

Question 1: As a collaborator on research designed to investigate possible determinants of interest in mathematics among high school students, how would you respond to your fellow researcher's proposal? What changes would you propose and why?

While I would be excited to contribute to an investigation into predictors of high school students' interest in mathematics, I would have several cautions for my fellow researcher related to the design of his correlational study. My chief concern is that the researcher may inadvertently identify spurious relationships in his study. The researcher appears to have validated a measure of the response variable "interest in mathematics among high school students", but seems to be unsure of how he should go about establishing the role of explanatory variables in predicting the response variable. Identifying suitable explanatory variables can be achieved in several ways, but the researchers' plan to collect data on as many variables as possible in hopes of finding statistical significance is fundamentally flawed. In the suggestions that follow, I outline several concerns with the researcher's plan of study and suggest changes to the proposal that will increase the potential contribution of the research to mathematics education. In particular, I recommend that the researcher begin his research with a thorough review of literature and consider reposing the research question along the following lines: "How are the variables, X_1 , X_2 , ... X_n , related to high school students' interest in mathematics?"

Correlational, or cross-sectional, designs comprise the most common type of research in the social sciences (Frankfort-Nachmias, & Nachmias, 2000). This type of design is commonly associated with surveys and can be used to either explain or establish causal relationships between properties and dispositions. In this case, the researcher is interested in finding "possible determinants" of the disposition he is calling "interest in mathematics" among high school students. Thus, I would support the researcher's plan to conduct a correlational study. The next choice that needs to be made is whether the study will seek to find causal relationships or simply explain existing relationships. Because the researcher is interested in finding predictive variables, I would recommend a cross-sectional design that includes constructing a statistical model (such as a general linear model) that the researcher believes will explain a large percentage of the variation in high school students' interest in mathematics.

An appropriate place to begin finding appropriate predictive variables for the study is through a comprehensive review of relevant literature. My colleague is not the first person to research students' interest-level in mathematics, so prior studies can provide a means for identifying important independent variables. It is important to note that increasing the number of explanatory variables will not, per se, positively affect which variables are found to play a statistically significant role in predicting the disposition under investigation. In fact, the inclusion of too many independent variables could result in the researcher identifying spurious relationships. For example, if X_1 strongly impacts both Y and X_2 , then a statistical correlation between X_2 and Y might be found, even though it is possible neither variable impacts the other.

Once a set of possible independent variables has been found, the researcher must develop reliable and valid measures for the independent variables. These could take the form of surveys, which could be assessed using test-retest, parallel-forms, or split-half methods for measuring the reliability of the tool. Similarly, the validity of the tool should be estimated so that the researcher is confident that their instrument(s) are measuring the variables he is interested in studying.

Next, I would recommend that the researcher clearly identify the population to which he seeks to generalize. By narrowing the scope of generalizability to, say, high school students in the large urban areas, the researcher will make sampling procedures more reasonable. The

researcher could then use random sampling procedures to produce a representative sample of the population, in which case the survey could be administered and data gathering and analysis can begin. Since the study was well-designed, I predict that my research colleague will be pleasantly surprised during the data analysis phase. He will be able to test hypotheses and be reasonably confident that any statistically significant explanatory variables represent nonspurious predictors of interest in mathematics among the high school students the chosen population. In fact, the (now well-informed) researcher may find that a large percentage of the variation in students' interest in mathematics is explained by the theoretically driven model—thus resulting in quality research and many compliments at the next conference he goes to!

Question 2: You are interested in extending the results of Van Dooren, Verschaffel, and Onghena (2002) by creating and investigating the effectiveness of a treatment to help preservice secondary teachers use evaluations that are more closely adapted to the nature of the task. Describe a research design that includes a treatment and a way to measure the goal. Outline the strengths and weaknesses of the design.

Improving prospective teachers' evaluations of student-work is a difficult task for a mathematics educator because many people evaluate students' problem-solving work more favorably if the work closely resembles their own preferred strategies. However, as Van Dooren, Verschaffel, and Onghena (2002) suggest in their abstract, preservice secondary teachers are particularly apt to have difficulty evaluating students' problem solving efforts in ways that more closely resembles the nature of the problem. Thus, an instructional intervention may be needed to improve prospective secondary teachers' evaluation of student work according to this criterion. Given the relatively little time classroom time devoted to giving prospective teachers guidelines for assessment, the question becomes: How can we incorporate and investigate the effectiveness of an instructional unit that improves preservice secondary teachers' evaluations of student work in ways that more closely resembles the nature of the problem-solving task? In the proposed research that follows, I suggest a short experimental study using a Solomon Four Group design that I claim will accomplish the research goal. After detailing the study, I explain why a Solomon Four Group design is appropriate and discuss some strengths and weaknesses of the proposed experimental design.

I recommend a classic experiment to evaluate the impact of an instructional treatment that attempts to improve preservice secondary teacher's evaluation of student's problem-solving work. The population for the proposed study includes preservice secondary teachers at a large university with a large secondary mathematics education program. In the study, about 100 students will register for a required course in their program entitled Secondary Teaching Methods in Mathematics. Although this sampling procedure is non-random, I believe the resulting sample will be representative of the population.

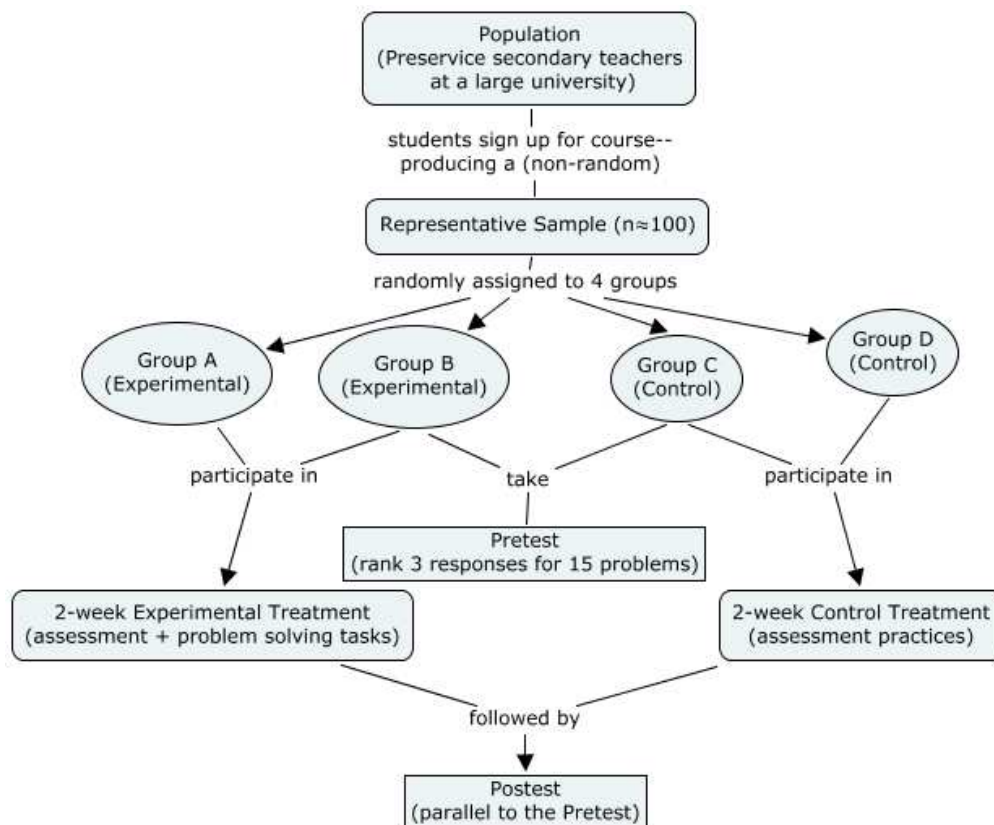


Figure 1. Proposed Solomon Four-Group experimental design to assess the effectiveness of a treatment to improve prospective secondary teachers' evaluation of student-work.

After the initial sample is gathered, the study will follow a variation of a classic experiment called the Solomon Four Group Design (See Figure 1 for a diagram of the research design). The 100 enrolled students will be randomly assigned to four smaller sections (labeled Classes A-D) of equal size during the week before the class begins. Classes A and B will be the experimental groups in the study; Classes C and D will be the control groups. During the first two weeks of the semester, a pretest will be administered to students in Classes B and C, thus ensuring that one control group and one experimental group participates in the pretest.

Midway through the semester, all four classes will take part in a 2-week unit on assessment practices. During this period, the control group will receive traditional instruction on best practices in assessment-- including mention that students' problem-solving strategies should be evaluated based on how closely the solution method relates to the task. The experimental group will receive similar instruction, supplemented by activities where students complete high-school level problems and assess one another's work. Special emphasis will be placed on which methods the preservice teachers prefer when solving problems on their own and how this may affect the criteria they use when evaluating students' work. The goal of the intervention is to help the future educators in the experimental group internalize the relevant aspects of evaluating students' solutions by placing the strategies in the context of the cognition of themselves and their peers. Near the end of the semester, all classes will take a posttest on evaluating students work that is parallel to the pretest (with different student responses).

So far, very little information has been given about the pre- and posttests. Each test will consist of three *correct* responses for each of five separate problem-solving tasks. These fifteen

responses will be chosen from real high-school students' work to closely reflect common strategies for solving the problems. Participants will rank the three responses for each problem according to their preferred method (1=good, 2=better, 3=best). The researcher will consult fellow mathematics education specialists and experienced teachers to ensure the validity of "correct" rankings (based on the degree to which the students' responses related to the nature of the task). The participants' rankings of student work will thus be evaluated based on a total "fit" statistic. Pre- and posttest data will be analyzed using ANOVA to determine if there were any effects of the pretest and/or the experimental treatment on the posttest performance.

The parallel nature of the pre- and posttests is the primary rationale for using the Solomon Four Group experimental design. Such a design can control for possible "testing effects" of the pretest on the preservice teachers' evaluation of student work during the posttest. Without controlling for effects of the pretest, any significant results attributed to the treatment could be threatened by internal validity issues. By design, the proposed study also incorporates many of the strengths of classic experiments, including random sampling (strong generalizability and procedures of control), time order, and controlled manipulation of independent variables. Thus, statistical differences found during the study could lead to causal statements.

Several weaknesses exist in the proposed experimental design. First, it may be very difficult (or impossible) to find a university with 100 preservice secondary teachers taking a methods course during one semester. Second, since the four classes would likely be taught by different instructors with varying dropout rate, there may be history and experimental mortality threats to the internal validity of the study. Third, the validity of the measure to evaluate the participants' rankings of student responses can be called into question. Although there are general characteristics associated with the "best" strategies for evaluating student work, the researcher will need to carefully choose problem-solving tasks and students' work so that there is wide agreement among professionals about which students' solutions were "best". Likewise, special care must be taken to choose parallel problems and student work for the pre- and posttests. Any differences between what is being measured in the two tests could lead to significant threats to the face validity and construct validity of the measures.

Question 3: An education agency would like to assess the impact of an intervention on middle school mathematics achievement on the CSAP exam. The evaluation must employ random sampling methods. (a) Summarize three quantitative articles that inform the investigation; (b) design a quantitative study to answer the research question; (c) include relevant issues in the sampling procedure, the research variables, a model of the research design, the data gathering process, issues with reliability and validity, and appropriate methods for analyzing the data.

If the goal is to evaluate a new intervention designed to improve CSAP scores in Colorado middle schools, then a review of literature can help to identify the significant factors influencing the assessment of such a program aimed at improving achievement on standardized tests. Three research articles add particular value to the design of such an evaluative study (Abbott, Joireman, & Stroh, 2002; Tomoff, Thompson, & Behrens, 2000; Good, & Grouws, 1979). Following a description of how these articles inform the design of the study, I provide a description of a large-scale retrospective cross-sectional study that could be used to evaluate the effects of the new intervention on students' CSAP scores. I will work under the assumption that the intervention will be implemented statewide on a volunteer basis for teachers, and that researchers are interested in statistically controlling for relevant factors that influence achievement so that the intervention can be assessed as directly as possible.

Abbott et al. (2002) examined fourth and seventh grade academic performance on the Washington Assessment of Student Learning (WASL) using Hierarchical Linear Modeling (HLM). Their research is particularly informative for an evaluative study because it highlights some school-level and district-level predictors of mathematics achievement among middle school students. The analysis of WASL data agreed with prior research that identified significant negative effects on 7th grade achievement attributable to school poverty (SES) and an interaction between school size and school poverty: “As can be seen, the negative relationship between school poverty and achievement is stronger in large districts, thus replicating Bickel and Howley’s (2000) findings (for 8th grade only).” (Abbott et al., 2002, p. 7) Thus, the work of Abbot et al. supports a research design that incorporates both school size and school poverty as indicators of middle school students’ achievement on a standardized test.

Tomoff, Thompson, and Behrens (2000) also used HLM to explore predictors of student achievement in the middle grades (i.e., grades 7 and 8). In particular, Tomoff et al. analyzed TIMSS data under the hypothesis that as many as 40 self-reported instructional behaviors of teachers might have significant effects on student achievement. HLM procedures reduced the number of predictors to 13 variables related to classroom practices in four categories: (1) creating projects and reports, (2) working in large and small groups, (3) practicing algorithms, and (4) doing textbook and worksheet exercises. Surprisingly, the NCTM-recommended practices (project creation and group work) had no significant effect on scores in both the overall test and the problem-solving subtest. Drill and practice of algorithms had a negative effect on students’ scores, while completing textbook and worksheet exercises had a positive effect on students’ scores. The strongest predictor of achievement scores was, however, the educational attainment of the students’ parents, which explained 35% of the variation in scores between classrooms. While their results may have limited direct generalizability to CSAP because of the low-stakes nature of the TIMSS exam, Tomoff et al. succeeded in providing an explanatory model for middle school achievement that included self-reported teacher practices. Since the goal of my investigation is to evaluate an intervention for teachers, the results of Tomoff et al. (2000) suggest a research design that incorporates teacher-level instructional behaviors as predictors of student performance. Their work also supports the inclusion of “parental educational attainment” as a predictor in the statistical model.

The work of Good and Grouws (1979) offers additional support for the contention that teachers’ instructional behavior can influence student achievement. In a follow-up to a large study of classroom behaviors associated with consistently effective teachers, Good and Grouws recruited 40 fourth grade teachers to participate in an experimental study measuring the impact of a new instructional program on student achievement. Teachers in the treatment group received two 90-minute training sessions and a 45-page manual detailing effective instructional behaviors, and then were observed to measure their instructional behaviors. The key aspects of the program included daily review, content development, seatwork, homework assignments, and special reviews. Treatment teachers implemented the new program well and students in the treatment group outperformed the control group. In fact, ten of the twelve classes that improved the most during the study were in the treatment group. This study informs my investigation into the effects of a teacher-based intervention in at least three ways; it (1) serves as an example of a scientific evaluation of an instructional intervention, (2) provides a protocol for assessing teachers’ instructional behaviors in relation to effective practices, and (3) suggests a strong link between observable teacher practices and their students’ achievement. On the other hand, I am assuming

that a statewide implementation of the program will take place on a volunteer basis, so the experimental methodology of Good and Grouws (1979) could not be applied.

If the goal is to answer the question “Does the intervention improve CSAP scores?”, then the research must incorporate statistical procedures to control for the many other variables that might improve CSAP scores. The three research articles I have described support a research design that produces a model of student achievement (dependent variable) that includes student-level indicators (e.g., prior CSAP scores, SES, and educational attainment of parents), teacher-level indicators (e.g., participation or non-participation in the interventional program, self-reported instructional behaviors, and observed instructional behaviors) and school-level indicators (e.g., school size, school SES, average school CSAP performance). I believe that a large-scale cross-sectional study that utilizes HLM techniques can produce an appropriate research design for the research objective.

A critical aspect of the proposed cross-sectional study is the sampling procedure. Since the population is middle school teachers in Colorado, and a large-scale implementation of the intervention will give every teacher access to the program, it is important to produce a representative sample that allows for meaningful construction of the statistical model. Thus, disproportionate stratified random sampling procedures may be appropriate. By categorizing Colorado middle schools according to strata with varying SES and school size, researchers could randomly select 10 middle schools from among the strata. Then, four teachers will be randomly sampled from each school—two teachers that chose to participate in the study and two that did not. Since the teacher is the basic unit of analysis, all students taught by teachers participating in the study would be eligible for participation in the study. However, students who did not complete the CSAP in the prior year will be excluded from the study. See Figure 2 for a model of the research design.

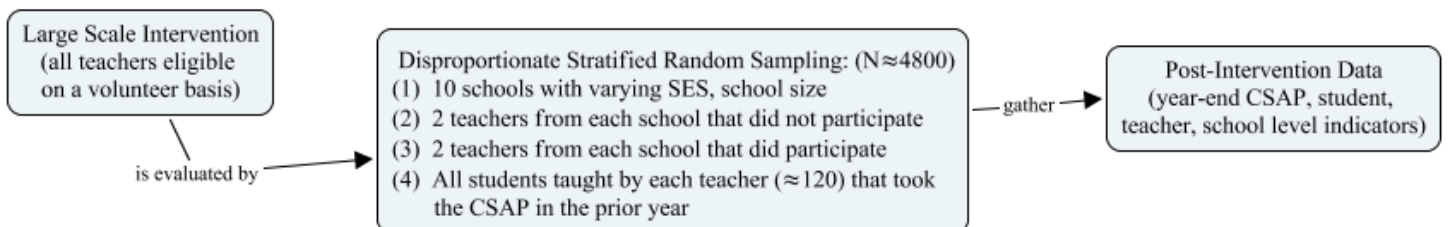


Figure 2. Research design of a large cross-sectional study to evaluate the impact of an intervention on students’ CSAP scores.

Following the sampling procedure, all teachers in the study will be asked to complete a questionnaire to provide self-reported information about their instructional practices. Then, the teachers that chose to participate in the study will receive the intervention, and trained specialists will observe all teachers toward the end of the semester to gather descriptive information related to their instructional practices. Students’ scores on the end of the year CSAP will provide the measure of the dependent variable, with data provided by the school will fill provide properties of the students who participated in the study.

There are several threats to the reliability and validity of the cross-sectional design. Incomplete data are likely for a large-scale study such as this that incorporates multi-level data from multiple sources. In addition, any unreliability or validity issues with the CSAP exam will necessarily exist in the analysis of the model. If, for example, the CSAP exam is found to produce a wide variation in scores during test-retest measures of reliability, then the hierarchical

linear model constructed to explain variation in the scores will not be able to explain this variation. While related to a post-test only control group design, the correlational study I've described does not incorporate experimental procedures such as random assignment to experimental and control groups, and so has limited power in its conclusions. It also is subject to significant threats to internal validity, including selection effects related to the volunteer aspect of teachers' enrollment in the intervention, history effects related to differing experiences among classes in the study, and experimental mortality caused by students transferring classes or not completing all measures in the study. However, the procedure for ensuring a representative sample through stratified probability sampling and random inclusion of participants in the evaluation suggests strong generalizability for the study's findings.

The correlational study I have described is founded upon and designed to be readily analyzed using HLM techniques. Besides producing descriptive statistics (means and standard deviations of performance), HLM software can analyze hypothesized linear relationships with regression methods at the student, teacher, and school levels (based on the variables described earlier). If the independent variable associated with inclusion in the intervention is found to have a significant effect on student achievement, the researchers will be able to answer the original question driving the research. Moreover, the model can provide meaningful data as to *how* the intervention might influence students' achievement through possible relationships between the intervention and teachers' instructional behaviors. In other words, the research design may be able to measure the extent to which the intervention changed teachers' classroom practices, which is associated with student achievement (Good, & Grouws, 1979; Tomoff, et al., 2000).

Question 4: Compare and contrast the articles by Oh, Rasmussen, and Allen (2005) and Ellington (2006). Reference at least two other articles and address how the two study's results support and contradict one another.

At first glance, the meta-analysis of graphing calculator studies performed by Ellington (2006) is starkly different from Oh, Rasmussen, and Allen's (2005) study of the delayed retention of knowledge in a differential equations course. However, an inspection of the variables used in both studies suggests a common thread: both Ellington (2006) and Oh, et al. (2005) were interested in the effects of interactive instruction techniques on students' procedural and conceptual knowledge. In the paragraphs that follow, I describe several conclusions that, when set in the context of additional research, are supported by the two articles. I will also describe conflicting results in the two articles while maintaining the perspective that the studies' findings should not be directly compared because of the limited generalizability of both investigations.

Before describing the conclusions of Ellington (2006) in relation to those of Oh, et al. (2005), it is helpful to contrast the research questions and methodologies employed in the two articles. Based on results from 42 studies meeting pre-set criteria for inclusion in the meta-analysis, Ellington analyzed 97 effect sizes of graphing calculator use on students' procedural, conceptual, and overall mathematics achievement. Half of the studies included college students, with most of the remaining studies including high school students. The research participants were mostly of mixed ability and the studies generally took place in algebra or Precalculus courses. All of the studies included traditional control groups and experimental groups that had access to graphing calculators *without* computer algebra systems. Some of the studies allowed students who were in the graphing calculator group to use the calculators on tests, and some did not.

In contrast to the Ellington's broad analysis of available studies on graphing calculator effects, the study conducted by Oh, et al. (2005) investigated the retention of differential equations knowledge among two small classes at an elite school in Korea. The inquiry-oriented (IO-DE) class (N=15) used discovery based materials and group work to "reinvent" differential equation procedures with the help of technology. Students in the IO-DE class had access to TI-89 calculators, which have computer algebra capabilities. The traditional (TRAD-DE) course (N=20) served as the control group for the comparative study, and the dependent variables were retention of procedural, conceptual, and overall knowledge 1 year after the students completed their respective courses. The focus of both Ellington and Oh, et al. on procedural, conceptual, and overall knowledge represents a point of commonality between their respective research. However, the choice of graphing calculator (TI-89) as well as the advanced content (differential equations) and the elite population of students in the Oh, et al. (2005) study combine to make it unreasonable to suggest any direct comparisons of Ellington's (2006) and Oh, et al.'s findings.

The major conclusion of the Oh, et al. study is that the students in the interactive and traditional courses both performed poorly on the delayed posttest of knowledge retention. There were no significant differences between the IO-DE and TRAD-DE students on the measures of procedural or overall retention. An ANCOVA procedure to adjust effects of prior posttest performance on the delayed posttest found a single significant difference: the IO-DE class retained more knowledge of qualitative/graphical approaches to differential equation problems than the TRAD-DE class. Ellington (2006) produced seemingly conflicting results in her study of graphing calculator effects. Students who used graphing calculators significantly outperformed their traditionally taught peers in measures of both procedural and conceptual understanding. Despite the differences in the studies, Ellington's results appear to contradict those of Oh, et al. (2005). However, when the differential equations study is categorized according to Ellington's criteria, the results of Ellington and Oh, et al. are very similar.

Ellington found when the studies included in the analysis were restricted to those that did not allow calculators during the testing of the students' understanding (like in the differential equations study), there were little significant differences in their procedural and overall understanding. In fact, for those studies not allowing calculators during testing, Ellington found that there was only a small significant difference between the control and graphing calculator groups in their performance on conceptual understanding. This coincides very well with Oh, et al.'s findings that the IO-DE class retained slightly more qualitative/graphical conceptual understanding than the TRAD-DE class. Perhaps it is no coincidence that retention of *graphical* knowledge was found to be significant in the differential equations study and that Ellington's study found students enjoy using graphing calculators. It might be true that the visual representations associated with both graphing calculators and the IO-DE instruction methods are enjoyable to students and make lasting impressions on their conceptual understanding of mathematics.

The study of Japanese and U.S. students' understanding of calculus conducted by Judson and Nishimori (2005) is similar to Oh, et al. if you consider (as the researchers did) the U.S. Portland class to be "interactive" and the Japanese Sapporo classes to be "traditional". In Judson and Nishimori's study, the U.S. and Japanese students performed similarly on conceptual tests, but Japanese students outperformed the Portland students on the traditional procedural algebra skills. Thus, the results of Judson and Nishimori provide additional insight into the results of the differential equations study by establishing a relationship between the retention results of Oh, et al. (2005) to the performance results of the Japanese calculus study. In other words, traditional

students may outperform interactive students on procedural tests (that do not allow graphing calculators during testing), but long-term retention of understanding may be similar for both groups.

An interesting study that produced very similar results to Ellington's study of graphing calculators (2006) is Hake's (1998) survey of mechanics understanding among students who participated in interactive engagement versus traditional instruction. Hake compared pre- and posttest data of interactive and traditional classes and found that interactive classes achieved significantly greater normalized performance gains on both procedural and conceptual measures of understanding in introductory physics. Hake's results can be seen as complimentary to Ellington's similar findings related to the effects of graphing calculator on conceptual and procedural understanding. In fact, the results of Ellington (2006), Hake (1998), Oh, et al. (2005), and Judson and Nishimori (2005) all support the conclusion that under proper testing conditions interactive instruction (e.g., use of graphing calculators) can produce gains in conceptual understanding above and beyond those found in traditional instruction methods.

Question 5: Use the table of student enrollment in Australian higher education during 1981-1988, cross classified by sex and age group, to answer the following questions: (1) Is there a statistical difference in the number of students based on year?, and (2) How do age and gender impact the number of students in Australian higher education?

I conducted a statistical analysis of student enrollment in Australian higher education by considering three explanatory variables: age group, sex, and year. In particular, two research questions were addressed: (1) Is there a significant difference in the number of students based on year?, and (2) How do age and gender affect the number of students in Australian education during the time period under investigation? To answer the first question, I calculated descriptive statistics, and performed a one-way analysis of variance (ANOVA) with year serving as the fixed effect and student enrollment serving as the dependent factor. The second question was addressed by completing a multiple effects ANOVA for student enrollment with hypothesized fixed effects that included age, gender, and age*gender interaction.

Descriptive statistics for the number of students in higher education by year are shown in Table 1. The overall mean was $M = 46084$ ($N = 64$) with a relatively large standard deviation ($SD = 13244.7$). The enrollment trends suggest (a) a slight increase in mean enrollment over time and (b) a corresponding increase in variation and range of values. In other words, as the box plot diagram in Figure 3 indicates, the central tendency and spread of the enrollment data both increased with time. A fixed effects ANOVA found no significant difference in the number of students by year ($F(7,56) = .624$, $p = .734$), reflecting a small amount of between-year variation. Table 1.

Descriptive statistics for student enrollment in Australian higher education by year (1981-1988).

Year	N	Mean	Std. Deviation	Minimum	Maximum
1981	8	41636.50	11681.252	20159	59698
1982	8	42544.50	11859.344	20418	59629
1983	8	43404.00	12387.545	20530	61120
1984	8	44554.63	12650.305	20828	61252
1985	8	46156.13	13072.828	21999	61337
1986	8	48657.38	13868.121	23443	61914
1987	8	49156.38	15067.596	23088	65741
1988	8	52565.63	16846.221	24768	72985

Total	64	46084.39	13244.744	20159	72985
-------	----	----------	-----------	-------	-------

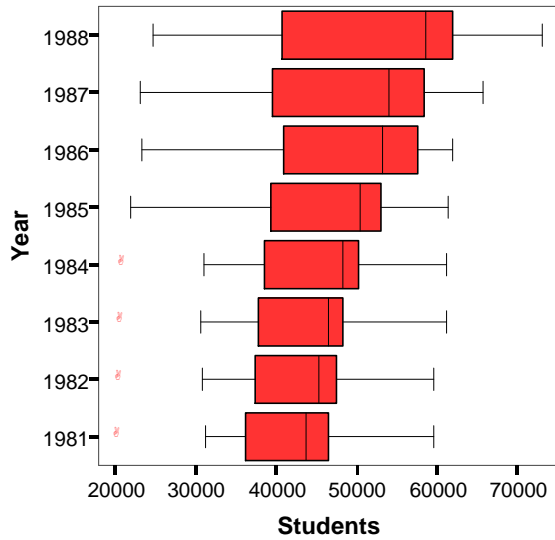


Figure 3. Box plots of student enrollment in Australian higher education by year.

In order to address the impact of age and gender on the number of students in Australian higher education, I began with a multiple effects ANOVA for student enrollment as the dependent variable, and age, gender, and age*gender interaction as the fixed effects in the model. See Table 2 for a summary of the resulting tests of the hypothesized between-subjects effects. The multiple effects ANOVA identified highly significant effects of age, gender, and age*gender interaction on the number of students in Australian higher education.

Table 2.

*Multiple Effects ANOVA for testing between-subjects effects of age, gender, and age*gender interaction on student enrollment in Australian higher education (1981-1988).*

Source	Type III Sum of Squares	df	Mean Square	F
Corrected Model	9658389297	7	1379769900	55.457***
Age	8619496684	3	2873165561	115.481***
Gender	251416700	1	251416700	10.105**
Age * Gender	787475913	3	262491971	10.550***
Error	1393274414	56	24879900	
Corrected Total	11051663711	63		

** p<0.01, ***p<0.001

Figure 4 shows box plots of the number of students in Australian higher education arranged by age and gender. The box plots display a large difference in the number of students between the ages 25 and 29 compared to the other age brackets. This can be considered to be the primary influence on the highly significant effects for age on the number of students (F(3,56)=115.5, p<0.001).

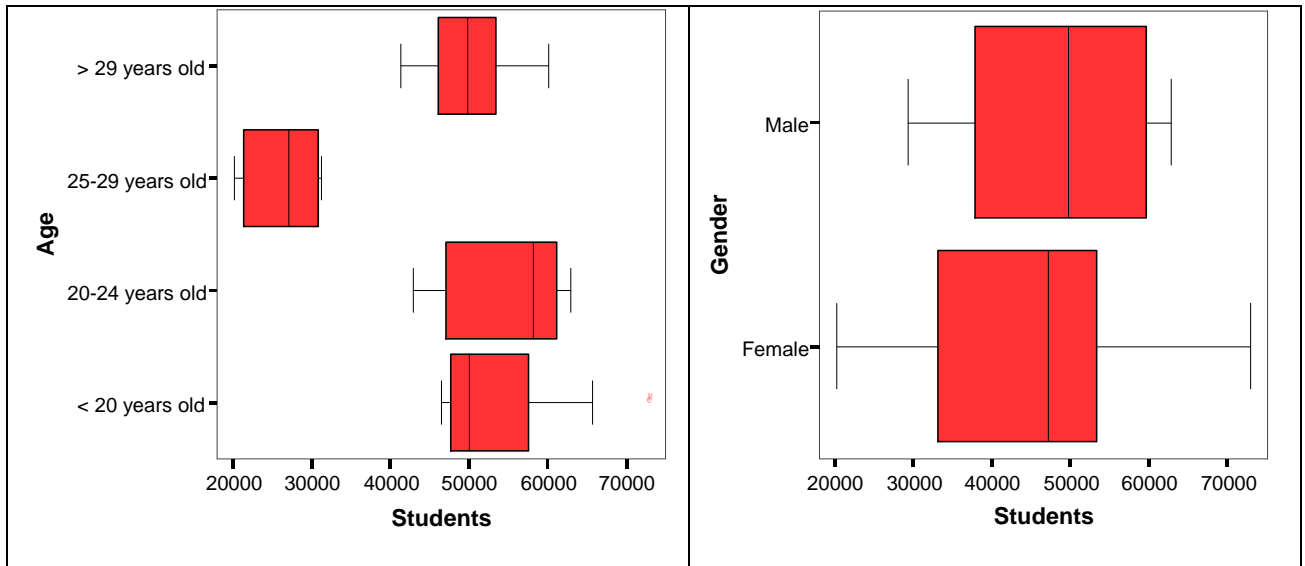


Figure 4: Box plots showing the number of students in Australian higher education according to age and gender.

The age*gender interaction effect can be found in the marginal means plot shown in Figure 5. If no interaction effects were present the marginal means plots of age and gender would be parallel—instead the female means exceed the male means for the youngest and oldest age groups, while the male means are larger for the two middle age groups.

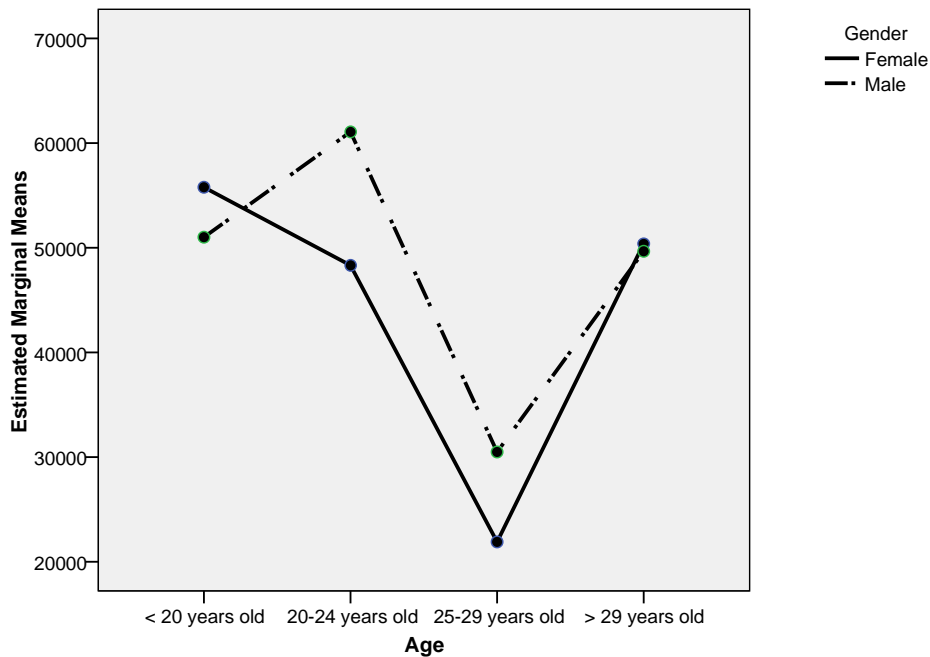


Figure 5. Estimated marginal means of students by age and gender.

Since the multiple effects ANOVA for number of students in Australian higher education produced significant effects related to age, gender, and age* gender interaction, the statistical

package can compute a linear regression model for the data. The resulting model explained 86% of the variation in the number of students:

$$Students = 54759.2 + 1958Gender - 4262.7Age + 802.3(Age * Gender)$$

Thus, there was a strong negative effect of age on the number of students (older age groups contribute significantly less to the total number than younger groups) and a positive effect of gender on the number of students. The phrase “positive effect of gender” is a little strange until you realize that Gender was coded in the statistical package as 0=Female and 1=Male. Hence, Males contribute more to the number of students than females. These regression results confirm the patterns evident in the box plots in Figure 4.

A residual plot of the linear regression resulted in a slight megaphone effect of the error terms associated with the model, indicating the need to transform the data. A logarithmic transformation provided a slightly better fit, but the best transformation I found was to transform the dependent variable “students” to a new variable “ratstudents”: $\frac{46084}{students}$. The resulting linear model explained over 95% of the variation in population and produced rather acceptable residual plot shown in Figure 6. However, the effects associated with this new model corresponded with those found in the original un-transformed equation, so there was little reason to switch models.

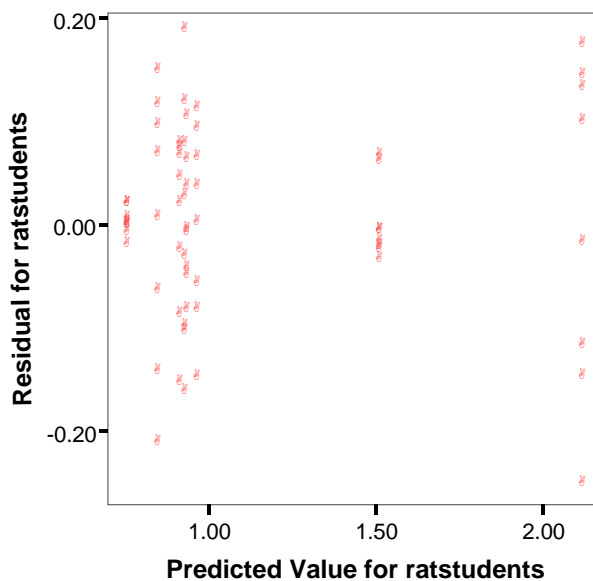


Figure 6. Residual plot for a linear model of the ratio of number of students to the grand mean $\left(\frac{46084}{students}\right)$, using age, gender, and age*gender as predictors.

References

Abbott, M. L., Joireman, J., & Stroh, H. R. (2002). *The influence of district size, school size and socioeconomic status on student achievement in Washington: A replication study using heirarchical linear modeling*. Lynwood, WA: Washington School Research Center.

Frankfort-Nachmias, C., & Nachmias, D. (2000). *Research methods in the social sciences*. (6th ed.). New York: Worth Publishers.

- Good, T. L., & Grouws, D. A. (1979). The Missouri Mathematics Effectiveness Project: An experimental study in fourth-grade classrooms. *Journal of Educational Psychology*, 71(3), 355-362.
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64-74.
- Harwell, M. R., Post, T. R., Maeda, Y., Davis, J. D., Cutler, A. L., Andersen, E., et al. (2007). Standards-based mathematics curricula and secondary students' performance on standardized achievement tests. *Journal for Research in Mathematics Education*, 38(1), 71-101.
- Judson, T. W., & Nishimori, T. (2005). Concepts and skills in high school calculus: An examination of a special case in Japan and the United States. *Journal for Research in Mathematics Education*, 36(1), 24-43.
- Tomoff, J., Thompson, M., & Behrens, J. (2000, April). *Measuring NCTM-recommended practices and student achievement with TIMSS*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. 443887)

Appendix A: SPSS Spreadsheet

Definition of Variables

prob5.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
2	Age	Numeric	8	0		{1, < 20 years	None	8	Right	Ordinal
3	Gender	Numeric	8	0		{0, Female}...	None	8	Right	Nominal
4	agegender	Numeric	8	2		None	None	11	Right	Scale
5	Students	Numeric	8	0		None	None	8	Right	Scale
6	ratstudents	Numeric	8	2		None	None	13	Right	Scale
7	pred_stud	Numeric	8	2	Predicted Valu	None	None	10	Right	Nominal
8	res_stud	Numeric	8	2	Residual for St	None	None	10	Right	Scale
9	pred_ratlstu	Numeric	8	2	Predicted Valu	None	None	10	Right	Nominal
10	res_ratlstud	Numeric	8	2	Residual for rat	None	None	10	Right	Scale

Data Table

Year	Age	Gender	agegender	Students	ratstudents	pred_stud	res_stud	pred_ratlstud	res_ratlstud
1981	1	1	1	46687	0.99	51016.75	-4329.75	0.91	0.08
1981	1	0	0	46460	0.99	55790.5	-9330.5	0.85	0.15
1981	2	1	2	59698	0.77	61075.63	-1377.63	0.75	0.02
1981	2	0	0	43007	1.07	48326.13	-5319.13	0.96	0.11
1981	3	1	3	31227	1.48	30499.13	727.88	1.51	-0.04
1981	3	0	0	20159	2.29	21904.13	-1745.13	2.11	0.17
1981	4	1	4	44558	1.03	49674.13	-5116.13	0.93	0.1
1981	4	0	0	41296	1.12	50388.75	-9092.75	0.93	0.19
1982	1	1	1	46977	0.98	51016.75	-4039.75	0.91	0.07
1982	1	0	0	48064	0.96	55790.5	-7726.5	0.85	0.11
1982	2	1	2	59629	0.77	61075.63	-1446.63	0.75	0.02
1982	2	0	0	43801	1.05	48326.13	-4525.13	0.96	0.09
1982	3	1	3	30972	1.49	30499.13	472.88	1.51	-0.02
1982	3	0	0	20418	2.26	21904.13	-1486.13	2.11	0.14
1982	4	1	4	46435	0.99	49674.13	-3239.13	0.93	0.06
1982	4	0	0	44060	1.05	50388.75	-6328.75	0.93	0.12
1983	1	1	1	47220	0.98	51016.75	-3796.75	0.91	0.07
1983	1	0	0	49062	0.94	55790.5	-6728.5	0.85	0.09
1983	2	1	2	61120	0.75	61075.63	44.38	0.75	0
1983	2	0	0	45009	1.02	48326.13	-3317.13	0.96	0.06
1983	3	1	3	30644	1.5	30499.13	144.88	1.51	-0.01
1983	3	0	0	20530	2.24	21904.13	-1374.13	2.11	0.13
1983	4	1	4	47753	0.97	49674.13	-1921.13	0.93	0.03
1983	4	0	0	45894	1	50388.75	-4494.75	0.93	0.08
1984	1	1	1	48301	0.95	51016.75	-2715.75	0.91	0.04
1984	1	0	0	50591	0.91	55790.5	-5199.5	0.85	0.07
1984	2	1	2	61252	0.75	61075.63	176.38	0.75	0
1984	2	0	0	46256	1	48326.13	-2070.13	0.96	0.03
1984	3	1	3	31034	1.48	30499.13	534.88	1.51	-0.03
1984	3	0	0	20828	2.21	21904.13	-1076.13	2.11	0.1
1984	4	1	4	49858	0.92	49674.13	183.88	0.93	-0.01
1984	4	0	0	48317	0.95	50388.75	-2071.75	0.93	0.03
1985	1	1	1	49617	0.93	51016.75	-1399.75	0.91	0.02
1985	1	0	0	54223	0.85	55790.5	-1567.5	0.85	0
1985	2	1	2	61337	0.75	61075.63	261.38	0.75	0
1985	2	0	0	47956	0.96	48326.13	-370.13	0.96	0
1985	3	1	3	30669	1.5	30499.13	169.88	1.51	-0.01
1985	3	0	0	21999	2.09	21904.13	94.88	2.11	-0.02
1985	4	1	4	51970	0.89	49674.13	2295.88	0.93	-0.04
1985	4	0	0	51478	0.9	50388.75	1089.25	0.93	-0.03
1986	1	1	1	52165	0.88	51016.75	1148.25	0.91	-0.03
1986	1	0	0	59198	0.78	55790.5	3407.5	0.85	-0.07
1986	2	1	2	61914	0.74	61075.63	838.38	0.75	-0.01
1986	2	0	0	51123	0.9	48326.13	2796.88	0.96	-0.06

1986	3	1	3	30869	1.49	30499.13	369.88	1.51	-0.02
1986	3	0	0	23443	1.97	21904.13	1538.88	2.11	-0.15
1986	4	1	4	54464	0.85	49674.13	4789.88	0.93	-0.09
1986	4	0	0	56083	0.82	50388.75	5694.25	0.93	-0.11
1987	1	1	1	56099	0.82	51016.75	5082.25	0.91	-0.09
1987	1	0	0	65741	0.7	55790.5	9950.5	0.85	-0.14
1987	2	1	2	60759	0.76	61075.63	-316.63	0.75	0
1987	2	0	0	52558	0.88	48326.13	4231.88	0.96	-0.08
1987	3	1	3	29251	1.58	30499.13	-1248.13	1.51	0.06
1987	3	0	0	23088	2	21904.13	1183.88	2.11	-0.12
1987	4	1	4	50022	0.92	49674.13	347.88	0.93	-0.01
1987	4	0	0	55733	0.83	50388.75	5344.25	0.93	-0.1
1988	1	1	1	61068	0.75	51016.75	10051.25	0.91	-0.16
1988	1	0	0	72985	0.63	55790.5	17194.5	0.85	-0.21
1988	2	1	2	62896	0.73	61075.63	1820.38	0.75	-0.02
1988	2	0	0	56899	0.81	48326.13	8572.88	0.96	-0.15
1988	3	1	3	29327	1.57	30499.13	-1172.13	1.51	0.06
1988	3	0	0	24768	1.86	21904.13	2863.88	2.11	-0.25
1988	4	1	4	52333	0.88	49674.13	2658.88	0.93	-0.05
1988	4	0	0	60249	0.76	50388.75	9860.25	0.93	-0.16