, multiplicative growth, and exponential equations. The pipeline design unit focuses on the geometric concepts of slope and rate of change.

The pre-calculus/trigonometry curriculum consists of seven units: (a) business plan, (b) electrical circuits, (c) wave motion, (d) aeronautical navigation, (e) optics, (f) introduction to statistics, and (g) dynamic systems. The business plan unit covers logarithms and bases and logarithmic functions. The electrical circuits unit covers the properties and applications of polynomials. In the wave motion unit, the content includes the trigonometric functions. The aeronautical navigation unit covers coordinate systems and vectors. The optics unit focuses on analytic geometry. In the introduction to statistics unit, the content includes the binomial theorem. The dynamic systems covers change with discrete dynamical systems, including constant, linear, and proportional change. The advanced placement calculus AB curriculum consists of five units: (a) water supply, (b) market growth, (c) amusement park design, (d) rocket design, and (e) loglinear analysis. The water Supply unit covers local linearity. The market growth unit covers functions and limits. The amusement park design unit includes the derivative and applications of differentiation. For the rocket design unit, the content includes the integral and applications of integration. The loglinear analysis unit focuses on transcendental functions.

Procedure. The evaluator will work with the high school teachers to develop two benchmark examinations, midterm and final examinations, for the geometry, algebra II, pre-calculus/trigonometry, and advanced placement calculus courses. In addition, the district personnel will create a mathematical problem-solving examination. The school administrative staff will gather the graduation exit examination mathematics subtest and advanced placement calculus examination scores. At the end of each academic year, the guidance office staff will collect the number of students who intend to major in engineering, the number of students who were admitted to a school of engineering, and the number of students who earned a bachelor's degree in an engineering field by contacting the former students.

Process Evaluation

Reach. The students' participation in the curriculum activities will be assessed using the teachers' daily attendance record.

Dosage. One hundred eighty lessons from the Mathematics Curriculum for Advanced Mathematical Proficiency will be taught in 55-minute sessions from August to May. When students are absent, they will receive make-up lessons. Each teacher will document that the lesson was taught in his or her daily lesson plan book. These daily lesson plans will be turned into the school administrative team for review.

Fidelity. With the weekly informal observation forms, school personnel will monitor the implementation process in the classroom. One of the following people will conduct these observations: school principal, assistant principal, curriculum director, or assistant curriculum director. A professional development workshop will be conducted to train the observers with the weekly informal observation form and behavioral checklist. Sample videos of classroom instruction will be utilized during the training workshop. After direct instruction and guided practice, independent practice will occur until the interrater reliability among the observers is consistent.

Professional Development Workshop. At each professional development workshop, all participants will complete an exit survey to determine the effectiveness of the session and determine future professional development needs. The exit survey was developed using a variety of preexisting instruments. Questions regarding instructional and student assessment methods were devised from the National Survey of Science and Mathematics Education (Westat, 2000). The areas of future professional development needs were based on the Local Systemic Change: Principal Questionnaire (Horizon Research, 2006). The items, which relate the importance for the skill to student success in mathematics, were collected from the Mathematics Teacher Questionnaire: Main Survey (TIMSS Study Center, 1998).

In addition, at each professional development workshop where lesson plans are developed, a lesson plan design rating system will be conducted. This rating system was adapted for this application using the Inside the Classroom: Observation and Analytic Protocol (Horizon Research, 2000). A team of three teachers who were not involved in the development of the lesson plan will review the lesson's design and content independently. Based on their evaluations and recommendations, the lesson plan will be revised or submitted to the curriculum unit.

Implementation Process. A formative evaluation will be conducted to assess the attitudes and instructional methods of the teachers throughout the implementation process. A demographic survey will collect information regarding education level, certification areas, and years of experience in public education. Qualitative interviews with the implementing teachers will ascertain their perceptions and gather feedback for program improvements. The series of interviews will be conducted during pre-planning, midterm, end of the course, and post-planning. Because adults are more likely to reject the new knowledge that contradicts their beliefs, the information gathered during these interviews will evaluate existing knowledge, beliefs, and motivations and will determine the extent to which the implementing teacher have ownership in the curriculum implementation process (Klingner, Ahwee, Pilonieta, & Menendez, 2003).

Outcome Evaluation

Participants

Comparison. During the academic year prior to implementation, the students who are enrolled in geometry, which will be primarily ninth- and tenth-grade students, will be assessed using the two benchmark examinations and the mathematical problem-solving examination. In addition, baseline information will be collected from the school's administrative staff regarding the scores from advanced placement calculus examinations and the scores from the graduation exit examination mathematics subtest. This grade ahead comparison will continue throughout the implementation process. Baseline information will be collected regarding the number of students during Year 0 who plan to major in engineering and the number of previous students who earned a bachelor's degree in an engineering field.

Intervention. Beginning with the second year of implementation, the students who are enrolled in geometry will be assessed using the two benchmark examinations and the mathematical problem-solving examination. In the third year of implementation, the students who are enrolled in algebra II will be assessed using the benchmark and mathematical problem-solving examinations. During the fourth year, the students who are enrolled in pre-calculus/trigonometry will complete the prescribed assessments and the graduation exit examination mathematics subtest. Lastly, in the fifth year of implementation, the students who are enrolled in AP calculus will complete the assessments and the AP calculus examination.

Design

To analyze the data associated with the implementation activities, a qualitative study of the implementing teachers and other faculty members' interview responses will monitor the effectiveness of the professional development workshops. Quantitative data will be analyzed using descriptives and frequencies.

Objective 1.1. With the scores from the midterm and final benchmark examinations, a 4 X 2 analysis of variance (ANOVA) will be conducted to determine if mathematical proficiency levels changed across implementation years and across mathematics courses. In addition, a sample of students who begin the geometry-calculus sequence in Year 0 will be tracked through Year 3 to assess mathematical proficiency with the comparison group. These results will be compared with the data from the students who begin the geometry-calculus sequence in Year 1 of the curriculum implementation. With a profile analysis, the repeated measure analysis will determine group differences and longitudinal trends between the intervention and comparison groups.

Objectives 1.2 and 1.3. To analyze the long-term outcomes for the Mathematics Curriculum for Advanced Mathematical Proficiency, with the scores from the advanced placement calculus examinations and the scores from the graduation exit examination mathematics subtests, longitudinal trends will be graphed using the percentage of passing scores and the average score with both examinations across the implementation years.

Objective 2.1. After the initial descriptives are assessed, a repeated measure ANOVA with one within-subject factor (time) and two between-subject factors (group and grade level) will be conducted to determine if mathematical problem-solving abilities have changed across implementation years and across grade level and group.

Objectives 3.1, 3.2, and 3.3. A frequency count of the number of students who intend to major in engineering at high school graduation, the number of students who were admitted to a school of engineering, and the number of students who earn a bachelor's degree in an engineering field will be assessed. Based on these frequency counts, a chi-square non-parametric analysis will be conducted to determine the observed numbers different from the expected numbers across implementation years.

Measures

Mathematical Proficiency. For summative evaluations, a benchmark examination will be given every 9 weeks to assess mathematical proficiency based on course content and performance standards. This measure will be created by the high school teacher staff and will contain items that are representative of the expectation instruction content for that time period. It will be a multiple-choice format that assesses conceptual and procedural mathematical knowledge.

Mathematical Problem Solving. At the end of the course, the mathematical problem-solving examination will be administered. The items for the mathematical problem-solving examination will be written, peer reviewed, field-tested, and data reviewed prior to placement on the final form. To training the evaluators and to ensure consistent scoring, a grade level group of educators who had extensive training and experience with the official scoring rubric scored student responses selected from the field test. When a consensus was reached among the scoring panel, these criteria responses were used to illustrate the scoring guide and the variety of possible solutions for each task during training and scoring.

In the spring of each academic year, the participants were given 45 minutes to complete the mathematical problem-solving examination. The examination consists of four tasks (i.e., one each from statistics and probability; algebraic relationships; measurement; and geometry). The students will be instructed to follow the student directions and to show all of their work. High school mathematics teachers will score the examinations after attending two days of training. At the training, the evaluators will work on the four sample tasks at their grade level. After further training with the criteria papers, each rater will qualify to score the examinations by accurately scoring a packet of examinations.

References

Horizon Research, Inc. (2000). *Inside the classroom: Observation and analytic protocol*. Chapel Hill, NC: Horizon Research, Inc.

Horizon Research, Inc. (2006). *Local systemic change: Principal questionnaire*. Chapel Hill, NC: Horizon Research, Inc.

Klingner, J. K., Ahwee, S., Pilonieta, P., & Menendez, R. (2003). Barriers and facilitators in scaling up research-based practices. *Exceptional Children*, *69*, 411-429.

TIMSS Study Center. (1998). *Mathematics teacher questionnaire: Main survey*. Chestnut Hill, MA: Boston College.

Westat. (2000). *National survey of science and mathematics education*. Rockville, MD: Westat.

Source:

Bell, J. L. (2008). *An examination of cognitive and non-cognitive factors and academic success in the pre-engineering curriculum at a four-year southeastern university* (Order No. 3333110). Available from ProQuest Dissertations & Theses Full Text. (304688037).