UNIVERSITY OF WASHINGTON

Practical Statistics for Human-Computer Interaction

Independent Study using SAS JMP and IBM SPSS.

Jacob O. Wobbrock, Ph.D.

The Information School | DUB Group

University of Washington

Seattle, WA 98195 USA

Acknowledgement

This work was supported in part by the National Science Foundation under CAREER grant IIS-0952786. Any opinions, findings, conclusions or recommendations expressed in this work are those of the author and do not necessarily reflect those of the National Science Foundation.

Table of Contents

Introduction

I have often found that students in my field of human-computer interaction (HCI), particularly from the computer science tradition, lack, as I did, the proper statistical skills to evaluate results of studies, whether experimental, quasi-experimental, observational, or response-based. These fundamentals can be hard to come by. On the one hand, formal coursework in statistics departments is, especially at the graduate level, overly theoretical for easy application to real-world HCI studies. On the other hand, students who simply read help documentation for a particular statistics package are likely to make numerous errors due to a lack of statistics fundamentals. What has been missing is the ground between these two extremes, a place for students to learn enough principled material to avoid blunders, while also learning the practical handling of real-world data sets in a statistics package.

To provide this middle ground is my purpose in creating this independent study. It is designed to be completed in a 10-week period, which fits nicely into the quarter system at many universities. While the modules are meant to be doable without aid of an instructor, periodic feedback from an instructor is presumed. I ask my own students to email me their modules each week, and after annotating them, I discuss any problems with them the following week. Students can also come to me when they are having difficulties along the way, although I encourage them first to reach the boiling point of frustration as they *independently* try to resolve their confusion. That is the nature of an independent study, after all.

Many, but not all, of the sample data sets that accompany this study are situated in the context of human-computer interaction. Others are from experiments in other fields. Most of the data sets are fictitious, designed by me to exhibit certain distributional properties and so forth.

The pedagogical goal for this study is that by its end, a student will be proficient in many common statistical analyses, and know how to recognize the proper analysis from simply viewing a data table, even without knowing how the experiment was run. This level of proficiency is sufficient for both comprehending statistical results in HCI papers, and for producing them in the first place.

A trial version of the SAS JMP statistics package is available from http://www.jmp.com/. A trial version of the IBM SPSS statistics package is available from http://www.spss.com/ . I readily admit the downside of having two proprietary statistics packages as the focus of this independent study, but these are the tools with which I am familiar. If someone wants to reproduce this study and its data sets for use with **R**, I would be most grateful.

I welcome comments, suggestions, and errata. You can reach me by email at wobbrock@uw.edu.

Jacob O. Wobbrock, Ph.D.
June 2010
Seattle, WA

01. Concepts and definitions

For the following statistics terms, discover the formal definition (*e.g.*, on-line or in a textbook), and then write a "common sense" definition reflecting the *intuition* of the concept. With each common sense definition, supply an example that shows you really know what the term means. You may have to read a considerable amount before gaining enough awareness to define each of these terms correctly. You might consider first writing the description for a single experiment, and then referring to various aspects of that experiment in your answers.

Answer. We'll refer to the following in our definitions: A human-computer interaction researcher wants to know about the effects of body posture on mobile text entry. She implements three popular text entry methods in a custom iPhone test bed: a virtual QWERTY keyboard, Palm OS Graffiti, and a phone keypad simulation running Tegic's T9. She recruits 20 subjects, having each one train for 30 minutes with one method chosen at random. Then she has each subject enter 20 test phrases in each of three postures—standing, walking, and jogging—the order of which was randomly determined. The outcomes of interest are words per minute and error rate.

1. Factor

   A factor is an experimental variable systematically changed to examine its effects, if any, upon an outcome of interest. In our example there are two factors, body *Posture* and text entry *Method*.

2. Level

   A level is a particular value that a factor can assume. In our example, the levels of *Posture* are standing, walking, and jogging. The levels of *Method* are QWERTY, Graffiti, and T9. Every factor must have at least two levels.

3. Independent variable

   An independent variable is the same thing as a factor.

4. Dependent variable

   A dependent variable is an experimental measure, response, or outcome of interest. In our example there are two dependent variables, words per minute and error rate.

5. Measure

   A measure is the same thing as a dependent variable.

6. Trial

   A trial is the experimental unit of activity over which one measure is taken. In our example, the entry of one text entry phrase would constitute one trial.

7. Covariate

   A covariate is a measurable feature of an experiment that, like a factor, may affect the dependent variable. Unlike a factor, however, a covariate is not manipulated; its levels take on their "natural" preset values and often cannot be changed. In our example, the gender and age of each subject would be covariates. Other covariates include how fast each subject walks or jogs, or the outside temperature present while each subject entered his phrases.

8. Within-subjects factor

   A within-subjects factor is one for which all levels are experienced by each subject. Since each subject entered phrases in all body postures, *Posture* is a within-subjects factor.

9. Between-subjects factor

   A between-subjects factor is one for which only one level is experienced by each subject. Since each subject used only one of three possible text entry methods, *Method* is a between-subjects factor.

10. Factorial design

    A factorial design refers to an experiment that has more than one factor, *i.e.*, more than one independent variable. Our example is a factorial design because it has two factors, *Posture* and *Method*.

11. Main effect

    A main effect refers to a finding of statistical significance for a factor in an experiment. If *Posture* exerts a significant effect on words per minute, we would say we have a main effect of *Posture* on text entry speed.

12. Interaction

    An interaction refers to the interplay of two or more factors such that the effect of a level of a factor depends upon the level of another. In our example, a significant *Posture*Method* interaction on text entry speed would mean that a given posture's speed relative to other postures' speeds depends on the text entry method being used; or, equivalently, that the speed of a text entry method relative to other text entry methods depends on the posture used.

13. Within-subjects design

    A within-subjects design is an experiment in which all factors are within-subjects factors. If all of our subjects had used all three text entry methods, the experiment could be said to have used a within-subjects design.

14. Between-subjects design

    A between-subjects design is an experiment in which all factors are between-subjects factors. If each subject had only experienced one posture along with only one text entry method, the experiment could be said to have used a between-subjects design.

15. Mixed factorial design

    A mixed factorial design is an experiment in which there are within-subjects factors and between-subjects factors. Our example experiment uses a mixed design.

16. Confound

A confound is a possible source of unaccounted for variation in the dependent variable. In other words, it is a feature in an experiment that was not manipulated as a factor, recorded as a covariate, or controlled to be the same for all subjects. In our example, a confound would be the prior text entry experience of the subjects. In a real experiment, this should be controlled for with a screening questionnaire (to recruit subjects only with the same prior experience, perhaps none), manipulated as a between-subjects factor, or if nothing else, recorded for each subject as a covariate.

17. Control

Verb: *to control* for something experimentally means to make it the same for all subjects. In our example, if we had recruited only subjects who were 24 years old, we would have controlled for age. Noun: *a control* is a subject in a control group, which is the group taking on the default, *de facto*, or untreated level of a between-subjects factor. Our example study did not have explicit controls, but most medical studies do. For example, one group is given a drug, a second group is given a placebo, and a third group is given nothing; this third group is full of controls.

18. Carryover effect

Carryover effects threaten to confound any within-subjects factor as effects from one level of the factor change the results for a subsequent level of the factor. In our example, a carryover effect may exist for *Posture*, if, say, subjects who jog first are then tired when entering phrases in the standing or walking postures. In general, common carryover effects involve fatigue, learning, and motivational changes. Carryover effects do not apply to between-subjects factors.

19. Counterbalancing

Counterbalancing is the process by which the levels of a within-subjects factor are administered among the subjects to avoid carryover effects from systematically confounding the results. In our example, the levels of *Posture* were counterbalanced by having their order chosen randomly for each subject. Randomization is one approach to counterbalancing; other approaches involve deliberately issuing all possible orders in an experiment (called "fully counterbalanced") or using a systematic partial ordering, *e.g.*, a Latin Square. It is good practice to test for order effects to ensure that counterbalancing worked. Testing for order effects is covered in module 8.

20. Nominal variable

A nominal variable is a factor or measure that takes on one of an unordered assortment of values. In our example, both *Posture* and *Method* are nominal variables.

21. Categorical variable

A categorical variable is the same thing as nominal variable.

22. Ordinal variable

An ordinal variable is a factor or measure that takes on one of an *ordered* assortment of values. Although ordered, no assumption is made that the ordering is linear, *i.e.*, that the gaps between successive values are known or regular. In our example, if the experimenter had subjects fill out 7-point Likert-type scales with subjective ratings of their opinions, these data would be codified with ordinal variables.

23. Continuous variable

A continuous variable is a factor or measure that takes on a number whose relationship to and distance from other numbers is known. (Numbers are often used in data tables for nominal and ordinal variables, too, but these numbers are merely symbols, and could be replaced by text strings if desired. Not so for continuous variables, where the number *is* the value for the variable, not merely a symbol for the value.) In our example, both words per minute and text entry error rate are continuous dependent variables.

24. Scalar variable

A scalar variable is the same thing as a continuous variable.

25. Fixed effect

A fixed effect is a factor in a statistical model whose levels were specifically chosen and about whose values we specifically care. Traditional ANOVA models can only contain fixed effects; as such, they are called fixed-effects models. In our example, both *Posture* and *Method* are fixed effects.

26. Random effect

A random effect is a factor in a statistical model whose levels were randomly selected from a larger population over which we wish to generalize, and about whose values we do not specifically care. Traditional ANOVA models cannot accommodate random effects. A common random effect is the factor representing the experimental unit, *e.g.*, *Subject*, because it is often the case that we randomly select subjects for an experiment from a larger population, and that we do not care about the specific subjects recruited (*i.e.*, the levels of the *Subject* factor). In our example, *Subject* is a random effect with 20 levels. A random-effects model only has random effects, but in practice, random-effects models are uncommon. More common is a mixed-effects model in which we have both fixed and random effects, and we seek to draw conclusions only about the fixed effects.

27. Mixed-effects model

A mixed-effects model is a statistical model containing both fixed and random effects. Mixed-effects model analyses of variance (usually not written as “ANOVA” to avoid confusion with traditional fixed-effects models) are powerful regression-based analyses that offer many advantages, both practical and theoretical, over pure fixed-effects models.

02. Understanding a data table

dinner.jmp

Open the JMP data table *dinner.jmp*. Examine the table and then answer the following questions.

1. Does it represent a within-subjects, between-subjects, or mixed design? How do you know?

   The data represent a between-subjects design. We know this because each subject is listed only once in the table, having received only one level of *Dinner* and one level of *Game*.

2. What columns are the independent variables? What columns are the dependent variables?

   The independent variables are the *Dinner* and *Game* columns. The dependent variable column is *Time*. (Technically, *Subject* is also an independent variable column with 30 levels, one for each subject, but this factor and its levels are not of specific interest and are thus ignored. We will see in future modules that *Dinner* and *Game* are called "fixed effects," while *Subject* is considered a "random effect" because of this distinction.)

3. What are the factors? What are the levels?

   One factor is *Dinner* with two levels, "steak" and "fish." The other factor is *Game* with two levels, "shooter" and "side-scroll."

4. Which factors are nominal? Ordinal? Continuous?

   Dinner is nominal. *Game* is also nominal. (*Time* is not a factor but it is a continuous dependent variable, or measure. *Subject* is a nominal factor with 30 levels, but as noted above, it is not a fixed effect like *Dinner* and *Game*, but a random effect.)

5. How would you represent this with the multiplicative pattern of describing experiments (*e.g.*, "a 3×2×3 within-subjects design")?

   This experiment was a 2×2 between-subjects design. This notation indicates that there are two factors because we have two multiplicands, and that each factor has two levels because both multiplicands are the number "2".

6. Given the variable modeling types, what type of analysis should be performed: ANOVA, Chi-Square, nominal logistic regression, ordinal logistic regression, or numeric regression? For each of these, say what the independent variable types and the dependent variable type must be.

   As we noted in question 4, the variable modeling types are two nominal factors and one continuous measure. (This can be thought of as an equation with two nominal *X*'s and one continuous *Y*, *i.e.*, $f(X_1, X_2) = Y$.) The appropriate analysis for nominal (or ordinal) inputs and a continuous output is an ANOVA, which stands for ANalysis Of VAriance. For a Chi-Square test, we require nominal (or ordinal) inputs and a nominal (or ordinal) output. For nominal logistic regression, we require continuous inputs and a nominal output. For ordinal logistic regression, we require continuous inputs and an ordinal output. For numeric regression, we require continuous inputs and a continuous output. In JMP, the Analyze > Fit Y by X menu item leads to a dialog box that contains a small chart of these analysis types. Please examine that chart in light of this answer.

7. What is a "balanced design?" Is this a balanced design? Why or why not?

A "balanced design" means that each level of each factor had assigned to it the same number of subjects. The *Game* factor is balanced because each of the two videogame types had 15 subjects assigned to it. However, the *Dinner* factor is not balanced because 16 subjects ate steak while only 14 subjects ate fish. Therefore, the experiment does not exhibit a balanced design.

8. Explain the structural differences between long-format and a wide-format data tables. Is *dinner.jmp* a long-format table or a wide-format table? How do you know?

A long-format table is one in which each table row represents only one trial. If subjects performed multiple trials, then a subject must be replicated on multiple rows. A wide-format table is one in which each table row contains all measures for a given subject. If a subject performed multiple trials, then each trial must be represented by a different column, and any given subject must only appear on one row. Because *dinner.jmp* has only one trial per row, it could be considered a long-format table. On the other hand, *dinner.jmp* also has all the measures for a given subject on one row, so it could be considered a wide-format table. This long/wide duality exists in *dinner.jmp* only because each subject was measured only once. In general, a purely between-subjects study will cause there to be no difference between a long-format and wide-format table. The distinction becomes vital, however, for studies that have any within-subjects factors, also called "repeated factors."

9. Give a plausible story about why and how this experiment was conducted.

A researcher wanted to know whether the dinner a computer gamer ate the night before would cause a significant difference in their videogame playing performance the next day. Further, the researcher wanted to know if such differences occurred for first-person shooter games and side-scrolling games. He recruited thirty computer gamers to take part, asking half to eat an 8 oz. steak dinner and half to eat an 8 oz. fish dinner the night before. (One fish-assigned subject mistakenly ate steak, resulting in 16 steak-eaters and 14 fish-eaters.) The next morning, 15 subjects played a first-person shooter videogame and 15 other subjects played a side-scrolling videogame. The time taken to complete a given level in each game was recorded as the sole continuous measure of performance.

10. Try doing an analysis and reporting your findings.

In JMP, go to Analyze > Fit Model and place *Time* in Y. In the model effects box, place *Dinner*, *Game*, and *Dinner*Game*. Click Run Model. (Because *dinner.jmp* uses a between-subjects design, we can ignore the *Subject* column. However, if the same subject was listed on multiple rows, we would need to place *Subject* in the model effects box and designate it as a random effect. This would result in a mixed-effects model, a topic covered in modules 3-5. Alternately, if we had subjects on multiple rows, we could transform our data table into wide-format and perform a traditional repeated measures ANOVA, which does not use random effects but treats all measures on the same row as coming from the same subject. As it is a purely between-subjects study, we can ignore both of these issues in *dinner.jmp*.)

The Effects Tests report contains our results, along with the Error line in the Analysis of Variance report. *Dinner* has a significant effect on *Time* (F(1,26) = 12.15, *p* < .01), with steak resulting in faster average times than fish (7.44 *vs*. 8.33 minutes). *Game* did not significantly affect *Time* (F(1,26) = 0.30, *n.s.*) and there was no *Dinner*Game* interaction (F(1,26) = 0.24, *n.s.*).

(You are not necessarily expected to provide a proper formal analysis at this stage. You will be asked to do so in the next module. But please try to understand where the above write-up gets its information in the statistical report.)

03. Introduction to ANOVA

dinner.jmp

For this module, we're still going to use *dinner.jmp*. You may have to look up certain things in the JMP help documentation. You may also have to refer to module 2 to remember how you interpreted the experiment design.

After opening *dinner.jmp*, please do the following:

Part 1, Descriptive statistics. Go to Analyze > Distribution. In Y, Columns put the dependent variable. In By, put only one independent variable. (Do this again in a separate analysis for the other independent variable.)

1. Eyeballing the descriptive statistics and histograms, what seems to be the case about the effect of each independent variable on the dependent variable?

   For both distributions, we turn on Uniform Scaling and Stack to facilitate visual comparisons. On the distribution of *Time* by *Dinner*, we can see that fish resulted in longer times than steak, which has a wider range and much lower minimum. It therefore appears that *Dinner* should exhibit a significant effect on *Time*. On the distribution of *Time* by *Game*, we see that for both videogame types, the spread of values looks quite similar. Therefore, it does not appear that *Game* should exhibit a significant effect on *Time*.

2. What might you expect the ANOVA to show? How do you know? Write your answer before going further.

   Based on visual comparisons of the distributions, we should expect an ANOVA to show a significant effect of *Dinner* on *Time*, a nonsignificant effect of *Game* on *Time*, and an unknown interaction result, as we were only asked to analyze each factor separately.

Part 2, ANOVA. Go to Analyze > Fit Model. This is where you provide a model that, when fitted to your data, will attempt to predict the outcome variable from the independent variables and their interactions. As before, put your outcome variable in Y. Put your two independent variables and their interaction (also called their "cross") in the Construct Model Effects box. Then click Run Model. Study the Summary of Fit, Analysis of Variance, Effect Tests, and the Residual by Predicted Plot.

3. Using published papers for reference, craft a paragraph that precisely reports the conclusions of this ANOVA. It should probably be about 2-4 sentences with statistics cited in-line in the appropriate places.

   Dinner has a significant effect on *Time* (F(1,26) = 12.15, $p < .01$), with steak resulting in faster average times than fish (7.44 *vs.* 8.33 minutes). *Game* did not significantly affect *Time* (F(1,26) = 0.30, *n.s.*) and there was no *Dinner**Game* interaction (F(1,26) = 0.24, *n.s.*).

Part 3, REML. In part 2, you ran a traditional between-subjects ANOVA, but the state-of-the-art is an analysis of variance based on regression that uses REstricted (or REsidual) Maximum Likelihood, or REML. While the mathematics of this approach is beyond the current scope, you should know that

REML analyses of variance can do everything ANOVA can do, and more. Unlike in traditional ANOVA, with REML you can have dropped/missing data. Unbalanced designs are easily handled. Also, a mixture of categorical and continuous independent variables can be used. And for within-subjects or repeated measures designs, REML allows for the modeling of each subject with covariance structures, which can take into account learning over time. These issues will be covered in more detail in later modules.

For simple analyses, ANOVA and REML will agree. Prove this to yourself: go to Analyze > Fit Model. Add the same model terms you did before but this time, also add *Subject* to the model effects. Then highlight *Subject* and click on the Attributes menu and select Random Effect. Be sure the Method being used is REML and the Emphasis is Effect Leverage. Then run the model.

4. Verify that the results are the same as before.

   Although JMP's output is tabulated slightly differently, the results from the traditional between-subjects fixed-effects ANOVA and the new mixed-effects REML analysis agree exactly.

5. Does the REML model fit any better or worse than the ANOVA model did? How do you know?

   The Summary of Fit report gives us R^2 and adjusted R^2, indicating how well the model fits, explains, or predicts the outcome. For both analyses, these numbers agree (R^2 = .32). This makes sense because the model in both analyses is the same. It is the analysis mechanics that are different, not the models themselves.

Part 4, Random Effects. REML analyses use what are called mixed-effect models, where you have some model effects that are called "fixed," and some, in this case *Subject*, that are called "random." Fixed effects are those whose levels are chosen intentionally and beyond which generalizations are not meant to be drawn. Random effects, on the other hand, are those whose levels are pulled randomly from a population over which the results are meant to generalize. The specific levels of a random effect, therefore, are not of any interest, only that they are randomly sampled. *Subject* is a classic random effect.

6. In light of the description here, why is *Subject* a random effect?

   All experiments have a unit upon which measures are taken, and in human-computer interaction, the experimental unit is usually an individual human subject. It is customary to care about the specific user interface a subject uses, or the specific task a subject performs, or a specific strategy the subject was given… But what is *not* customary is to care about the specific subjects themselves, *i.e.*, about the levels of the *Subject* factor. While there may be things *about* each subject we care to record (*e.g.*, their gender, age, handedness, experience with computers, and so forth), we do not care about the distinction between, say, s01 and s30. We might even say, "a subject is a subject is a subject." Therefore, *Subject* is a classic random effect whose specific levels we do not regard as important. Rather, we wish to generalize to the whole population of possible subjects.

7. Provide an example of another random effect from a hypothetical or real experiment. (Hint: think of soil plots in agricultural studies.)

   In agricultural studies, different types of fertilizers, say, may be compared for how well they grow corn. Corn must be planted in various plots of soil, but we care only about the distinction among fertilizers, not into which patch of dirt the corn was planted. In fact, we wish to draw conclusions about fertilizers regardless of the patch of dirt in which they are used. Our data table might have a column called *Soil Plot* with levels "p01" through "p30", say, and these are exactly analogous to our human subjects in human-computer interaction studies. Soil plots are therefore random effects in agricultural experiments. If the same plot is given multiple seeds, we can think of this as the same human subject performing multiple trials in a repeated measures study design. In a long-format data table, the multiple seeds would result in the same soil plot identifier repeated on multiple rows of the table.

04. Statistics tools

posts., postsctrl.*, postsbtwn.*, postswthn.*, postsmix.*, postsmfx.**

Now that you have some initial exposure to statistical concepts, data sets, and analyses of variance, we are going to spend one module familiarizing ourselves with two industrial-strength statistics packages, SAS JMP and IBM SPSS. Many researchers use two (or more) tools when analyzing data. Doing so helps to compare results to ensure correctness. (Another popular tool is **R**, which is free but uses a command-line interface. You are welcome to use **R** but it is currently not supported as part of this independent study.)

Unlike the rest of the modules in this independent study, this module provides low-level details on carrying out analyses using these tools, and requires less independent discovery on your part. The goal here is to increase your comfort and familiarity with both JMP and SPSS, and to see how different statistics packages "name and frame" their features. The rest of the modules in this independent study will be based mainly on JMP, but with occasional forays into SPSS. After this module, cross-package exploration and comparison should be easy and is heretofore encouraged.

Free trial versions are available for both JMP and SPSS. JMP can be downloaded at http://www.jmp.com/. SPSS can be downloaded at http://www.spss.com/. (This module is based on JMP 7.0 and SPSS 19.0, but later versions of these products should suffice.) The file extensions for data sets in JMP are **.jmp*, and for SPSS are **.sav*.

Follow the instructions for each type of analysis *carefully*, and write up your answers to the questions that appear after each analysis. You are meant to perform each analysis in JMP and SPSS concurrently. This module can be completed with only one tool, but the comparative benefits are lost.

Part 1, *t*-test. Open the files *posts.** of a hypothetical study of 40 college students' Facebook posting behavior using one of two platforms: Apple's iOS or Google's Android OS. The data show the number of Facebook posts subjects made during a particular week using their mobile platform.

SPSS. Be sure View > Value Labels is checked for the data labels to show correctly. Perform a *t*-test by going to Analyze > Compare Means > Independent-Samples T Test. Transfer *Posts* into the Test Variable(s) box and *Platform* into the Grouping Variable. Click Define Groups. Enter "1" for Group 1 and "2" for Group 2. (This works because the *Platform* variable is defined as having one of two integer values, with 1 corresponding to "iOS," and 2 corresponding to "Android." To see this, go to the Variable View pane, and examine the Values column entry for *Platform*.) Click Continue. Then click OK.

JMP. Perform a *t*-test by going to Analyze > Fit Y by X. Put *Posts* into the Y, Response box and *Platform* into the X, Factor box. Click OK. Then access the menu from the little red arrow on the Oneway Analysis of Posts by Platform heading. (You can also always access the same menu by right-clicking anywhere on headings with red arrows.) From the menu, choose Means and Std Dev. Also choose Means/Anova/Pooled t. (For the *t*-test, you'll use only the results in the section labeled t Test. You can collapse unwanted sections using the little blue arrows.)

1. Is this a between-subjects or within-subjects experiment? Why?

   This is a between-subjects experiment because each subject was given only one level—iOS or Android—of the only factor in the experiment, *Platform*.

2. What is the independent variable named?

   The independent variable is *Platform*.

3. How many levels (*i.e.*, values) does the independent variable have? What are they?

   Platform has two levels, iOS and Android.

4. How many subjects were in this experiment?

   Forty subjects took part in this experiment, s01 through s40.

5. How many subjects were exposed to each level of the independent variable? Is the design balanced (*i.e.*, are the numbers equal)?

   Twenty subjects were exposed to each platform; therefore, the design is balanced.

6. What are the mean and standard deviation number of posts for each level of the independent variable?

   On average, Android resulted in 30.10 posts (*SD* = 8.80), while iOS resulted in 24.95 posts (*SD* = 7.04).

7. Assuming equal variances, what is the *t* statistic for this *t*-test? (Hint: this is also called the *t* Ratio.)

   The *t* statistic for this test is -2.04.

8. How many degrees of freedom are there (*df*s)?

   There are 38 degrees of freedom.

9. What is the two-tailed *p*-value resulting from this *t*-test? Is it significant at the $\alpha = .05$ level?

   The two-tailed *p*-value is .048. This is significant at the customary $\alpha = .05$ level because .048 < .050.

10. The formulation for expressing a significant *t*-test result is: *t*(*df*s) = *t*-statistic, $p < .05$. For a nonsignificant result, it is: *t*(*df*s) = *t*-statistic, *n.s.* Write your result in the proper formulation.

    t(38) = -2.04, $p < .05$.

11. What can we conclude about the results of this study? Which platform seems to result in more Facebook posts, if any? Write a sentence summarizing the finding from this analysis.

    College students who used Google's Android posted significantly more than those who used Apple's iOS (30.10 *vs.* 24.95, *t*(38) = -2.04, $p < .05$).

Part 2, One-way ANOVA (F-test). Still using the files *posts.**, we will perform an F-test, which can do everything a *t*-test can do, and more. An F-test, which is the most common analysis of variance, can handle multiple independent variables, or factors, and these factors can have more than two levels. By comparison, a *t*-test can only have one factor with two levels, which is not very useful for many experiment designs.

SPSS. To perform an F-test on the data in *posts.sav*, go to Analyze > Compare Means > One-Way ANOVA. Into the Dependent List, transfer *Posts* as your measure. Into the Factor box, transfer *Platform*. Click Options and check the Descriptive box. Click Continue and click OK.

JMP. To perform an F-test on the data in *posts.jmp*, go to Analyze > Fit Y by X. Put *Posts* into the Y, Response box and *Platform* into the X, Factor box. Click OK. From the red-arrow menu, choose Means/Anova/Pooled t. Now you will use the Analysis of Variance section and the Means for Oneway Anova section.

12. Do the number of observations (*N*'s) and means agree with those produced by the *t*-test? What are they? (If they do not agree, there is an error somewhere!)

    The number of observations agree at *N* = 40, with 20 for each platform.

13. In the ANOVA table, what is the F-statistic? What is the *p*-value? Is it significant at the α = .05 level?

    The F-statistic is also called the F Ratio and is 4.18. The corresponding *p*-value is .048. It is significant at the customary α = .05 level because .048 < .050.

14. How does this *p*-value compare to that produced by the *t*-test? Did you expect it to be otherwise?

    The *p*-value is the same as that produced by the *t*-test. We should expect it to be the same because although the *t* and F tests are different, they always produce the same results when analyzing a single factor with two levels.

15. *SPSS only*. If this experiment had more than one factor (see parts 5-8, below), we would use Analyze > General Linear Model > Univariate. Use this function on the same data and see if you achieve the same results. (Hints: you don't use *Subject*, and you don't have to select any options within a subdialog beneath the Univariate dialog.) Report the line on which your F-test result appears.

    We do achieve the same results. In our SPSS output, the line labeled "Platform" contains our F value of 4.177. The "Corrected Model" line also contains this F value because *Platform* is the only factor in the entire model.

16. *JMP only*. If this experiment had more than one factor (see parts 5-8, below), we would use Analyze > Fit Model. Use this function on the same data and see if you achieve the same results. (Hint: you don't use *Subject*, but put *Posts* as your response and *Platform* as your model effect.) Report the section(s) in which your results appear.

We do achieve the same results. In our JMP output, the line labeled "Platform" in the Effect Tests report contains our F ratio of 4.1772. This F value also appears in the Analysis of Variance report for the whole model because *Platform* is the only factor in the model.

17. The general formulation for expressing an F-test result is: F(df_{num},df_{den}) = F-ratio, $p < .05$ (or *n.s.*). Report the F-test result, filling in df_{num} and df_{den} with the numerator and denominator (also called "error") degrees-of-freedom, respectively.

    F(1,38) = 4.18, $p < .05$.

Part 3, Mann-Whitney *U* test, Wilcoxon rank sums test. The use of a *t*-test or an F-test requires the upholding of some assumptions, which will be covered in more detail in module 9. One of these assumptions is that the variance in the measures of subjects in each group is similar. Another is that the measures form a roughly normal distribution, or would if enough measures were taken from the population. When these assumptions are violated, a nonparametric test should be used. (Nonparametric tests are not a specific test but a class of test.) Such a test does not presume an underlying normal distribution. They are therefore sometimes called "distribution free." Such tests are useful when, for example, analyzing Likert scale data, which is often non-normal. The equivalent of a between-subjects (independent samples) *t*-test using nonparametric statistics is the Mann-Whitney *U* test. The Mann-Whitney *U* test is a test for an experiment containing one between-subjects factor with two-levels. More detail on nonparametric tests will be given in module 9; this exercise is just to give you some initial exposure.

SPSS. Perform a *U* test on *posts.sav* by going to Analyze > Nonparametric Tests > Legacy Dialogs > 2 Independent Samples. Transfer *Posts* into the Test Variable List. Transfer *Platform* into the Grouping Variable box. Click Define Groups and enter "1" for Group 1 and "2" for Group 2. Click Continue. Click Options and check the Descriptive box. Then check the Mann-Whitney U box. Click OK.

JMP. JMP does not perform a Mann-Whitney *U* test, but a similar test, which will give equivalent results, is the Wilcoxon rank sums test. (This is not to be confused with the popular matched-pairs Wilcoxon signed-rank test.) Perform a Wilcoxon rank sums test on *posts.jmp* by going to Analyze > Fit Y by X. Put *Posts* into the Y, Response box and *Platform* into the X, Factor box. Click OK. From the red-arrow menu, choose Nonparametric > Wilcoxon Test. The results are in the section labeled 2-Sample Test, Normal Approximation.

18. Do the Counts (*N*'s) agree with those produced by the *t*-test and F-test? What are they? (If they do not agree, there is an error somewhere!)

    The number of observations agree at *N* = 40, with 20 for each platform.

19. *SPSS only***.** What is the Mann-Whitney *U* value produced?

    The Mann-Whitney *U* value is 135.50.

20. *JMP only***.** What is the Wilcoxon *W* value produced? (Hint: this is the *S* value displayed.) Does it agree with the SPSS value for Wilcoxon W?

The Wilcoxon *W* value produced by JMP is 345.50. Yes, this value agrees with the Wilcoxon W value produced by SPSS.

21. This test produces a *Z*-value as its statistic. What is the *Z*-value?

    The Z-values are slightly different between SPSS and JMP. In SPSS, *Z* = -1.747. In JMP, *Z* = -1.734.

22. What is the two-tailed *p*-value? Is it significant at the α = .05 level? How does the result compare to the *t*-test and F-test results for the same data from parts 1 and 2, above?

    The *p*-value is .08, which is not significant at the α = .05 level because .08 > .05. The *p*-value of .08 is about .03 greater than the *p*-value for the parametric *t* and F tests, above. That indicates that this nonparametric procedure is, for this data, statistically less powerful.

23. The formulation for a Mann-Whitney *U* test result is *U* = U-value, *Z* = Z-value, *N* = *n*, *p* < .05 (or *n.s.*). The formulation for a Wilcoxon rank sums test is *W* = W-value, *Z* = Z-value, *N* = *n*, *p* < .05 (or *n.s.*). Report both tests, filling in the appropriate values. Note that there are no degrees of freedom in these nonparametric tests.

    The result for the Mann-Whitney *U* test is *U* = 135.50, *Z* = -1.75, *N* = 40, *n.s.* The result for the Wilcoxon rank sums test is *W* = 345.50, *Z* = -1.73, *N* = 40, *n.s.*

24. What can we conclude about the results of this study? Which platform seems to result in more Facebook posts, if any? Write a sentence summarizing the finding from this nonparametric analysis.

    The mobile operating system used by college students did not have a significant effect on the number of Facebook posts made according to [a Mann-Whitney *U* test (*U* = 135.50, *Z* = -1.75, *N* = 40, *n.s.*)][a Wilcoxon rank sums test (*W* = 345.50, *Z* = -1.73, *N* = 40, *n.s.*)].

Part 4, One-way ANOVA with 3+ levels. As noted above, the F-test can handle more than one factor, and also, more than two levels per factor. A one-way ANOVA refers to a single factor design. Similarly, a two-way ANOVA refers to a two-factor design, *i.e.*, two independent variables. In this part, we will still conduct a one-way ANOVA, but this time, our factor will have three levels. Thus, it cannot be analyzed with a *t*-test, which can only handle two levels of a single factor.

SPSS. Open *postsctrl.sav*. This data set is the same for the iOS and Android levels, but now has added 20 new college students as a control group who did not use a mobile device for posting on Facebook but were told to use their desktop computer instead. Thus, the *Platform* factor now has three levels: iOS, Android, and desktop. To perform an ANOVA, repeat the steps you did in part 2 using Analyze > Compare Means > One-Way ANOVA. However, this time, before executing the analysis, click Post Hoc and select LSD and Bonferroni. Click Continue and click OK.

JMP. Open *postsctrl.jmp*. To perform an ANOVA, repeat the steps you did in part 2 using Analyze > Fit Y by X. (On the red-arrow menu, you will choose Means/Anova, as Means/Anova/Pooled t has been automatically renamed to drop the "Pooled t" part because a *t*-test is no longer possible.)

25. Was this a one-way, two-way, or three-way analysis of variance? What is/are the factor(s)? What are each factor's levels?

    This was a one-way between-subjects ANOVA with a single factor *Platform* with three levels, iOS, Android, and desktop.

26. How many data points are there for each level of *Platform*? What are the means and standard deviations for each level? (*JMP* hint: use both Means/Anova and Means and Std Dev from the red-arrow menu.)

    Each level of *Platform* has *N* = 20 data points. On average, Android resulted in 30.10 posts (*SD* = 8.80), iOS resulted in 24.95 posts (*SD* = 7.04), and using only the desktop resulted in 31.70 posts (*SD* = 7.62).

27. In this case, the overall, or *omnibus*, F-test is testing for whether any differences exist among the levels of the independent variable. Is the F-test significant at the α = .05 level? What is the F-ratio? What is the *p*-value? Construct a proper reporting formulation for this result.

    The F-ratio is 4.03, and with *p* = .023, it is significant at the α = .05 level. Thus, there is a significant difference in the number of posts written on different platforms (F(2,57) = 4.03, *p* < .05).

28. *SPSS only.* The omnibus F-test does not tell us whether all three levels of *Platform* are different from one another, or whether just two levels (and which two?) are different. For this, we need *post hoc* comparisons, which are justified only when the omnibus F-test is significant. Examine the Post Hoc Tests output. The LSD output does *not* correct for making multiple *post hoc* comparisons (remember that with α = .05, we can expect 1 in 20 tests to be significant purely by chance). Which pairwise comparisons are significant without *post hoc* correction?

    Without correction, we see that iOS and Android resulted in a significantly different number of posts (*p* = .04), that iOS and the desktop resulted in a significantly different number of posts (*p* = .01), but that Android and the desktop did not result in a significantly different number of posts (*p* = .52).

29. *JMP only.* Use the red-arrow menu to choose Compare Means > Each Pair, Student's t. This performs *post hoc* pairwise *t*-tests for each pair of levels within our factor. These tests are *not* corrected for multiple comparisons (remember that with α = .05, we can expect 1 in 20 tests to be significant purely by chance). Which pairwise comparisons are significant without *post hoc* correction? (Hint: use the Level, -Level bottommost table produced to make this judgment, or use the red-arrow menu on the Comparisons header to select Detailed Comparisons Report.)

    Without correction, we see that the desktop and iOS resulted in a significantly different number of posts (*p* = .01), that Android and iOS resulted in a significantly different number of posts (*p* = .04), but that the desktop and Android did not result in a significantly different number of posts (*p* = .52).

30. A Bonferroni correction divides α by the number of *post hoc* comparisons. In this case, with three *post hoc* comparisons, we would use α = .05 / 3 = .0166. (Equivalently, we can multiply our *p*-values by 3 and then continue using α = .05, which is what SPSS's Bonferroni output does.) Under the Bonferroni correction, which pairwise comparisons are significant? Did any comparisons move from significant before correction to nonsignificant after? If so, which?

With a Bonferroni correction, the desktop and iOS remain significantly different (p = .0087 < .0166). However, Android and iOS are no longer significantly different (p = .0427 > .0166). And of course, desktop and Android remain nonsignificant (p = .5221 > .0166).

31. Assuming a Bonferroni correction, what can we conclude from this study? Write a sentence summarizing the findings from this analysis.

There is a significant main effect of *Platform* on *Posts* (F(2,57) = 4.03, p < .05). Pairwise comparisons with Bonferroni correction show that only the desktop and iOS are significantly different (t(57) = -2.72, p < .01).

32. This study added one control group. However, one other obvious control group is missing. If we were to add a second control group of 20 college students, what should that control group be?

We should include another group of 20 college students who are not told to use a mobile device *or* a desktop computer for posting on Facebook, so we can ascertain posting behavior apart from being placed into one of our three other conditions.

Part 5, Two-way between-subjects ANOVA. It is often the case that we wish to examine the effects of more than one factor, and we also care about the interaction among factors. Because multiple factors are involved, this is called a factorial design, expressed as $N_1 \times N_2 \times \ldots \times N_n$ for an arbitrary number n of factors, and where each N_i is an integer indicating the number of levels of that factor. In practice, it is difficult to interpret experiments with more than three factors, especially if those factors each have more than two levels.

For this part, we will examine an augmented version of our current study that adds another factor. Open *postsbtwn.**. You will see another column labeled *Day* with values "weekday" and "weekend." These values correspond to the days of the week the subject was allowed to post to Facebook.

SPSS. To perform a two-way between-subjects ANOVA, go to Analyze > General Linear Model > Univariate. You should be able to discover which variables go where. (Hint: only the Dependent Variable and Fixed Factor(s) boxes need to be filled in. *Subject* is not used. It is included in the data tables for clarity, and will come into play later when we use mixed-effects models in part 8, below.) Click Options and transfer all the contents of the Factor(s) and Factor Interactions box to the Display Means for box. Check Compare main effects. Check Descriptive statistics and Estimates of effect size. Click Continue and click OK.

JMP. To perform a two-way between-subjects ANOVA, go to Analyze > Fit Model. You should be able to discover which variables go where. Be sure that your model effects contain not only the independent variables, but also their interaction. (Hint: You should have one Y value and three items as model effects. *Subject* remains unused for now. It is included in the data tables for clarity, and will come into play later when we use mixed-effects models in part 8, below.) Click Run Model.

33. Was this a one-way, two-way, or three-way analysis of variance? What is/are the factor(s)? What are each factor's levels? Express the design using $N_1 \times N_2 \times \ldots \times N_n$ notation.

This was a 2×2 between-subjects design with factors *Platform* and *Day*. *Platform* has two levels, iOS and Android. *Day* also has two levels, weekday and weekend.

34. For each identified factor, was it between-subjects or within-subjects? How do you know?

    Both *Platform* and *Day* are between-subjects factors because each subject only experienced one level of each.

35. What were the means and standard deviations for the number of posts on weekdays? weekends?

    Weekdays saw an average of 26.30 posts (*SD* = 8.43). Weekends saw an average of 28.75 posts (*SD* = 8.17).

36. Write the F-test result for the *Platform* factor. Is it significant at the α = .05 level? What is its *p*-value? Does this differ from the finding prior to the inclusion of the *Day* factor (part 2, above)? If so, how?

    Platform did not exhibit a significant effect on *Posts*, although there was a trend in this direction (F(1,36) = 4.06, *p* = .052). This *p*-value is about .004 greater than we found for *posts.** in part 2, but just enough to move from statistically significant to statistically nonsignificant. Results with *p*-values between .05 and .10 are called "trends" or "marginal results."

37. Write the F-test result for the *Day* factor. Is it significant at the α = .05 level? What is its *p*-value?

    Day did not exhibit a significant effect on *Posts* (F(1,36) = 0.92, *n.s.*). The *p*-value is .34.

38. Write the F-test result for the *Platform*Day* interaction. Is it significant at the α = .05 level? What is its *p*-value?

    There was no significant *Platform*Day* interaction (F(1,36) = 0.00, *n.s.*). The *p*-value is .98.

39. Within each factor, why don't we need to perform any *post hoc* pairwise comparison tests?

    Within each factor, there is no reason to perform any pairwise comparisons because each factor has only two levels; a pairwise comparison would therefore just be a repeat of the main effect test for each factor. However, in this case, even *if* there were more than two levels in one or both factors, we lack the significant main effect results to justify *post hoc* pairwise comparisons.

40. Interpret these results and craft three sentences describing the results of this experiment, one for each factor and one for the interaction. What can we say about the findings from this study? (Hint: *p*-values between .05 and .10 are often called "trends" or "marginal results," and are often reported, although they cannot be considered strong evidence. Be wary of ever calling such results "marginally significant." A result is either significant or it is not; there is no "marginal significance.")

    Platform did not exhibit a significant effect on *Posts*, although there was a trend in this direction (F(1,36) = 4.06, *p* = .052), with Android resulting in 30.10 posts (*SD* = 8.80) and iOS resulting in 24.95 posts (*SD* = 7.04). *Day* did not exhibit a significant effect on *Posts* (F(1,36) = 0.92, *n.s.*), with weekdays averaging 26.30 posts (*SD* = 8.43) and weekends averaging 28.75 posts (*SD* = 8.17). There was no significant *Platform*Day* interaction (F(1,36) = 0.00, *n.s.*).

Part 6, Two-way within-subjects ANOVA. Thus far, we have only considered experiments where one subject was measured once on only one level of each factor. But often we wish to measure a subject more than once, perhaps for different levels of our factor(s), or over time, in which case *time* itself becomes a factor. Such designs are called "repeated measures" designs, and the factors on which we obtain repeated measures are called within-subjects factors (as opposed to between-subjects factors). For repeated measures studies, we can still use an ANOVA, but now we use a "repeated measures ANOVA," and our data table inevitably looks different: for a wide-format table, there are now multiple measures per row (each row still corresponds to just one subject, as it has thus far).

Our current hypothetical study on Facebook posts has been modified to be a purely within-subjects study. Imagine that each college student was issued either an iOS or Android device for one week, and then the other device for the next week. Also, each college student's posts were counted separately on weekdays and weekends. Instead of needing 40 college students as before, we now only need 10 students for the same data, which is shown in *postswthn.**. Open those files and see the wide-format data tables.

SPSS. To perform a traditional repeated measures ANOVA, go to Analyze > General Linear Model > Repeated Measures. For repeated measures analyses, the data table must be in wide-format, which has all measures for a given subject in a single row. (This contrasts with long-format, which has only one measure per subject in a given row, and for repeated measures designs, it must therefore have multiple rows for the same subject, which will occur in part 8, below.) Thus, we must tell SPSS which factors are represented by each column. For the first factor, enter the factor that changes the *slowest* across columns, *e.g.*, *Platform*. It should be clear by now that it has two levels—iOS and Android—so enter "2" in Number of Levels. For the second factor, enter the factor that changes *fastest* across columns, *e.g.*, *Day*. It also has two levels: weekday and weekend. Once these factors are added, click Define. Next, highlight the four result columns and transfer them to the Within-Subjects Variables box. Next click Options. Transfer everything in the Factor(s) and Factor Interactions box into the Displays Means for box. Check the Compare main effects box. Also check Descriptive statistics and Estimates of effect size. Click Continue and click OK.

JMP. To perform a traditional repeated measures ANOVA, go to Analyze > Fit Model. On the Fit Model dialog, put all four results columns into the Y box. For Personality, choose Manova. No model effects are used in the Construct Model Effects box, as everything is within-subjects here (that box would be used if we had any between-subjects factors). Click Run Model. Now click Choose Response > Compound, which is used for two layers of repeated measures. Because our factors are encoded in the columns themselves, we must tell JMP which factors are represented by each column. In the top textbox, change "factor1" to the factor that moves *fastest* across columns, namely *Day*. In the left textbox, change "factor2" to the factor that moves *slowest* across columns, namely *Platform*. Check the boxes for Create Interaction Effect also and Univariate Tests Also. Click OK.

41. Was this a one-way, two-way, or three-way analysis of variance? What is/are the factor(s)? What are each factor's levels? Express the design using $N_1 \times N_2 \times \ldots \times N_n$ notation.

    This was a 2×2 within-subjects design with factors *Platform* and *Day*. *Platform* has two levels, iOS and Android. *Day* also has two levels, weekday and weekend.

42. For each identified factor, was it between-subjects or within-subjects? How do you know?

    Both *Platform* and *Day* are within-subjects factors because each subject experienced all levels of both.

43. Verify that the means and standard deviations for the number of posts in each condition match the values produced in the previous version of the data (part 5, above).

The means in *postswthn.** exactly match those from *postsbtwn.**. Specifically, the four cell means and standard deviations are easily calculated using Table > Summary in JMP:

	weekday	***weekend***
Android	28.90 (8.96)	31.30 (8.93)
iOS	23.70 (7.39)	26.20 (6.83)

44. *SPSS only*. Write the statistical result for the *Platform* factor. (Hint: the table you care about is called Tests of Within-Subjects Effects. Use the Sphericity Assumed line, and remember that the word "error" is synonymous with "denominator" when reporting F-test results.)

Platform exhibited a significant effect on *Posts* (F(1,9) = 22.10, *p* < .01). The *p*-value shown is .001.

45. *JMP only*. Write the statistical result for the *Platform* factor. (Hint: you can collapse all of the Whole Model sections and ignore them. The section you care about is called Intercept within the Platform report. Use the Univar unadj Epsilon line, ignoring the Value column.)

Platform exhibited a significant effect on *Posts* (F(1,9) = 22.10, *p* < .01). The *p*-value shown is .0011.

46. Write the statistical result for the *Day* factor.

Day did not exhibit a significant effect on *Posts* (F(1,9) = 0.37, *n.s.*). The *p*-value is .559.

47. Write the statistical result for the *Platform*Day* interaction.

There was no significant *Platform*Day* interaction (F(1,9) = 0.00, *n.s.*). The *p*-value is .949.

48. What can we conclude about the results from this study?

From this study, we can conclude that *Platform* had a significant effect on *Posts* (F(1,9) = 22.10, *p* < .01), but that neither *Day* (F(1,9) = 0.37, *n.s.*) nor *Platform*Day* (F(1,9) = 0.00, *n.s.*) exhibited significant effects. Judging from means, it seems that Android had significantly higher numbers of posts compared to iOS (30.10 *vs.* 24.95).

49. In general, was the significance level as judged by the *p*-values for each factor stronger or weaker (*i.e.*, numerically lower or higher) than for the same data from 40 subjects in part 5, above? Why do you think that is?

Judging by the *p*-values, the results for the within-subjects study seem stronger than the results for the between-subjects study, even though numerically the studies contain the same data. The within-subjects *p*-value for *Platform* was .001 compared to .052 for the between-subjects *p*-value. This is the most dramatic difference between the two sets of results. For the *Day* factor, the within-subjects *p*-value was .56, surprisingly a bit higher than the between-subjects *p*-value of .34. The interaction *p*-values are nearly the same at .95 *vs.* .98. In general, we should expect within-subjects studies to be more statistically powerful than between-subjects studies because subjects are, in effect, compared to themselves, and any given subject is more like himself than he is like any other subject. Within-subjects designs therefore reduce variance compared to between-subjects designs, and thus result in more power for detecting differences. The downside of using within-subjects designs is that they are subject to carryover effects and require careful counterbalancing. This makes them impractical in many circumstances.

Part 7, Two-way mixed factorial design. After seeing purely between-subjects and purely within-subjects designs, you may wonder whether mixed factorial designs involving factors of each type are possible. Indeed they are, and *postsmix.** contains our same data in a hypothetical mixed design. Imagine that each college student is again assigned only one mobile device for a week, an Apple iOS device or Google Android device, but that the student is allowed to use it both on weekdays and on weekends. This is still considered a repeated measures design because at least one factor is within-subjects, but now it is a mixed factorial design, meaning at least one factor is also between-subjects.

SPSS. To perform a traditional repeated measures ANOVA, go to Analyze > General Linear Model > Repeated Measures and fill in the dialogs based on what you learned in the previous parts. Don't forget to add your between-subjects factor.

JMP. To perform a traditional repeated measures ANOVA, go to Analyze > Fit Model and fill in the dialog based on what you learned in the previous parts. (Hint: you'll have two Y responses.) Don't forget to add your between-subjects factor in the model effects box. Choose Manova as the Personality and click Run Model. Select Choose Response > Repeated Measures and enter for "Y Name" the factor embedded in your Y response columns, namely *Day*. Check the Univariate Tests Also box and click OK.

50. Was this a one-way, two-way, or three-way analysis of variance? What is/are the factor(s)? What are each factor's levels? Express the design using $N_1 \times N_2 \times \ldots \times N_n$ notation.

    This was a 2×2 mixed factorial design with a between-subjects factor *Platform* and a within-subjects factor *Day*. *Platform* has two levels, iOS and Android. *Day* also has two levels, weekday and weekend.

51. For each identified factor, was it between-subjects or within-subjects? How do you know?

    Platform was a between-subjects factor because each subject experienced only one of its two levels. *Day* was a within-subjects factor because each subject experienced all of its two levels.

52. Verify that the means for the number of posts in each condition match the values produced in the previous version of the data (parts 5 and 6, above).

    The means and standard deviations do indeed agree with those shown in the answer to question 43. In JMP, Tables > Summary can be used to quickly compute these means and standard deviations.

53. *SPSS only.* Write the statistical result for the *Platform* factor. (Hint: you care about the Platform line in the Tests of Between-Subjects Effects.)

    Platform exhibited a significant effect on *Posts* (F(1,18) = 5.72, $p < .05$). The p-value shown is .028.

54. *JMP only.* Write the statistical result for the *Platform* factor. (Hint: your result for Platform is in the Between Subjects > Platform report.)

    Platform exhibited a significant effect on *Posts* (F(1,18) = 5.72, $p < .05$). The p-value shown is .0280.

55. *SPSS only.* Write the statistical result for the *Day* factor. (Hint: you care about the Tests of Within-Subjects Effects.)

    Day did not exhibit a significant effect on *Posts* (F(1,18) = 0.71, *n.s.*). The *p*-value shown is .410.

56. *JMP only.* Write the statistical result for the *Day* factor. (Hint: use the Univar unadj Epsilon line in the Day report and ignore the Value column.)

    Day did not exhibit a significant effect on *Posts* (F(1,18) = 0.71, *n.s.*). The *p*-value shown is .4098.

57. Write the statistical result for the *Platform**Day* interaction. (*SPSS* hint: you care about the Tests of Within-Subjects Effects, and Error(Day) provides your denominator degrees of freedom.)

    There was no significant *Platform**Day* interaction (F(1,18)=0.00, *n.s.*). The *p*-value is .986.

58. What can we conclude about the results from this study?

    From this study, we can conclude that *Platform* had a significant effect on *Posts* (F(1,18) = 5.72, *p* < .05), but that neither *Day* (F(1,18) = 0.71, *n.s.*) nor *Platform**Day* (F(1,18) = 0.00, *n.s.*) exhibited significant effects. Judging from means, it seems that Android resulted in significantly more Facebook posts than iOS (30.10 *vs.* 24.95).

59. In general, was the significance level as judged by the *p*-values for each factor stronger or weaker (*i.e.*, numerically lower or higher) than for the same data from 40 subjects in part 5, above? than from the 10 subjects in part 6, above? In each case, why do you think that is? Remember that there were 20 subjects in this mixed design.

    The following table shows the *p*-values for each effect from the three analyses in parts 5-7, sorted by descending number of subjects:

	No. Subjects	*Platform*	*Day*	*Platform*Day*
Between	40	.052	.344	.985
Mixed	20	.028	.410	.987
Within	10	.001	.559	.949

What we can see from this table is that the mostly significant *Platform* result becomes ever-stronger as the number of subjects, and therefore the amount of inter-subject variability, decreases. This is what we should expect. As noted in the answer to question 49, using within-subjects designs usually results in more statistical power because any given subject is more like herself than she is like any other subject. Although the *Day* and *Platform**Day* effects do not show a similar trend as subjects decrease, these effects are nonsignificant and remain that way.

Part 8, Mixed-effects models. Although similar in terminology to a "mixed factorial design," a mixed-effects model analysis of variance means something else entirely. "Mixed effects" refer to the mixture of fixed and random effects, where "effects" are synonymous with factors. A fixed effect is everything we have considered thus far: it is a factor whose specific levels we care to draw conclusions about, and whose values were therefore determined *a priori* during experiment design. Parts 1-7 of this module have dealt only with fixed effects. Indeed, fixed effects ANOVAs represent the traditional method for analyzing parametric data. However, with the advent of personal computing, more computationally intensive analyses have become available. One is to perform analyses of variance (no longer written as "ANOVA" to avoid confusion) using an underlying regression-based model, covariance matrices, and a method known as REstricted (or REsidual) Maximum Likelihood, or REML. You briefly encountered REML in module 3.

A REML analysis has many advantages over a traditional fixed-effects ANOVA. Unlike in traditional ANOVA, a REML analysis can have dropped or missing data. Unbalanced designs are also handled. Also, a mixture of nominal and continuous factors can be used, instead of only nominal factors. And for repeated measures designs, REML allows for separate modeling of each subject over trials using covariance matrices, which takes into account learning. REML analyses are therefore preferred to traditional repeated measures ANOVAs for analyzing repeated measures.

So what is a *random effect* that makes such models "mixed?" Unlike a fixed effect whose levels you specifically care about, a random effect is a factor whose levels are sampled randomly from a larger population about which we care to draw inferences. Accordingly, we do not care about the specific levels of a random effect. Classic examples of random effects are the actual college students used in our Facebook posting study, or the plots of soil into which different types of seeds are planted. We wish to generalize to all college students and all soil plots, respectively, and do not care specifically about which humans were used or in which plots we planted.

Unlike the wide-format data tables used in traditional repeated measures ANOVAs, the table for mixed-effects model analyses must be in long-format. For a given dependent variable, long-format tables have only one measure per row, and with repeated measures, subjects therefore must be listed on multiple rows. We represent the within-subjects data from part 6 in long-format as *postsmfx.**. Remember that there are only 10 unique subjects for this hypothetical within-subjects experiment. So why are there 40 rows in our long-format table? Because each subject is repeated four times. Now our *Subject* column finally matters. Thus, our columns are the same as for our fully between-subjects data in *postsbtwn.**, but now with only 10 unique subjects, not 40.

SPSS. To perform a mixed-effects model analysis of variance on *postsmfx.sav*, go to Analyze > Mixed Models > Linear. Transfer *Subject* into the Subjects box and *Platform* and *Day* into the Repeated box, as these are repeated factors. For Repeated Covariance Type, choose Compound Symmetry. (This choice will be revisited in module 8.) Click Continue. The Dependent Variable is *Posts*. The Factor(s) are *Platform* and *Day*. Then click Fixed and highlight *Platform* and *Day* in the Factors and Covariates box. Click Add to show these factors plus their interaction in the Model box. Click Continue. Click EM Means, highlight everything in the Factor(s) and Factor Interactions box and transfer it to the Display Means for box. Check the Compare main effects box. Click Continue and click OK.

JMP. To perform the mixed-effects model analysis of variance on *postsmfx.jmp*, go to Analyze > Fit Model. Put *Posts* in the Y box. Highlight both *Platform* and *Day* and choose Macros > Full Factorial to construct your model effects. Next, for the first time, also put *Subject* in the model effects box. Then highlight *Subject* and use the Attributes red-arrow menu to select Random Effect. Subject should become "Subject& Random." Be sure Personality is Standard Least Squares, Emphasis is Effect Leverage, and Method is REML. Click Run Model. You will mostly care about the results in the Fixed Effects Tests section.

60. What are the fixed effects? What are the random effects? How many levels does each of these effects have? What are those levels?

    The fixed effects are *Platform* and *Day*. The former has two levels, iOS and Android. The latter has two levels, weekday and weekend. The random effect is *Subject*. It has 10 levels, s01 through s10.

61. Verify that the means and standard deviations for the number of posts in each condition agree exactly with those from the wide-format within-subjects data from part 6, above.

    The means for *Posts* can be easily computed using Tables > Summary in JMP. The means and standard deviations agree exactly with those shown in the answer to question 43.

62. Write the statistical result for the *Platform* factor. How does it compare to the result from the traditional repeated measures ANOVA from part 6, above? (Do not just compare the *p*-values, but also the F-ratios and the numerator and denominator degrees-of-freedom.)

    Platform exhibited a significant effect on *Posts* (F(1,27) = 4.41, *p* < .05). The *p*-value is .045. The F ratio from the traditional repeated measures ANOVA was 22.10, quite a bit higher than 4.41 for the same within-subjects data. The *p*-values are .045 for the mixed-effects model and a much lower .001 for the traditional repeated measures ANOVA. The denominator degrees-of-freedom are also different, with 27 for the mixed-effects model and only 9 for the repeated measures ANOVA. In general, it seems as if it has been easier, in this case, to detect statistical significance with the repeated measures ANOVA than with the mixed-effects model analysis of variance.

63. Write the statistical result for the *Day* factor. How does it compare to the result from the traditional repeated measures ANOVA?

    Day did not exhibit a significant effect on *Posts* (F(1,27) = 1.00, *n.s.*). The *p*-value is .327. The F-ratio from the traditional repeated measures ANOVA was 0.37, lower than 1.00 for the same within-subjects data. The *p*-value of the mixed-effects model is .327 compared to .559 for the repeated measures ANOVA. And the denominator degrees-of-freedom is 27 compared to 9. In this case, it seems that the mixed-effects model produced a stronger, although still nonsignificant, statistical result than the repeated measures ANOVA.

64. Write the statistical result for the *Platform*Day* interaction? How does it compare to the result from the traditional repeated measures ANOVA?

    There was no significant *Platform*Day* interaction (F(1,27) = 0.00, *n.s.*). The *p*-value is .984. The F ratio from the traditional repeated measures ANOVA was 0.0044 compared to just 0.0004 for the mixed-effects model. The *p*-value is .9839 for the mixed-effects model compared to .9486 for the repeated measures ANOVA. The denominator degrees-of-freedom is again 27 compared to just 9.

65. What can we conclude about the results from this study?

    From this study, we can conclude that *Platform* had a significant effect on *Posts* (F(1,27) = 4.41, *p* < .05), but that neither *Day* (F(1,27) = 1.00, *n.s.*) nor *Platform*Day* (F(1,27) = 0.00, *n.s.*) exhibited significant effects. Judging from means, it seems that Android produced significantly more Facebook posts than iOS (30.10 *vs.* 24.95).

66. Share any thoughts you have about how mixed-effects models compare, either statistically or conceptually, to traditional ANOVAs.

    It is clear from the description in part 8, above, that mixed-effects models are very powerful and offer advantages over traditional repeated measures ANOVAs with only fixed effects. It seems from these data that despite mixed-effects models retaining larger denominator degrees-of-freedom than repeated measures ANOVAs, mixed-effects models do *not* make it easier to achieve statistical significance. On the contrary, the F-ratio for the significant *Platform* factor is lower, and the *p*-value correspondingly higher, for the mixed-effects model than for the repeated measures ANOVA from part 6, above. It turns out that mixed-effects models use wider confidence intervals than repeated measures ANOVAs to compensate for these greater *df*s, thereby fending off an increase in Type I errors. A convenience of using mixed-effects models in human-computer interaction studies that use logging is that such logging usually produces long-format tables appropriate to mixed-effects model analyses.

Part 9, Contrast tests. As we briefly discussed in part 4, above, a significant omnibus test for a given factor simply indicates that there exists *some* significant difference among levels of that factor. If the factor has only two levels, then we know where that difference lies. But if the factor has more than two levels, a *post hoc* comparison, called a "contrast test" or "pairwise comparison," is probably in order using the Bonferroni correction.

Another circumstance in which contrast testing is useful is *across* factors in a factorial design. For example, in *postsmfx.**, suppose we wish to know whether there was a significant difference between numbers of posts in on weekdays using Android, and numbers of posts on weekends using iOS. How would we test for such a difference? Contrast testing provides the answer.

SPSS. First, it is illustrative to graph the interaction. Using the *postsmfx.sav* data table from part 8, go to Graphs > Chart Builder. (Click OK on the dialog that appears.) Highlight Line in the Choose from box, and drag the Multiple Line graph into the chart preview area. From your Variables box, drag *Posts* onto the Y-Axis? box. Drag *Day* onto the X-Axis? box. Drag *Platform* into the Set color area. Then click OK. Visit your SPSS output to see the graph. You can then visually compare (Android, weekday) to (iOS, weekend).

By default, SPSS will not perform contrast tests *across* factors. To do this, we need to use the SPSS Syntax Editor. After running your mixed-effects model analysis of variance from part 8, go to the SPSS Statistics Viewer and click on the code snippet that SPSS generated for you to perform that analysis. Use CTRL+C to copy it. Then use File > New > Syntax to open the syntax editor. Use CTRL+V to paste the code snippet there. Erase any initial lines before the `MIXED` command. You should see the last four lines of code using the `EMMEANS` command. For *Platform* and *Day*, these lines are accompanied by the `COMPARE` statement, which will compare levels separately within *Platform* and *Day*. No such statement accompanies *Platform*Day*. We can add the `COMPARE` statement to this line, specifying the factor whose levels we wish to compare *within* the levels of the other factor. Thus, `COMPARE(Platform)` will compare platforms within each level of *Day*. However, this cannot be used to compare *across* factors of an interaction. For that, we need the `TEST` subcommand. With this command, we can specify exactly the contrast we want by giving weights.[1] In the end, your script should look like this. Only the last two lines have been modified, with the modifications shown in bold:

```
   MIXED Posts BY Platform Day
     /CRITERIA=CIN(95)  MXITER(100)  MXSTEP(10)  SCORING(1)  SINGULAR(0.000000000001)
HCONVERGE(0, ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
     /FIXED=Platform Day Platform*Day | SSTYPE(3)
     /METHOD=REML
     /REPEATED=Platform*Day | SUBJECT(Subject) COVTYPE(CS)
     /EMMEANS=TABLES(OVERALL)
     /EMMEANS=TABLES(Platform) COMPARE ADJ(LSD)
     /EMMEANS=TABLES(Day) COMPARE ADJ(LSD)
     /EMMEANS=TABLES(Platform*Day) COMPARE(Platform) ADJ(LSD)
     /TEST = 'Contrast of Android,weekday to iOS,weekend' Platform*Day 0 -1 +1 0 .
```

☠ Unfortunately, for reasons I have not been able to discern, the `TEST` subcommand for the *postsmfx.sav* data set will not execute. SPSS complains that the "L matrix is not estimable." Searching the Web and SPSS documentation did not provide any answers. The above detail is included here for future reference, as the `COMPARE` option is sufficient when we want to test *within* a factor, and apparently the `TEST` command usually works just fine. (If you decode this mystery, please let me know!)

Note: If upon running this script you receive an error concerning the `CRITERIA` subcommand, it is probably because you have a broken line due to line-wrapping. Closely examine your syntax for the `CRITERIA` subcommand and ensure that the script is not broken across lines.

JMP. First, it is illustrative to graph the interaction. On your statistical output from part 8, find the report labeled Platform*Day. Use its red-arrow menu and choose LSMeans Plot. You will see that the *x*-axis is assigned the *Day* factor, and the lines themselves are assigned the *Platform* factor. Use the red-arrow menu again to uncheck LSMeans Plot. Now, hold down SHIFT *before* clicking, and use the red-arrow menu again to choose LSMeans Plot. This time, you can see the graph has been "sliced" the other way, with *Platform* on the *x*-axis and *Day* assigned to the lines. Now re-slice the graph the original way (without SHIFT) and then proceed to perform a contrast test.

Use the red-arrow menu to choose LSMeans Contrast. Here, you choose opposing positive and negative weights to indicate which levels you wish to combine and compare. To test (Android, weekday) *vs.* (iOS, weekend), toggle weights +1 and -1 next to each of these, respectively. Then click Done.

67. Describe the graph(s) you created. What main effect does the slant in the two lines describe? What main effect does the vertical gap between the two lines describe? What effect do the parallel lines describe?

    The graphs in SPSS and JMP both show two upward-slanting parallel lines. From left-to-right, the lines move from weekday to weekend. Because the lines slant upward, it is clear that the number of posts, indicated on the *y*-axis, increases on the weekends. Because both lines slant upward, we see that this occurs for both platforms about equally, as each line encodes one platform. The line for Android is consistently above the

[1] See http://www.ats.ucla.edu/stat/spss/faq/test_subcommand_mixed_mod.htm for more detail.

line for iOS, indicating that on average, Android resulted in more posts than iOS. Because these lines are parallel, we see that each platform was affected by the change from weekdays to weekends in the same way, so there is no significant *Platform*Day* interaction. (Such an interaction would require these lines to be non-parallel, indicating an asymmetric effect.)

68. *JMP only*. What is the result for the contrast test you performed for (Android, weekday) *vs.* (iOS, weekend)? Report and interpret this result as you would in a paper.

A pairwise comparison among college students using Android on weekdays versus iOS on weekends was nonsignificant (F(1,27) = 0.61, *n.s.*). (The specific *p*-value was .443, but we would not write this part in a paper.)

69. When performing only one *post hoc* contrast test, why is a Bonferroni correction unnecessary?

When performing only one *post hoc* contrast test, a Bonferroni correction is unnecessary because it would cause us to divide our significance threshold α = .05 by 1, yielding an unchanged threshold.

05. Repeated measures

cockpits.jmp

In this module, we are going to look again at within-subjects designs due to their prevalence in human-computer interaction studies. As you know, these are also called repeated measures designs because each subject is measured more than once. This is in contrast to between-subjects designs, where each subject is measured only once. Remember that mixed factorial designs have both between- and within-subjects factors. Note that "repeated measures" is a term sometimes used to refer to longitudinal studies, where subjects are measured more than once *over time*, perhaps even with *time* itself being the only independent variable to vary from one measure to the next (*i.e.*, all other factors' levels are kept the same). We will be considering longitudinal studies in module 8.

This module uses data table *cockpits.jmp*, which is the same data as one of the JMP sample files, *cholesterol.jmp*, but recast to represent a fictitious human factors study of combat pilots. Open *cockpits.jmp* table before answering the following questions.

Part 1, Table examination. Carefully examine the data table. Also, right-click on Notes > Edit at the top-left of the data table to read study notes for this fictitious study. (Such notes can be created in JMP by using the red-arrow menu on "cockpits" and choosing New Table Variable. Also, notes can be added to individual columns by right-clicking on the column name, choosing Column Info > Column Properties > Notes.)

1. Is this table in wide-format or long-format? What would the other look like (*i.e.*, how many rows and columns, and what might the column names and variable types be)?

   The table *cockpits.jmp* is in wide-format because each pilot is listed on only one row, and yet there are six measures per row. The long-format version of this table would have 120 rows, with columns named, *e.g.*, *Pilot*, *CockpitUI*, *Month*, *TimeOfDay*, and *Minutes*. All but *Minutes* would be nominal factors, and *Minutes* would be a continuous response measure.

2. Characterize the design of this experiment in the appropriate experimental terms: what are the factors and their levels? For each factor, is it within-subjects or between-subjects?

   The data is from a 4×3×2 mixed design. *CockpitUI* is a nominal between-subjects factor with 4 levels: A, B, Control, and Placebo. *Month* is a nominal within-subjects factor with 3 levels: April, May, and June. *TimeOfDay* is a nominal within-subjects factor with 2 levels: day and night. There are 20 pilots each measured six times as part of this study.

3. Describe the experiment in everyday terms. What is a plausible account of what was done here? Also, explain the importance of the Placebo level of *CockpitUI*.

   The Notes for the data table provide the answer to the first part: In this fictitious study, 20 trainee pilots were tested six times on one of four cockpit user interfaces for how long (in minutes) it took them to complete a challenging dogfight in a flight simulator during day and night conditions. Pilots trained daily and were tested on the first day of three months. The A and B interfaces were novel cockpit designs created by two competing Navy human factors teams. The Control interface was the *de facto* cockpit interface currently in use. The Placebo interface was one the pilots were led to believe was novel, but which actually was not, as it exhibited only trivial alterations from the Control interface.

The Placebo cockpit is important for ruling out that any improvements with novel cockpit designs A and B are due to a novelty effect and not due to better design. Presumably, the pilots would already have been familiar with the Control interface, and thus their excitement would be less than for interfaces A and B. By leading pilots to believe the Placebo interface was *also* novel, a novelty effect can be examined and, if A and B should perform well, ruled out as the cause of that increased performance.

Part 2, Descriptive statistics. Using JMP, go to Analyze > Distribution, and place the six result columns in Y, Columns. Then on the Distributions menu, choose Stack and Uniform Scaling. Visually compare the plots.

4. Does the data seem normal? (Non-normal data is not amenable to ANOVA and must be transformed before an ANOVA is performed. Transformations are the topic of module 6.)

   The data appear more normal for the earlier months (*i.e.*, April) than the later months (*i.e.*, June). By June, the data has become bimodal (*i.e.*, with two humps), as presumably the treatment groups separate from the control and placebo groups, which are "left behind" near where all pilots began in April. Overall, this data does seem amenable to ANOVA, especially if analyzed within treatment groups.

5. Can you eyeball the results of the within-subjects significance tests from these distributions? Make your predictions *before* proceeding. (Hint: you should be making predictions about two factors and their interaction here, but there is another factor not visible in these distributions. Which is that factor? Why isn't it visible here?)

   The progression from April to June is the *Month* factor, and dogfight completion times certainly appear to decrease as *Month* changes. Thus, there should be a significant effect of *Month* on minutes taken to complete the dogfight. The day/night pairs within each month do not show much change, so we can be skeptical that there is a significant *TimeOfDay* effect. The change across day/night does not itself seem to differ across months, so we also can be skeptical of a *Month*TimeOfDay* interaction. The factor not visible in these distributions is the between-subjects *CockpitUI* factor, as all treatments are lumped together. We could remedy this by adding *CockpitUI* to the By box on the Analyze > Distribution dialog, but this produces an unwieldy number of reports.

Part 3, Repeated measures ANOVA. Now we will perform a repeated measures ANOVA. Note that in papers and other write-ups, calling this a "repeated measures ANOVA" is important, not just an "ANOVA." Reviewers want to ensure this data was analyzed by taking repeated measures into account, *i.e.*, that measures on the same subject were not regarded as independent. To perform it, go to Analyze > Fit Model. Put your six result columns in Y. Put your one between-subjects factor in Construct Model Effects. (Remember that your subject term, *Pilot*, is not used as a model effect unless we are using mixed-effects models with long-format tables.) For Personality, choose Manova. Then click Run Model.

6. Look at the graphs in the Least Squares Means section. (You can drag a graph's borders to enlarge it.) Eyeballing the graphs, can you now make better predictions about which effects are significant? Remake your predictions here. (Hint: you should be making predictions about three factors and all possible interactions among them.)

> The Overall Means graph shows a steady progression of faster dogfight completion times across months. This suggests a main effect of *Month* on completion time. This graph also shows a slight but reliable increase in completion time within each month when moving from daytime to nighttime dogfights. There may therefore be a *TimeOfDay* main effect. This slight increase seems proportional across months, however, so there should be no *Month*TimeOfDay* interaction. The CockpitUI graph shows the overall data broken out by level of *CockpitUI*. We can clearly see that Control and Placebo interfaces showed no decrease in completion time, but experimental interfaces A and B did exhibit decreased completion times overall. Thus, there is a main effect of *CockpitUI*. Because different interfaces behaved differently across months, there should be a significant *CockpitUI*Month* interaction. Regardless of interface, there seems to be the same slight increase in completion times at night, so there should be no *CockpitUI*TimeOfDay* interaction. As for a three-way *CockpitUI*Month*TimeOfDay* interaction, there should be none because *TimeOfDay* does not seem to change the nature of the *CockpitUI*Month* interaction.

First, we have to test the between-subjects factor, *CockpitUI*. Using the Choose Response menu, select Sum. Click Run. The report you care about is under the CockpitUI heading. It shows the main effect of *CockpitUI*.

7. Report and interpret the *CockpitUI* effect. Do not just report the statistical result, but also interpret its meaning using means and standard deviations. (Hint: you may find it convenient to use Microsoft Excel to help compute these values.)

> The different cockpit interfaces showed a significant effect on pilots' dogfight completion times (F(3,16) = 274.97, *p* < .0001). On average, pilots flying in cockpit A had a completion time of 220.63 minutes (*SD* = 43.91), pilots flying in cockpit B had a completion time of 233.63 minutes (*SD* = 36.82), pilots flying the Control cockpit had a completion time of 279.73 minutes (*SD* = 4.16), and pilots flying the Placebo cockpit had a completion time of 280.63 (*SD* = 4.96). Thus, cockpits A and B seemed to reduce dogfight completion time compared to the Control and Placebo cockpits.

Now we want to examine within-subjects factors and their relation to each other and to the between-subjects factor. From the Choose Response menu, choose Compound. This menu item is used when *two* within-subjects factors are represented across columns. (If we had only one within-subjects factor, we could choose Repeated Measures instead.) The factor that changes fastest across columns should be named at the top, and its levels specified. In this case, put "TimeOfDay" in the top textbox and indicate that it has 2 levels. Put "Month" in the left textbox and then study the grid: it should look properly laid out and labeled. Also, check the boxes for Create Interaction Effect also and Univariate Tests Also. Then click OK.

Find the TimeOfDay results section. (It is in small font under the first section called Compound.) The *TimeOfDay* main effect is under the Intercept section, and is on the Univar unadj Epsilon line. We ignore the Value column, but we do care about the Exact F (the F-test value), NumDF (numerator degrees of freedom), DenDF (denominator degrees of freedom), and Prob>F (the *p*-value) columns.

8. Report and interpret the *TimeOfDay* effect. Do not just report the statistical result, but interpret its meaning, also. (Hint: you need to look at means for levels of *TimeOfDay* to do so.)

> Whether dogfights were during the daytime or nighttime had a significant effect on completion time (F(1,16) = 360.94, *p* < .0001). On average, daytime completion times were 251.65 minutes (*SD* = 39.53), while

nighttime completion times were 255.67 minutes (*SD* = 39.22). Thus, dogfight completion times were consistently lower in the daytime, although only by about 4 minutes.

In the CockpitUI section in the TimeOfDay results, we find the *CockpitUI*TimeOfDay* interaction. Again we care about the result on the Univar unadj Epsilon line.

9. Report and interpret the *CockpitUI*TimeOfDay* interaction. What, conceptually, is this interaction testing?

The *CockpitUI*TimeOfDay* interaction was nonsignificant (F(3,16) = 0.63, *n.s.*). The *p*-value shown is .6038. Conceptually, this interaction is testing whether the effect on dogfight completion time that each cockpit interface had depended on whether that cockpit was used during the daytime or nighttime. To find no significant interaction means that the time of day affected all cockpits' performances proportionally.

Now find the *Month* results section. Notice here that we have a Sphericity Test, which is a test that takes place in repeated measures ANOVA but is unnecessary in mixed-effects model analyses of variance. If the Sphericity Test is significant, we cannot use an unadjusted ANOVA result for that particular factor, and instead must use a corrected result, such as the Greenhouse-Geisser (G-G) results. The G-G results can be found on the Univar G-G Epsilon line. For the *Month* factor, the Sphericity Test is nonsignificant (*p* = .27), so we do not have a violation of sphericity and can report the unadjusted result as we have been doing.

10. In the Intercept section for Month, find, report, and interpret the *Month* main effect. What is this testing? What does the result indicate?

The pilots' extended training over months from April through June had a significant effect on dogfight completion times (F(2,32) = 253.15, *p* < .0001). On average, completion times in April were 279.15 minutes (*SD* = 4.45), in May were 248.60 minutes (*SD* = 35.23), and in June were 233.23 minutes (*SD* = 48.18). Thus, dogfight completion times decreased as training progressed over months.

11. Now find the *CockpitUI*Month* interaction, and report and interpret that result.

The *CockpitUI*Month* interaction was significant (F(6,32) = 92.05, *p* < .0001), indicating that the effect different cockpit interfaces had on dogfight completion times differed depending on the month (*i.e.*, depended on how long the pilots had been training with each interface). Our cockpit graph (*see* question 6) clearly shows that interfaces A and B had very different completion times in months May and June than did the Control and Placebo interfaces, supporting this finding.

Find the *TimeOfDay*Month* interaction results.

12. Is a Greenhouse-Geisser correction necessary? How do you know?

The Greenhouse-Geisser correction is unnecessary because sphericity is not violated according to a nonsignificant sphericity test (*p* = .37).

13. In this report, the Intercept section is the result for the *TimeOfDay*Month* interaction. Is it significant? Report and interpret the result.

The *TimeOfDay*Month* interaction was nonsignificant (F(2,32) = 1.59, *n.s.*). The *p*-value shown is .2203. This result indicates that the effect that the daytime/nighttime difference had on dogfight completion times was proportional across months.

14. What is this interaction testing? What does it mean?

We already know from our finding in question 8 that time of day (daytime *vs.* nighttime) affected dogfight completion times—dogfights took consistently longer at night. This interaction tests whether that finding is affected or altered by the month in which pilots were measured. In other words, we are testing whether there is a disproportionate effect of the daytime/nighttime change in completion times depending on the month.

Finally, find the results for the *CockpitUI*TimeOfDay*Month* three-way interaction.

15. Report and interpret the *CockpitUI*TimeOfDay*Month* interaction.

The *CockpitUI*TimeOfDay*Month* interaction was nonsignificant (F(6,32) = 1.50, *n.s.*). The *p*-value shown is .2101. This means that the time of day did not alter the nature of the significant *CockpitUI*Month* interaction, *i.e.*, this interaction was consistent across times of day.

16. What is this three-way interaction testing? (You will quickly realize that gaining an intuition for a three-way interaction is difficult, but not impossible. But four-way interactions just might be impossible to intuitively comprehend!)

The three-way interaction is testing whether a third interacting factor affects the interaction between the other two factors. We can slice this however we wish. For example, since we know there was a significant *CockpitUI*Month* interaction (*see* question 11), which was due to different cockpit interfaces having different effects on dogfight completion time based on the month, then we can regard this three-way interaction as questioning whether the additional *TimeOfDay* factor alters that extant *CockpitUI*Month* interaction. Is that interaction different based on the time of day? Examining the CockpitUI graph suggests that it is not; the *CockpitUI*Month* relationship seems consistent for all daytime/nighttime pairs.

Part 4, Avoiding confounds from carryover effects. In experiments with repeated measures, carryover effects from one level of a factor to another can confound results if not properly neutralized through counterbalancing. Full counterbalancing requires that all possible orderings of a factor's levels are administered to subjects an equal number of times. Partial counterbalancing is also possible if many levels of repeated factors are present. A popular scheme for partial counterbalancing is called a Latin Square. Carryover effects are unavoidable, but counterbalancing distributes these effects evenly across levels—that is, provided there is no *asymmetric skill transfer*, where the carryover effect differs depending on which levels came first. In cases where asymmetric skill transfer is present, a between-subjects factor should be used instead of a within-subjects factor.

17. For each factor in this study, discuss whether it requires counterbalancing. For those factors that require it, how could counterbalancing be achieved? What are the implications if counterbalancing is not employed? What confounding issues could explain the results if not counterbalanced out? What would be an example of asymmetric skill transfer? How can it be handled?

CockpitUI is a between-subjects factor and so counterbalancing does not apply. *Month* is a within-subjects factor but represents the passage of time, which cannot be counterbalanced. Put another way, more pilot training cannot be "administered" before less pilot training. In fact, change over time is of direct interest in the study and we would not *want* this factor to be counterbalanced out.

However, *TimeOfDay* does require counterbalancing. This factor is a within-subjects factor indicating whether a dogfight was run during daytime or nighttime conditions on the first day of each month. To fairly assess whether the time of day affects dogfight completion times, we cannot run all daytime dogfights before all nighttime dogfights (or *vice versa*) because, *e.g.*, pilot fatigue (or pilot learning) in the second fight could affect the results, rather than the change in time of day. To counterbalance the effects of fatigue (or learning), half of the *first* dogfights must be daytime fights and half nighttime fights for each cockpit interface and within each month. Also, the same pilots should not always have one or the other time of day first, but each experience an even number of both so that individual pilot skills are not a confounding issue. Thankfully, our study was conducted in a flight simulator, allowing the *TimeOfDay* factor to be controlled via simulation.

Asymmetric skill transfer would occur in this case if, *e.g.*, daytime runs somehow prepared pilots better for nighttime runs than *vice versa*. In this case, pilot learning from daytime runs would asymmetrically improve performance on nighttime runs, but the reverse improvement would not occur. Counterbalancing is powerless to completely mitigate such an asymmetric effect, and a between-subjects factor for *TimeOfDay* should be employed.

Part 5, Change table format. Recall from modules 3 and 4 that mixed-effects models are useful, often even preferable, for repeated measures designs. But to perform a mixed-effects model analysis of variance, you need a long-format table. Recall your answers from question 1, above.

18. What columns should your long-format table have?

    Our long-format table should have columns for *Pilot*, *CockpitUI*, *Month*, *TimeOfDay*, and *Minutes*.

19. How many rows should it have?

    Our long-format table should have 120 rows.

20. For how many rows should each subject be repeated? Be sure you understand exactly what the long-format version of this data should look like. You might consider first sketching it out on paper before proceeding.

    Each of 20 pilots was measured 6 times (3 months × 2 times of day). This gives us our 120 rows, and also indicates that each individual pilot will be repeated on 6 separate rows.

Although Tables > Stack provides support for transforming a wide-format table into a long-format one (and Tables > Split does the reverse), Tables > Stack is not well designed and causes more headaches than it relieves. (Tables > Split, on the other hand, is very useful.) For a table this small, it is feasible to manually create a long-format table using copy and paste. You can do this in Microsoft Excel and paste it back into an empty data table in JMP, or do it directly in JMP. Be extremely careful not to corrupt your data!

21. Once your long-format table is completed, state the number of columns and rows it has, and what its column names are. (These names should correspond exactly to the factor names from your analysis in part 3, above.)

    After it is finalized, our table has 5 columns with the following names: *Pilot*, *CockpitUI*, *Month*, *TimeOfDay*, and *Minutes*. All are nominal except for *Minutes*, which is continuous. The table has 120 rows for 20 pilots, with each pilot repeated on 6 rows.

22. Sanity-check your transformed data by computing means and standard deviations (use either Tables > Summary, or easier, Analyze > Distribution). Ensure that these values match those from your wide-format table. If all the means, standard deviations, quartiles, *etc.* agree, you'll know you transformed your data table correctly from wide-format to long-format.

    For convenience, we use Analyze > Distribution on *cockpits.jmp*, and Tables > Summary on *cockpits.long.jmp*, our new long-format table, calculating the mean and standard deviation of *Minutes* grouped by *Month* and *TimeOfDay*. We compare the Moments report for each (*Month*, *TimeOfDay*) pair to the values calculated by Tables > Summary. All values agree.

Part 6, Mixed-effects model. Now perform a mixed-effects model analysis of variance using the REML method from modules 3 and 4. Recall that we avoid writing "ANOVA" because that connotes a traditional ANOVA involving only fixed effects. (Hint: go to Analyze > Fit Model, enter your one Y measure. You now have three columns as independent variables. To enter them swiftly, highlight *CockpitUI*, *Month*, and *TimeOfDay* in Select Columns, and then choose Macros > Full Factorial. Then add *Pilot* by itself. Now highlight *Pilot* in Construct Model Effects and *CockpitUI* in Select Columns and click Nest[2] to produce *Pilot[CockpitUI]*. Finally, highlight this term in Construct Model Effects and use the Attributes red-arrow menu to choose Random Effect. The final built term should be *Pilot[CockpitUI]& Random*. Be sure the REML method is used with Standard Least Squares personality and Effect Leverage emphasis.)

Compare the results of the mixed-effects model analysis of variance with the traditional repeated measures ANOVA.

23. Are the same factors significant? With similar F- and *p*-values?

    For the repeated measures ANOVA, the *CockpitUI*, *Month*, *CockpitUI*Month*, and *TimeOfDay* effects were significant. In the mixed-effects model, these effects are also significant, and no other effects are significant. The table below shows the F- and *p*-values for both analyses.

[2] Nesting and nested effects are covered in module 7. The reason for nesting *Pilot* in the between-subjects factor *CockpitUI* is to avoid pooling pilots who used different cockpits. Nesting causes only pilots who used the same cockpit interfaces to be pooled. It turns out that because *Pilot* encodings are all unique for each subject, JMP produces the correct results in this case even if nesting is not used, *i.e.*, using a *Pilot& Random* term. However, if each pilot identifier had been assigned *within* each cockpit interface (*e.g.*, the first pilot in cockpits A, B, Control, and Placebo all get pilot identifier "1"), then nesting would be essential.

	repeated measures ANOVA				***mixed-effects model***			
	df$_{num}$	*df*$_{den}$	*F*	*p*	*df*$_{num}$	*df*$_{den}$	*F*	*p*
CockpitUI	3	16	274.97	<.0001	3	16	274.97	<.0001
Month	2	32	253.15	<.0001	2	80	622.88	<.0001
*CockpitUI*Month*	6	32	92.05	<.0001	6	80	226.49	<.0001
TimeOfDay	1	16	360.94	<.0001	1	80	13.70	.0004
*CockpitUI*TimeOfDay*	3	16	0.63	.60	3	80	0.02	.99
*Month*TimeOfDay*	2	32	1.59	.22	2	80	0.03	.97
*CockpitUI*Month*TimeOfDay*	6	32	1.50	.21	6	80	0.03	1.00

From the table, we can see that the significant effects are indeed the same in both analyses. However, the between-subjects main effect is the only one for which the F-ratio and *p*-values exactly match. The F-ratios for *Month* and *CockpitUI*Month* are larger in the mixed-effects model, but the *TimeOfDay* F-ratio is higher in the repeated measures ANOVA. For the nonsignificant effects, the mixed-effects model gives results further from significance than the repeated measures ANOVA judging by *p*-values.

24. How do the degrees-of-freedom compare?

From the table, we can see that the numerator degrees-of-freedom agree in all cases, but the denominator degrees-of-freedom do not. Mixed-effects models preserve greater denominator degrees-of-freedom than repeated measures ANOVAs, but compensate by using wider confidence intervals, making significance no easier to detect. Note that the mixed-effects model has the same denominator degrees-of-freedom for all effects except for the between-subjects factor *CockpitUI*. The repeated measures ANOVA, on the other hand, has varying denominator degrees-of-freedom based on the effects being tested.

25. Judging by the F- and *p*-values, with which analysis does it seem easier to achieve significance?

In general, with the exception of the *TimeOfDay* effect, it seems that the mixed-effect model makes significant effects stronger and nonsignificant effects weaker.

26. Write a few sentences describing the statistical analysis you just performed as you would for a publication.

A mixed-effects model analysis of variance was used with *Minutes* as the continuous response and *CockpitUI*, *Month*, and *TimeOfDay*, and their interactions, as nominal fixed effects. *Pilot* was modeled as a nominal random effect nested within *CockpitUI*.

Part 7, More than two within-subjects factors. If you have more than two repeated factors, JMP cannot perform a traditional repeated measures ANOVA because the Analyze > Fit Model > Manova > Choose Response > Compound procedure we used in part 3 cannot handle more than two factors. With 3+ repeated factors for a traditional repeated measures ANOVA, we can use SPSS, which can handle arbitrary numbers of within-subjects factors in a wide-format table.

Run SPSS and paste your original wide-format JMP data from *cockpits.jmp* into a spreadsheet. Name and set the column values properly in the Variable View, including using integers and value labels for *CockpitUI*, and then use Analyze > General Linear Model > Repeated Measures. (Hint: when you execute this command, you'll provide names for your within-subjects factors and their number of levels, and then map your columns to these. Don't forget to also add the between-subjects factor to your model.) Although the data in *cockpits.jmp* has only two repeated factors and not three, you will see easily how SPSS could be used to support any number of repeated factors.

Run the ANOVA in SPSS and explore the output report. In particular, you want to pay attention to the following sections in the SPSS output:

- *Within-Subjects Factors*. Verify that your within-subjects variables and levels are properly built.
- *Between-Subjects Factors*. Verify that your between-subjects variable(s) appear there with the proper number of levels and data points.
- *Mauchly's Test of Sphericity*. Compare to the Sphericity Test outputs from JMP.
- *Tests of Within-Subjects Effects*. This is where your main results will be—note that the Error(*x*) rows give your denominator degrees-of-freedom for factor *x*.
- *Tests of Between-Subjects Effects*. The Error line gives your denominator *df*s, and the line with the name of your effect gives your numerator *df*s.

Compare your SPSS repeated measures ANOVA output to that from JMP for the repeated measures ANOVA in part 3.

27. How do your results compare among these two reports?

The results are very similar but not exactly the same. The following table shows the results from repeated measures ANOVAs in JMP and SPSS.

			JMP		***SPSS***	
	df_{num}	df_{den}	*F*	*p*	*F*	*p*
CockpitUI	3	16	274.97	<.0001	274.98	.000
Month	2	32	253.15	<.0001	252.53	.000
*CockpitUI*Month*	6	32	92.05	<.0001	91.92	.000
TimeOfDay	1	16	360.94	<.0001	379.61	.000
*CockpitUI*TimeOfDay*	3	16	0.63	.60	0.90	.47
*Month*TimeOfDay*	2	32	1.59	.22	1.50	.24
*CockpitUI*Month*TimeOfDay*	6	32	1.50	.21	1.41	.24

28. Why might exact values not always match?

The exact values may vary because of different calculations being made under-the-hood by the two statistical packages. In general, JMP uses Type I sum of squares. By default, SPSS uses Type III sum of squares, but this can be set by the user. After changing SPSS to use Type I sum of squares in the Model dialog, however, the above results are unchanged, so this issue does *not* account for the difference here. Instead, perhaps these discrepancies are due to accumulated rounding and truncating differences.

29. Do the study conclusions change for any of these analyses?

No, the conclusions of our statistical tests, and therefore the findings from this study, do not change whether we use a repeated measures ANOVA in JMP or in SPSS.

06. Transformations

debugging.jmp, *errors.jmp*, *errorsrare.jmp*

In this module, we will explore the use of common transformations on data for the purposes of making them amenable to analyses of variance. We will focus on the common log$_e$ transform, but a reading included with these modules discusses a wider range of possible transformations than you'll be exposed to here.

Part 1, Examine the data table. Open *debugging.jmp* and examine it. Before going further, answer the following orientation questions:

1. Is this table in wide-format or long-format? How do you know?

   The *debugging.jmp* table is in long-format. We can tell this because each subject is listed on two rows, and each row contains just one measure per subject.

2. How many subjects participated in this experiment?

   There were 20 subjects in this experiment.

3. What are the factor(s) in this experiment? For each factor, say what variable type it is and whether it is between-subjects or within-subjects.

   There are two factors in this experiment, *Language* and *IDE*. These are both nominal. *Language* is a between-subjects factor, while *IDE* is a within-subjects factor.

4. What is the response in this experiment? What variable type does it have?

   The response in this experiment is *Seconds*. It is continuous.

5. What type of analysis is warranted based on the factor and response variable types?

   Based on the nominal factor inputs and continuous response output, an analysis of variance, whether using a traditional ANOVA or a mixed-effects model, is warranted. Note that we require a repeated measures analysis because *IDE* is a within-subjects factor.

6. Write up a plausible explanation of what this study was examining.

   This study appears to have been examining the time it took for programmers to debug a program given to them in a specific language, either C# or Java, and using, in turn, two separate "integrated development environments" (IDEs). Carryover effects would be problematic if the program debugged was the same in each IDE, so perhaps two or more programs were in play, and each subject was assigned a program randomly without replacement from among these. A more complete version of this data table should include a column called *Program* as a covariate to enable us to examine whether the programs themselves affected *Seconds*.

Part 2, Explore the distributions. Use the Analyze > Distribution menu to explore this data. First explore it as a whole, and then explore it using the By box to visually compare the two factors to each other. (Hint: use Uniform Scaling and Stack to make visual comparisons easier.)

7. Describe what you see on the overall data. Does it look normally distributed to you?

   The overall data does not look normally distributed. Rather, its most frequent values are at its left end, and it shows a long tail of less frequent values to the right.

8. When looking at two plots by *IDE*, what do you see? Do you think there is an effect of *IDE* on *Seconds*?

   When examining the data distributed by *IDE*, we see that Visual Studio has a longer tail than Eclipse. It appears that there is an effect of *IDE* on *Seconds* given this difference.

9. When looking at two plots by *Language*, what do you see? Do you think there is an effect of *Language* on *Seconds*?

   When examining the data distributed by *Language*, we see that although Java has a slightly longer tail than C#, the distributions look quite similar. It does not appear that there is an effect of *Language* on *Seconds*.

Part 3, Test for normality. Again use Analyze > Distribution and examine the whole data set. Use the red-arrow menu on Seconds and choose Fit Distribution > Normal.

10. A normal distribution, based on a Gaussian function, is fully parameterized by two parameters. In general, what are they? And what values do they show for this fitted curve?

    The two parameters necessary to define a normal distribution are its mean and its standard deviation. These are also called its location (μ) and dispersion (σ). For our fitted normal, our mean is μ = 3984.50 seconds and standard deviation is σ = 3678.88 seconds.

Next, use the red-arrow menu on Fitted Normal and select Goodness of Fit. This causes JMP to perform a Shapiro-Wilk *W* test of normality.

11. Conceptually, how do you think the Shapiro-Wilk *W* test works? What is it doing? Report the results of the test, and say whether your *Seconds* data can be considered normal or not. Does it appear to violate the normality assumption of ANOVA?

    We can think of the Shapiro-Wilk *W* test as performing an analysis of variance between the observed data in our table and sample data generated from the fitted distribution with the given mean and standard deviation. If these data sets are significantly different, then we conclude that our observed distribution is non-normal. The Shapiro-Wilk *W* test gives W = .78 and p < .0001 for our data, which leads us to admit that our distribution is significantly different from normal. Thus, our data appears to violate the normality assumption of ANOVA.

Next, use the red-arrow menu on Seconds to fit a LogNormal distribution. If a distribution is normal under a logarithmic transform, it is said to be lognormal. Examine the Goodness of Fit.

12. What kind of test is performed? Report the results of the test.

    In testing the goodness-of-fit of a lognormal distribution, Kolmogorov's *D* test is performed. It shows a nonsignificant result (*D* = .09, *p* = .15).

13. According to the test, can *Seconds* be considered lognormal?

    The test is nonsignificant; therefore we retain the null hypothesis that our data is from the lognormal distribution. So yes, we can consider *Seconds* to be lognormal.

Part 4, Apply a log transform. Return to your data table and create a new column. Name it log(Seconds) and enter the Formula editor. Under Functions (grouped), choose Transcendental > Log. Under Table Columns, choose Seconds. (You should now see Log(Seconds) as the formula for this column. Note that Log by itself uses base *e*, which is typical by convention. If you wanted to supply a different base, you would press the comma (,) key after choosing Log and a subscript would appear.) Click OK. You should now see your new column populated.

14. Use Analyze > Distribution on your newly created column and perform the same two tests for normality that we did in part 3. Report the test results and conclusions. Be sure to name each test appropriately.

    After log-transforming our *Seconds* data into *log(Seconds)* and fitting a normal distribution to it, a Shapiro-Wilk *W* test indicates that the transformed data can be considered normal (*W* = .97, *p* = .39). The fitted normal has mean μ = 7.92 and standard deviation σ = 0.93. After fitting a lognormal distribution to *log(Seconds)*, a Kolmogorov's *D* test indicates that the data can be considered lognormal (*D* = .10, *p* = .15) (*i.e.*, the original *Seconds* data can be log-transformed twice and still be considered normal). The fitted lognormal has mean μ = 2.06 and standard deviation σ = 0.12.

15. When time is a dependent variable, especially times of many seconds or minutes, it often exhibits a lognormal distribution. Think about the nature of time and human performance, and give a brief explanation of why time is often lognormal.

    Time as a dependent variable is often lognormally distributed because it is difficult for subjects to perform a task faster than some threshold, but subjects can always go arbitrarily slower, giving rise to a long tail and a lognormal distribution.

Part 5, Perform the analysis. Perform two analyses of variance, one on your original response variable, and one on your newly transformed response. The analysis you perform is up to you. (Your choices are either a traditional fixed-effects ANOVA or a mixed-effects model analysis of variance. Be sure to describe the analysis you performed and the model you used.)

16. What are the conclusions of these analyses?

    For completeness, we will first perform a traditional repeated measures ANOVA on *Seconds* and then a mixed-effects model analysis of variance on *Seconds*. We will follow these two analyses with two parallel analyses on *log(Seconds)*.

    First, we create a wide-format table with *Subject*, *Language*, *Eclipse*, and *VisualStudio* as columns. *Language* is a between-subjects factor with two levels, C# and Java. *IDE* is an implicit within-subjects factor encoded in the continuous response columns Eclipse and VisualStudio. We use Analyze > Fit Model > Manova > Choose Response > Repeated Measures as covered in module 5. Our results show a significant main effect of *IDE* (F(1,18) = 7.43, *p* < .05), a nonsignificant main effect of *Language* (F(1,18) = 1.72, *n.s.*), and a nonsignificant *IDE*Language* interaction (F(1,18) = 1.65, *n.s.*).

Second, we return to our long-format table and use a mixed-effects model analysis of variance with model effects for *Language*, *IDE*, *Language*IDE*, and *Subject[Language]& Random*. Using Analyze > Fit Model and the REML method we see a poor fit of the model (R^2=-.19), hinting at the unsuitability of these data for analysis of variance. We also see that our within-subjects effects, *IDE* and *Language*IDE*, both show 1 denominator degree-of-freedom, which suggests a problem. Thus, we pronounce this analysis flawed and disregard it. Fortunately, we know that *Seconds* should be log-transformed before analyzing it!

Third, we run our traditional repeated measures ANOVA on *log(Seconds)*. Our results again show a significant effect of *IDE* (F(1,18) = 7.97, *p* < .05), a nonsignificant effect of *Language* (F(1,18) = 1.93, *n.s.*), and a nonsignificant *IDE*Language* interaction (F(1,18) = 0.10, *n.s.*).

Fourth, we run our mixed-effects model analysis of variance on *log(Seconds)*. Our results show a much better model fit (R^2=-.87) than for *Seconds*. We also see identical outcomes as with our repeated measures ANOVA. There is a significant effect of *IDE* (F(1,18) = 7.97, *p* < .05), a nonsignificant effect of *Language* (F(1,18) = 1.93, *n.s.*), and a nonsignificant *IDE*Language* interaction (F(1,18) = 0.10, *n.s.*).

17. How do their findings compare?

The traditional repeated measures ANOVA yielded the same conclusions when analyzing *Seconds* as *log(Seconds)*. However, examining the F-ratios shows some differences, such as the 1.65 *vs.* 0.10 for the *IDE*Language* interaction. This suggests that perhaps the analysis of *Seconds* is more prone to a Type I error than the analysis of *log(Seconds)*. The mixed-effects model analysis of variance was disregarded for *Seconds* but yielded proper results for *log(Seconds)*, results that agreed exactly with the repeated measures ANOVA on *log(Seconds)*.

18. Write up the description and the results for this study as you would for a publication. One long paragraph is sufficient.

Twenty C# and Java programmers were recruited to debug two similar programs using two popular “integrated development environments” (IDEs), Eclipse and Visual Studio. Programs were assigned randomly to each programmer and to each IDE. Half of the programmers were proficient in C#; the other half were proficient in Java. This study was a 2×2 mixed design with a between-subjects factor of *Language* (C#, Java) and a within-subjects factor of *IDE* (Eclipse, Visual Studio). The outcome measure was the time it took for each programmer to successfully debug each program, measured in seconds. This temporal response was found to violate normality (Shapiro-Wilk *W* = .78, *p* < .0001) but was sufficiently lognormal (Kolmogorov’s *D* = .09, *p* = .15) and so was log-transformed before analysis. A repeated measures ANOVA showed a significant main effect of *IDE* on *log(Seconds)* (F(1,18) = 7.97, *p* < .05). An analysis of means shows that Eclipse resulted in faster debugging times than Visual Studio (2406.2 *vs.* 5562.8 seconds). There were no significant effects of *Language* (F(1,18) = 1.93, *n.s.*) or of *IDE*Language* (F(1,18) = 0.10, *n.s.*) on *log(Seconds)*.

Part 6, Using a constant. Next, open the file *errors.jmp* and examine the data table.

19. What does this experiment seem to have been about?

This experiment seems to have recruited 10 subjects to point at on-screen targets, perhaps using a novel input device or cursor. Targets were placed at a certain *Distance* and had a certain *Size*, and the pointing errors for targets in each *Distance* × *Size* condition were recorded. *Total Errors* per condition was therefore the continuous response variable. *Distance* and *Size* were both continuous within-subjects factors.[♣] *Distance* had levels 32, 64, 96, and 128 pixels. *Size* had levels 8 and 16 pixels. The study was therefore a 4×2 within-subjects design.

20. What does the dependent variable seem to measure? Does it seem reasonable that it might be lognormally distributed? Why or why not?

The dependent variable seems to measure the number of pointing errors per subject in each *Distance* × *Size* condition. We do not know from the data table how many pointing trials were administered in each *Distance* × *Size* condition. From the data, it seems reasonable that perhaps 100 pointing trials were administered in each condition, so *Total Errors* would measure the number of trials out of 100 that were misses. It seems reasonable that these data may be lognormally distributed because it may be difficult to perform close to error-free, but possible to make an arbitrarily high number of errors, resulting in a long tail at high error counts.

21. Examine the distribution of *Total Errors* and fit normal and lognormal distributions to it. What happens? Why? Write up what you find from analyzing the distribution.

The overall distribution of *Total Errors* appears lognormal with frequent values near zero and increasingly less frequent values forming a long tail to the right. A fitted normal results in a distribution with mean μ = 11.90 and standard deviation σ = 14.57. A significant Shapiro-Wilk W test indicates that *Total Errors* cannot be assumed to come from a normal distribution (W = .68, p < .0001). Attempting to fit a lognormal distribution produces an error: "Lognormal Distribution not available for non-positive data." This error occurs because log(0) is undefined. In other words, we cannot perform a log-transform in the presence of zeroes.

22. How could you remedy your inability to fit a lognormal distribution to the data?

We could add +1 to each of our error counts before taking the log, thereby removing zeroes from the data. Note that log(1) = 0, so what *were* zeroes will once again become zeroes after the log-transform is applied. Thus, the frequency of zeroes in the log-transformed data will match exactly the frequency of zeroes in the original data.

23. Attempt to create a log-transformed column for your data. What happens? Why? Again, what could you do to make this work? (Write your response before proceeding.)

If we simply add a formula column called *log(Total Errors)* defined as such, we get a column that contains missing values, which appear as bullet characters (•). We could instead create a new formula column called *log(Total Errors + 1)* defined as such.

[♣] It is defensible to model *Distance* and *Size* as either continuous, ordinal, or nominal factors. For starters, conducting a comparable fixed-effects repeated measures ANOVA using a wide-format table implicitly treats *Distance* and *Size* as nominal. In a mixed-effects model analysis of variance, we have additional options for continuous or ordinal. A continuous modeling type makes sense because we know pixel distances are not just categories, or even ordered categories with unknown gaps between levels, but scalar values with measurable gaps between levels.

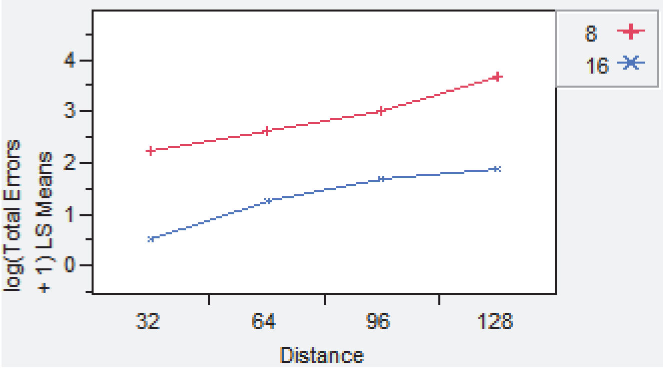
One option is to change from doing $y' = \log_e(y)$ to instead doing $y' = \log_e(y + 1)$. This can be generalized to $y' = \log_e(y + c)$, for some constant c. It is often the case that $c = 1$ suffices, although statisticians have more complex approaches to determining c based on particular data (*see* Berry 1987).

24. Using $c = 1$, apply the log-transformation. Then re-examine the distribution using Analyze > Distribution. Does it work now? Perform a normal fitness test and report the results.

We create a new formula column called *log(Total Errors + 1)* defined as such. Then we provide this column to Analyze > Distribution. The distribution visually looks quite normal. A fitted normal gives a distribution with mean μ = 2.11 and standard deviation σ = 0.97. A nonsignificant Shapiro-Wilk *W* test indicates that we can assume these data are from a normal distribution (*W* = .99, *p* = .54). Since our data was log-transformed before fitting this normal, this result effectively tells us that *Total Errors* was lognormally distributed.

25. Perform a mixed-effects model analysis of variance on both the original response and your newly transformed response. As you did in part 5, write up the experiment, procedure, analysis, and findings as you would for a publication (*see* question 18).

Ten subjects were recruited to point to on-screen targets using a new optical mouse. (Data was previously collected in a similar study for an old roller-ball mouse.) Each subject performed 100 pointing trials for a given combination of target distance and target size. With four distances and two sizes, this study was a 4×2 within-subjects design. *Distance* was a continuous within-subjects factor with levels of 32, 64, 96, and 128 pixels; *Size* was a continuous within-subjects factor with levels of 8 and 16 pixels. The outcome measure, *Total Errors*, was the number of pointing errors occurring in each 100-trial set. This response was found to violate normality (Shapiro-Wilk *W* = .68, *p* < .0001) but was sufficiently normal after a $y' = \log_e(y + 1)$ transform (*W* = .99, *p* = .54). (The +1 was necessary due to the presence of zeroes in the response.) A mixed-effects model with response *Total Errors* yields R^2=.72 and shows significant effects for *Distance* (F(1,67) = 57.17, *p* < .0001), *Size* (F(1,67) = 89.92, *p* < .0001), and *Distance***Size* (F(1,67) = 30.35, *p* < .0001). A mixed-effects model analysis of variance with response *log(Total Errors + 1)* yields R^2=.96 and shows significant effects for *Distance* (F(1,67) = 559.20, *p* < .0001), *Size* (F(1,67) = 1232.12, *p* < .0001), but not for the *Distance***Size* interaction (F(1,67) = 0.02, *n.s.*). Given that we know *Total Errors* is lognormally distributed and given that the R^2 model fit is much better for the *log(Total Errors + 1)* response, the interaction should be considered nonsignificant. Indeed, an examination of means shows that more distant and smaller targets resulted in more errors, but about equally for both target sizes, suggesting no interaction.



Part 7, Challenges. Now open *errorsrare.jmp* and examine the data table. Pay special attention to the frequency (or lack thereof) of errors in the distributions you examine.

26. What problems can you anticipate with applying a log-transform (or, frankly, most any other transform)?

    The distribution of *Total Errors* in *errorsrare.jmp* shows that there are a great many zeroes in the data. Although we can apply a log-transform of the kind $y' = \log_e(y + 1)$, this will simply reproduce these zeroes, and thus not change our distribution much; it will most certainly not make it normal. In the end, a log-transform can change the spread of values (*i.e.*, the distance between them), but not their frequency. If there are a great many raw *n*'s in the data, then there will simply be a great many $\log_e(n)$'s in the transformed data.

27. Apply a workable log-transform as you did previously. Re-examine the distribution of data using Analyze > Distribution. Did the log-transform succeed in normalizing the data? Why or why not?

    We create a new formula column named and defined as *log(Total Errors + 1)*. Analyze > Distribution shows that the spread of values has changed but not the frequency of zeroes and what used to be ones (now $\log_e(1 + 1) = .693$). As we stated above, the log-transform affects the spread of values (*i.e.*, their relative distance from one another), but not their frequency. Many raw zeroes become many zeroes again under $y' = \log_e(y + 1)$.

In reality, these error data are not amenable to ANVOA because of the high frequency of zeroes. Even using a log-transform, one cannot change the frequency of outcomes, just their distance from one another. We will see in module 9 how to use nonparametric techniques to analyze data such as these.

28. Despite its inappropriateness, use a mixed-effects model analysis of variance on both your raw outcome and your log-transformed outcome with Analyze > Fit Model and report your findings.

    An analysis for *Total Errors* shows a main effect of *Distance* ($F(1,67) = 64.75$, $p < .0001$), a main effect of *Size* ($F(1,67) = 84.15$, $p < .0001$), and a *Distance*Size* interaction ($F(1,67) = 40.83$, $p < .0001$). The R^2 of this model on the raw data is .73. An analysis for *log(Total Errors + 1)* shows a main effect of *Distance* ($F(1,67) = 102.75$, $p < .0001$), a main effect of *Size* ($F(1,67) = 251.33$, $p < .0001$), and a *Distance*Size* interaction ($F(1,67) = 25.11$, $p < .0001$). The R^2 of this model on the raw data is .84.

Another challenge of using a transformation is *interpreting* your results. In research write-ups, it is essential to be clear when reporting results based on a transformed response variable. The more intricate the transform, the harder it is to straightforwardly interpret the results.

29. Look back at your answers in previous parts of this module. Were you careful to indicate that you were interpreting findings from a log-transformed response? Or did you write your answers just as you would have had you not transformed your response?

    Indeed, it is important when we say factor *A* had a significant effect on outcome *B*, that if *B* was log-transformed, we state that factor *A* had a significant effect on $\log_e(B)$. If we make comparisons among means of *B*, or produce graphs of *B* by *A*, and yet the analysis was conducted on $\log_e(B)$, we need to make this clear. To be safe, we might only compare log-transformed means and depict log-transformed values on our graphs if our entire analysis hinges on having log-transformed *B* in the first place. However, such graphs are harder for readers to appreciate.

Part 8, Other transforms. Acquire the Hoyle (1973) reading and familiarize yourself with the 19 transformations it reviews in section 4. You'll see that some of them are quite involved.

30. Name the three transformations you think are most relevant to human-computer interaction data. Why those three? What kind of data would warrant their use?

In human-computer interaction, the most common transformation is Hoyle's #4.3, the logarithmic transformation, defined either as $y' = \log_e(y)$ or $y' = \log_e(y + 1)$. This transform changes the spread of values such that long tails get "pulled in" and lognormal distributions become normal. A second relevant transformation is Hoyle's #4.17, normal scores, which deal with assigning ranks to data and then performing analyses on ranks. This results in a form of nonparametric analysis covered in module 9. Another third relevant transformation is Hoyle's #4.15, logit transformation, which is used in the analysis of dichotomous outcomes (0/1) in binomial regression. Binomial regression is covered in module 10.

07. Nesting

designers.jmp, *higgins.jmp*

Nesting is an issue that arises in many types of experiments in which a hierarchical or containment structure is used to organize factors. It also arises in mixed factorial designs when long-format tables are used, as experimental units (Subjects) must be nested in between-subjects factors. With nesting, we prevent the introduction of unwanted variation by pooling nested factors only within levels of nesting factors, rather than across levels. Before beginning the exercises, it is important to give you an intuition regarding nested effects so that you can recognize them when they should be built. (A rule of thumb is to be on the lookout whenever you have between-subjects factors. As stated, nested effects occur in mixed factorial designs that have both between- and within-subjects factors, although they can occur for other reasons, *e.g.*, with covariates.)

Part 1, Text entry example. As a fictitious example, consider a study to examine the effects of bodily posture on mobile text entry performance. A researcher examined the text entry performance of subjects while they were sitting, standing, walking, and riding on a bus. Each subject assumed each of the postures in the study in a random order, but because of the learning necessary, each subject used only one of the following text entry methods: *mini-Qwerty*, *soft-Qwerty*, *Graffiti*, *EdgeWrite*, or *ShapeWriter*. As the difference in performance of these methods is already well known, the research question concerns the posture effect, not the relative performance of the methods, *per se*.

1. What are the between-subjects and/or within-subjects factors for the above study?

   The between-subjects factor is text entry *Method*. The within-subjects factor is body *Posture*.

2. How many levels does each factor have? What are those levels?

   Method has five levels: mini-Qwerty, soft-Qwerty, Graffiti, EdgeWrite, and ShapeWriter. *Posture* has four levels: sitting, standing, walking, and riding.

3. What would be a reasonable response variable? *Record your answers before proceeding.*

   A reasonable response for a text entry experiment would be words per minute (wpm). Another reasonable response would be text entry error rate.

Think in terms of the variance of the response variable. The use of five separate text entry methods introduces a lot of variance into the data. ShapeWriter, for example, can reach 40 wpm or higher while Graffiti may only reach 20 wpm. What to do about this? On one hand, the researcher could decide to simply use *one* text entry method for the whole experiment. But this reduces the generalizability of the results: the effects of posture, then, could be said to only apply to the method tested, and not to mobile text entry in general. So it would be nice to use a selection of methods to increase generalizability.

To account for this unwanted variance, we must nest *Subject* within *Method*, written (counter-intuitively for computer programmers) as *Subject[Method]*. (In general, we write factor *A* is nested in factor *B* as *A[B]*. *A* is said to be the nested factor, and *B* is said to be the nesting factor.)

Mathematically, nesting causes the levels of the nested factor (A) to be modeled *separately* for each level of the nesting factor (B), rather than *across* levels of the nesting factor. In the example above, each subject will be grouped based on whatever *Method* they used. So the subjects that use ShapeWriter will be modeled together, but separately from the subjects that use Graffiti, and so on.

The explanation of nesting from the SAS JMP 7.0 help documentation reads as follows, with letters *A* and *B* from our *A[B]* notation added for clarity:

> JMP uses the terms inside the parentheses (B) as grouping terms for each group (A). For each combination of levels of the nesting terms (B), JMP constructs the effect on the outside of the parentheses (A).

The intuition for when to nest can be thought of in terms of the meaning of the nested effect. If the levels of a factor (*e.g.*, *Subject*) differ in their experimental "meaning" or "interpretation" according to the levels of another factor (*e.g.*, *Method*), then the first should be nested in the second (*e.g.*, *Subject[Method]*). In the example above, although each subject is, in some sense, equivalent to each other subject, the activity that a subject does in the experiment is entirely different based on the text entry method they are assigned. Nesting *Subject[Method]* ensures that each subject is modeled according to the text entry method that they used, allowing us to better isolate the effects of posture. (Note that we must still make *Subject* a random effect; in JMP it will look like "Subject[Method]& Random" in the model effects box.)

Part 2, Agriculture example. Mixed designs are sometimes called *split-plot designs* due to their common use in agriculture. Consider an experiment from Milliken & Johnson (1984) reported in Higgins *et al.* (1990) to test the effects of fertilizer and moisture levels on the dry matter resulting from peat moss. See if you can spot the experimental unit (the "subject"), the between-subjects factor, and the within-subjects factor from Higgins' description:

> In the experiment, four peat pots, with a different level of fertilizer randomly assigned to each, were placed in a tray. Each tray was treated with one of four different moisture levels, the levels being randomly assigned to trays. The response variable was amount of dry matter (Higgins *et al.* 1990, p. 189).

I'll clarify that in this experiment, there were 12 trays, 4 moisture levels, and 4 types of fertilizer.

4. What is the experimental unit in this study? How many of these are there?

   In this study, the experimental unit, equivalent to the "subject," is *Tray*, *i.e.*, the 12 separate trays in the study.

5. What is the between-subjects factor? What are its levels?

   The between-subjects factor is *Moisture*, which has four levels, presumably something like low (1) to high (4) but this information is not given in the description.

6. What is the within-subjects factor? What are its levels?

   The within-subjects factor is *Fertilizer*, which has four levels. It is unclear whether this is four different quantities of one fertilizer, or four different fertilizer brands administered in equal amounts.

7. What constitutes a single "trial?" Indicate how a trial can be defined in any experiment.

   A trial is defined as the action or item on which a single measure is taken. It is also that which produces a single row in a long-format data table. In this experiment, a trial is each of the four peat pots on a tray. Each tray, being a "subject," can be said to have "performed" four trials via each peat pot. The amount of dry matter in each pot is the response for each trial.

Part 3, Education example. Other examples of nesting occur in educational research, where different sites are used. Consider a fictitious study of high school students' performance on a test. If five different high schools are used, then we will probably want to nest *Student[School]* to reflect that each student came from a different high school. Nesting prevents the pooling of students across schools.

Part 4, Manufacturing example. Wikipedia's entry on *Restricted randomization* has an informative section called *Example of nested data* (http://en.wikipedia.org/wiki/Split-plot_design). Here is its description of a fictitious manufacturing experiment that has *two* nested effects. Note at the end where it also mentions the use of *random effects*, with which you should be already familiar.

> Consider a batch process that uses 7 monitor wafers in each run. The plan further calls for measuring a response variable on each wafer at each of 9 sites. The organization of the sampling plan has a hierarchical or nested structure: the batch run is the topmost level, the second level is an individual wafer, and the third level is the site on the wafer.
>
> The total amount of data generated per batch run will be 7 * 9 = 63 observations. One approach to analyzing these data would be to compute the mean of all these points as well as their standard deviation and use those results as responses for each run.
>
> Analyzing the data as suggested above is not absolutely incorrect, but doing so loses information that one might otherwise obtain. For example, site 1 on wafer 1 is physically different from site 1 on wafer 2 or on any other wafer. The same is true for any of the sites on any of the wafers. Similarly, wafer 1 in run 1 is physically different from wafer 1 in run 2, and so on. To describe this situation one says that sites are nested within wafers while wafers are nested within runs.
>
> As a consequence of this nesting, there are restrictions on the randomization that can occur in the experiment. This kind of restricted randomization always produces nested sources of variation.
>
> The objective of an experiment with this type of sampling plan is generally to reduce the variability due to sites on the wafers and wafers within runs (or batches) in the process. The sites on the wafers and the wafers within a batch become sources of unwanted variation and an investigator seeks to make the system robust to those sources.

> Because the wafers and the sites represent unwanted sources of variation and because one of the objectives is to reduce the process sensitivity to these sources of variation, treating wafers and sites as random effects in the analysis of the data is a reasonable approach. In other words, nested variation is often another way of saying nested random effects or nested sources of noise. (Wikipedia, accessed June 1, 2010)

Part 5, Study of designers. Open the JMP sample data file *designers.jmp*. This data table is equivalent to JMP sample file *animals.jmp*, recast to serve as an example relevant to human-computer interaction. Examine the table notes and the columns carefully. Then answer:

8. How many designers took part in this study?

   Six individual designers took part in this study. Three designers were females, and three designers were males.

9. What are the between- and within-subjects factors in this study? What and how many are their levels?

   The between-subjects factor in this study is *Gender*, with two levels: F(emale) and M(ale). The within-subjects factor is *Season*, with four levels: fall, winter, spring, and summer.

10. What is the response variable in this study?

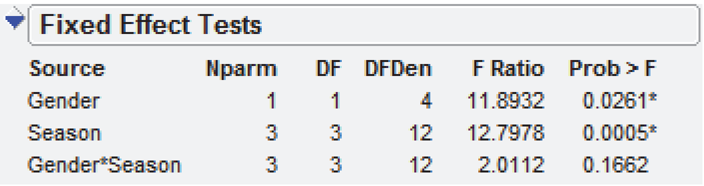
    The response in this study is a continuous numeric variable, *Breakthroughs*.

11. Give a plausible explanation for what this study may have been about.

    This fictitious study is of UX designers and their self-reported breakthroughs across seasons while working for Microsoft in the greater Seattle area. The researchers' interest may have been to see whether designers feel they experience different numbers of breakthroughs in their work based on the time of year by season.

12. Now perform the proper analysis on this data. You should use Analyze > Fit Model and use a mixed-effects model analysis of variance. (Hint: nesting is required.) Reproduce your model effects, your response variable, and your results table. (You can either retype it or clip out the relevant window regions from JMP. If you are on a Windows machine, you can use Start Menu > Accessories > Snipping Tool.)

    The model effects are *Gender*, *Season*, *Gender*Season*, and *Designer[Gender]& Random*. The response variable is *Breakthroughs*. The results table is shown below:



Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Gender	1	1	4	11.8932	0.0261*
Season	3	3	12	12.7978	0.0005*
Gender*Season	3	3	12	2.0112	0.1662

13. Which factor had to be nested? In which other factor? Why?

    Designer was nested in *Gender*, which makes our analysis robust to unwanted variance by modeling females separately from males. *Designer* was also a random effect, as is usual for our subjects variable, and so our final model effect appears as "Designer[Gender]& Random."

14. Look at the levels of *Designer* in the data table. Why does it make intuitive sense that nesting was necessary?

    The levels of *Designer* are just nominal labels 1, 2, and 3, and these are reused within each type of *Gender*. If we only had *Designer* in the model and not *Designer[Gender]*, the model would conflate, for example, the first female with the first male, the second female with the second male, and so on, folding these designers together as if they were the same people. By nesting *Designer* in *Gender*, we effectively model subjects as the combination of both terms, *e.g.*, F.1, M.1, F.2, M.2, and so on. We could avoid this issue by not reusing the same subject labels within different genders, *i.e.*, by making all levels of *Designer* unique. But regardless, we should still nest *Designer* in *Gender* to not be dependent on our factor encodings.

15. Based on the premise for this study and your findings, what can you conclude?

    Our mixed-effects model analysis of variance indicates that both the time of year (F(3,12) = 12.80, *p* < .001) and whether the designer was female or male (F(1,4) = 11.89, *p* < .05) had a significant effect on number of self-reported breakthroughs. Generally, self-reported breakthroughs increased in spring and summer, but were lower in fall and winter, perhaps due to feelings of optimism or increased productivity from higher energy. Males reported more breakthroughs than females, perhaps due to a higher desire to report or a lower threshold for declaring a "breakthrough." There was no significant interaction between these factors, meaning that the time of year proportionally affected the self-reported breakthroughs of females and males.

Part 6, Peat moss exercise. Now open *higgins.jmp*, which corresponds to the data from the peat moss experiment in part 2, above. Based on the description in part 2, and now in examining the data table, answer the following:

16. Is this a wide-format or long-format table? How do you know?

    The *higgins.jmp* data table is in wide-format. Each of 12 trays, the "subjects," was measured four times, once for each fertilizer (type or amount, we cannot be sure). Each tray was kept at a single moisture level. All of the measures for a tray appear in a single row in the table, and the *Fertilizer* factor is encoded across the four response columns.

17. What are the factors and levels? (Do they match your answers given in questions 5 and 6, above? If not, why not?)

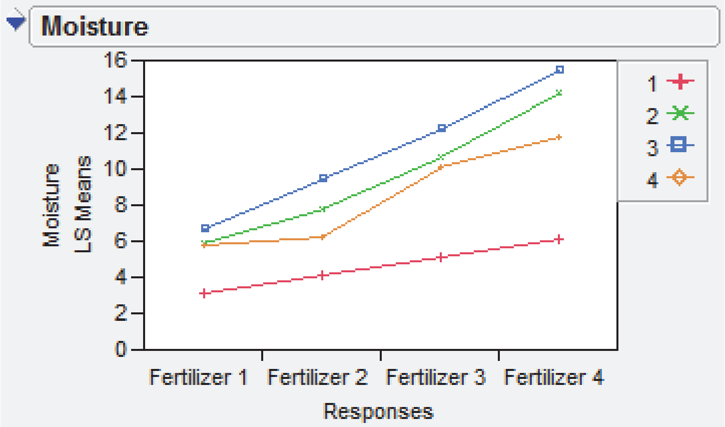
    Moisture is a between-subjects factor with levels 1-4, and *Fertilizer* is a within-subjects factor with levels 1-4. Indeed, these answers match those given in questions 5 and 6, above.

18. Recall from module 4, part 7 that you analyzed a two-way mixed factorial design with a traditional repeated measures ANOVA. Apply that same procedure here to analyze *higgins.jmp*. (Remember to consider Sphericity; if there is a violation, report the test and use an adjustment.) Write up the findings from this experiment as you would for a publication. First introduce the experiment in your own words.

Twelve trays each containing four peat pots were each subjected to one of four moisture levels while each pot was given a different level of fertilizer. The measure was the amount of dry matter produced in each pot after a given amount of time. A repeated measures ANOVA shows that *Moisture* exhibited a significant main effect on dry matter (F(3,8) = 26.30, *p* < .001), with moisture level 3 producing the most dry matter at 10.92 units (*SD* = 3.54), followed by moisture level 2 at 9.64 units (*SD* = 3.32), level 4 at 8.47 units (*SD* = 3.01), and level 1 at 4.58 units (*SD* = 1.44).

Fertilizer also exhibited a significant main effect on dry matter (F(3,24) = 133.16, *p* < .0001), with fertilizer level 4 producing the most dry matter at 11.85 units (*SD* = 3.85), followed by fertilizer level 3 at 9.53 units (*SD* = 2.88), level 2 at 6.88 units (*SD* = 2.41), and level 1 at 5.34 units (*SD* = 1.79).

These main effects must be interpreted with caution, however, as there was a significant *Moisture*Fertilizer* interaction (F(9,24) = 5.74, *p* < .001). The graph below captures this interaction by showing how each fertilizer amount from 1 to 4 magnified the effect of each moisture level on the amount of dry matter produced.



19. Note that as in module 4, part 7, explicit nesting was not an issue; it is also not (yet) relevant to this analysis. Speculate as to why not. (Hint: the data itself warrants nesting, but a traditional repeated measures ANOVA makes explicit nesting unnecessary because of the inherent structure of the wide-format data table. Why?)

    There was no need for nesting here because the nested factor would be, in this case, *Tray[Moisture]*, since each tray was assigned to only one level of moisture and we wish to model the trays in each moisture level apart from trays in other moisture levels. In a traditional repeated measures ANOVA, the experimental unit (*i.e.*, *Tray*) is not a model effect; rather, one row in the wide-format data table contains all measures taken on one experimental unit, in this case, a tray.

After you have analyzed *higgins.jmp* using a traditional repeated measures ANOVA, transform the *higgins.jmp* data table into long-format. You can use Tables > Stack or Microsoft Excel's Paste Special > Transpose command. When you are done, be sure to compute means and standard deviations over both tables to verify that your wide-format to long-format transformation was without error.

20. How many columns does your long-format table have? What are their names?

    The four columns in the long-format data table are *Tray*, *Moisture*, *Fertilizer*, and *Dry Matter*.

21. How many rows?

    There are 48 rows in the long-format data table. This corresponds to the measures from four fertilizers on 12 trays.

22. Do means and standard deviations agree between your wide-format and long-format tables?

    The means and standard deviations do indeed agree between the wide-format and long-format versions of *higgins.jmp*. They were confirmed using Analyze > Distribution and Tables > Summary.

Next, conduct a mixed-effects model analysis of variance on the long-format table using Analyze > Fit Model.

23. What are your model effects?

    The model effects are Moisture, Fertilizer, Moisture*Fertilizer, and Tray[Moisture]& Random.

24. What is your model's response?

    The response variable for this model is *Dry Matter*.

25. What is your random effect?

    The random effect is the experimental unit, *Tray*. Due to nesting, in the model effects box it is written as "Tray[Moisture]& Random."

26. You should be nesting one factor within another: which are you nesting and why? Give the rationale. (Hint: remember that *Subject* is indeed a factor, just not usually one of interest.)

    Tray, the factor corresponding to the experimental unit, is nested within in *Moisture*. In the model, this is shown as "Tray[Moisture]& Random." We nest because each tray was assigned to only one level of moisture and we wish to model the trays in each moisture level separately from the trays in the other moisture levels to remove unwanted variance.

After you have run your model, inspect the results in light of what you obtained using your traditional repeated measures ANOVA on *higgins.jmp*.

27. Compare the results between your repeated measures ANOVA and your mixed-effects model analysis of variance. Do the results agree? Compare and contrast your outcomes.

    The table below shows the results of the two analyses. As one can quickly see, the results agree exactly for these data.

	repeated measures ANOVA				*linear mixed-effects model*			
	df_{num}	df_{den}	*F*	*p*	df_{num}	df_{den}	*F*	*p*
Moisture	3	8	26.30	.0002	3	8	26.30	.0002
Fertilizer	3	24	133.16	<.0001	3	24	133.16	<.0001
*Moisture*Fertilizer*	9	24	5.74	.0003	9	24	5.74	.0003

Next, we are no longer going to use a mixed-effects model to analyze our long-format table, but instead use a traditional ANOVA but still on the long-format table. To do this, simply remove the "&Random" attribute from your random effect that you had in the previous model. (Personality should still be Standard Least Squares and Emphasis should still be Effect Leverage.)

28. Reproduce your model effects here.

    The model effects are now Moisture, Fertilizer, Moisture*Fertilizer, and Tray[Moisture].

29. Report the results of this analysis.

    Moisture exhibited a main effect on *Dry Matter* (F(3,24) = 120.33, *p* < .0001), as did *Fertilizer* (F(3,24) = 133.16, *p* < .0001) and *Moisture*Fertilizer* (F(9,24) = 5.74, *p* < .001). There was also a significant effect of *Tray[Moisture]* (F(8,24) = 4.58, *p* < .01), indicating that different trays resulted in different amounts of dry matter, even within the same moisture levels.

30. How do the results for each factor and interaction compare to the mixed-effects model results? Do any of the results from question 29 seem incorrect? Which? How do you know?

    The results for the repeated factor *Fertilizer* and for the *Moisture*Fertilizer* interaction remain unchanged from the repeated measures ANOVA and the mixed-effects model analysis of variance. These results seem correct. However, the between-subjects factor *Moisture* has changed a great deal, moving from 8 denominator degrees-of-freedom to 24 and achieving a much stronger F-ratio (120.33 *vs.* 26.30). This seems incorrect in light of our previous results. By removing the random effect from *Tray[Moisture]*, it seems we have ruined the veracity of the *Moisture* effect.

Finally, remove the nesting such that your formerly nested effect is no longer nested. (Symbolically, if you had *A[B]* then just use *A* in the model for that term. You will still have *B* separately.)

31. Reproduce your model effects here.

    The model effects are now Moisture, Fertilizer, Moisture*Fertilizer, and Tray.

32. What happens when you run the model? (Hint: look at the top of the report for Singularity Details.) Why do you think this happens?

    When the model is run, a singularity is produced, which indicates an error. A singularity happens whenever there are combinations of factors' levels suggested by the model that do not actually occur in the data. By having *Tray* unnested and nonrandom in the model, we are requiring the data to allow for combinations of *Tray* and *Moisture* that do not exist. Specifically, only one level of *Moisture* exists for each level of *Tray*, and this causes a problem. (This problem is resolved when a still unnested *Tray* is made into a random effect because *Tray* is no longer treated as fixed; for safety, however, *Tray* should still be nested. It is only by virtue of not reusing any of the same identifiers for trays across different moistures that "Tray& Random" suffices here; *see* the answer to question 14.)

33. In general, how can a singularity problem be remedied?

    A singularity can be remedied by nesting properly. It may also be remedied by using random effects, but as noted, this is by virtue of using unique labels for the experimental units, and it is risky to rely on this. Instead, nest properly and make random effects where appropriate, nested or otherwise.

Part 7, Create your own experiment. Create a mock human-computer interaction experiment where nesting is warranted. Try to make the experiment somewhat plausible and interesting.

34. Write your experiment description.

    A human-computer interaction researcher wanted to study the effects of internet connection speed on chat behavior in massively multiplayer online role-playing games (MMORPGs). She recruited 40 subjects, each with one of four interaction connection speeds: dial-up modem, DSL, cable modem, or T1. She then had them play for two hours in each of three MMORPGs and logged the number of text characters sent via each game's text chat feature.

35. Identify any between-subjects factors, within-subjects factors, random effects, and their levels.

    The between-subjects factor is *Connection Speed* with four levels: dialup, DSL, cable, and T1. The within-subjects factor is *Game* with three levels, each corresponding to a specific game. *Subject* would be a random effect as usual.

36. Which factor(s) should be nested within which other factor(s)? Why?

    Subject should be nested within *Connection Speed*, which will appear in the model as "Subject[Connection Speed]& Random." We do not want to pool all subjects together, but instead, model them according to the connection speed they used.

Part 8, Nesting in mixed factorial designs. JMP provides an online sample to illustrate the equivalence of wide-format and long-format repeated measures analyses. The sample is *#30584: Analyzing Repeated Measures in JMP Software*, and is available at http://www.jmp.com/support/notes/30/584.html. It is also provided in PDF format as part of this module as `Sample 30584.pdf`. Open the sample and read it, paying special attention to the wide-format and long-format tables that are figures in the sample.

37. Which figure represents the wide-format data? Which figure represents the long-format data? For each, how do you know?

    In JMP sample #30584, Figure A represents the wide-format data, and Figure B represents the long-format data. In the former, there is only one row per subject ID, and three measures, *LogHist0*, *LogHist1*, and *LogHist3*, all in each row. In the latter, there is only one measure per row, and thus subject IDs must be repeated on three rows.

38. What are the fixed effects and their levels? Which are between-subjects effects and which are within-subjects effects?

    The fixed effects are *drug* and *Time*, visible in the long-format table in Figure B. The between-subjects effect is *drug* with levels morphine and trimeth. The within-subjects effect is *Time*, with levels LogHist0, LogHist1, and LogHist3.

39. What is the response measure? How is it encoded in the wide-format and long-format tables?

    The response measure is the blood concentration of histamine in the given subject, which happens to be a dog. In the wide-format table, this measure is encoded across three columns, *LogHist0*, *LogHist1*, and *LogHist3*. In the long-format table, this measure is encoded as a single column, *LogHist*.

40. In either table, which column identifies the experimental unit?

    The experimental unit is encoded as *ID*. Each dog in the study was given a unique *ID*.

41. In the analysis for the long-format table, what is the random effect?

    The random effect is *ID*.

42. In the analysis for the long-format table, which effect is nested? In which effect? Why?

    The nested effect is also *ID*, and it is nested in *drug* because we do not want to pool dogs across different drugs, but model them within each drug separately. As *ID* is a random effect, it is written in the model effects box as "ID[drug]& Random."

08. Longitudinal data

macbtwn.jmp, macwthn.jmp

Whether it is throwing a javelin or saying the alphabet backwards, most types of human performance improve with practice. The ability for people to enter text by typing on a QWERTY keyboard, for example, follows a well-known "learning curve," where performance, measured as words per minute (WPM), improves steadily over time. Studies where practice and learning play a crucial role are often best conducted *longitudinally*, meaning "over time." A typical practice is to have subjects attend sessions that are no closer than 2 hours apart, no farther than 48 hours apart, and without more than 2 on the same day. Such studies may have within- or between-subjects factors of interest, but all will have a factor named *Session* or *Day* or *Time* that codifies the *x*-axis in Figure 1, reproduced from MacKenzie & Zhang (1999).

Finding "the elusive crossover point" is important when practice and learning play a central role. This point tells us the time it takes for one technique's performance to match that of another, often more familiar, technique. Also, longitudinal data of the kind shown in Figure 1 allows us to fit learning curves modeled by the Power Law of Learning:

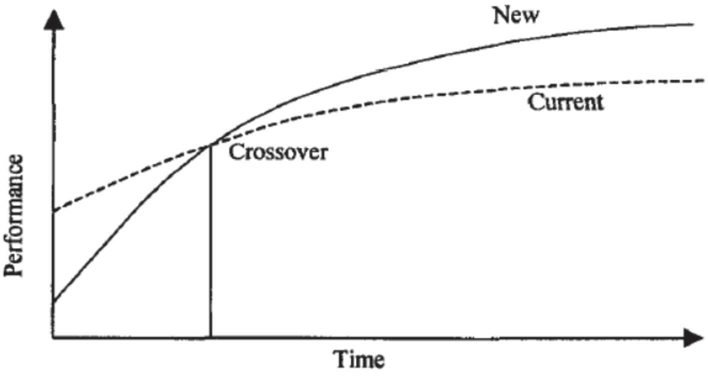


Figure 1. The elusive crossover point

Performance = a × Session^{b} ,

or, more generally,

Y = aX^{b} .

Here, the regression coefficient a is the y-intercept and represents initial performance. The other regression coefficient, b, is usually a value between 0.0 and 0.5, and reflects how steep the learning curve is. The response Y is the y-axis measure (*e.g.*, WPM), and the variable X is the temporal variable on the x-axis (*e.g.*, *Session*). As part of this module, we will see how to fit a learning curve to longitudinal data using the Solver in Microsoft Excel.

From a modeling standpoint, the temporal variable is always a within-subjects variable, as every subject is measured at every time point along the x-axis. Although the measures do not happen at the exact same times, relative to each subject, a given *Session* number is the same from one person to the next.

Questions of central interest are:

- How is longitudinal data analyzed statistically?
- Is it possible to statistically operationalize a notion of "when learning stops?" If so, how?
- How is a learning curve of the form Y = aX^{b} fitted?
- How can future performance be estimated based on a learning curve?

Part 1, Repeated measures ANOVA. Consider an example inspired by an experiment conducted by MacKenzie & Zhang (1999). The authors invented an optimized stylus keyboard layout called OPTI, and compared it to a stylus keyboard based on the familiar QWERTY layout. A longitudinal study was used. The main performance measure was words per minute (WPM), although error rates were also recorded (but will be ignored for our purposes). Open *macbtwn.jmp* and answer the following questions.

1. Is this data table in wide-format or long-format? How do you know?

   This data table is in wide-format because each subject occupies just one row in the table, with all measures over sessions 1-20 appearing in a single row for each subject.

2. How many subjects took part in this study?

   Ten subjects took part in this study.

3. What are any factors and their levels? For each factor, say whether it is between- or within-subjects.

   The study has two factors. *Method* is a between-subjects factor with two levels: OPTI and QWERTY. *Session* is a within-subjects factor with 20 levels: 1 through 20.

4. What is the formal multiplicative notation for describing this experiment?

   This experiment is a 2×20 mixed factorial design.

5. In any text entry experiment, we would assume subjects each entered multiple text phrases. Why, then, do you think the data table just shows one measure per subject per session? How was this achieved? What are the implications of making this choice in terms of both *power* and *variance*?

   Even though subjects probably entered multiple text phrases in each session, the data table shows just one measure per subject per session due to some kind of reduction strategy. For example, the measure may reflect the mean, median, minimum or maximum WPM achieved during a session. The choice to reduce each session to a single measure per subject is a choice that reduces power because less overall data is gathered, but it may also reduce variance depending how spread out the intra-session measures would have been. If, in a given session, a subject would have exhibited a range of text entry speeds, that variance will be lost via reduction. On the other hand, if most phrases would have been entered at very similar speeds, the intra-session variance would be minimal but the statistical power would be increased by having more data points.

Now we will analyze the data. The procedure is identical to that for module 4, part 7. Follow that procedure for this longitudinal data and then answer the following questions.

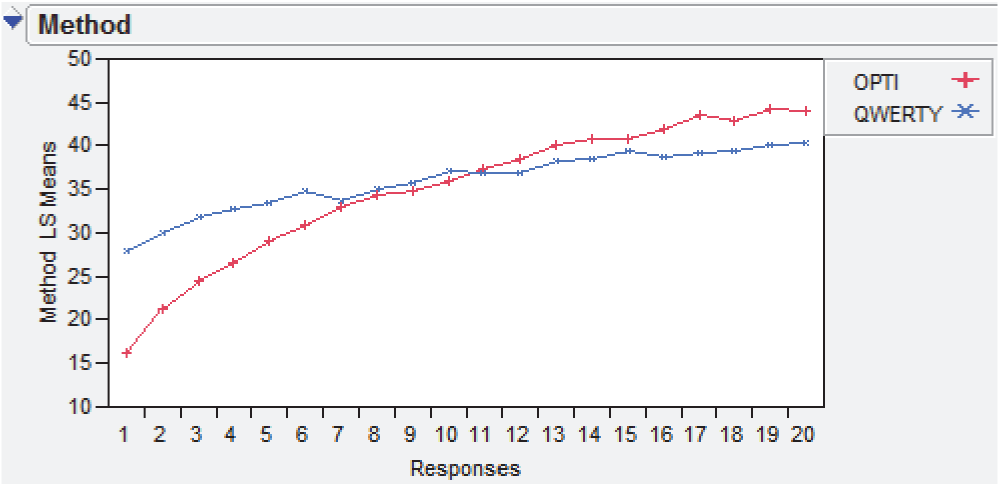
6. When using JMP's Manova platform, what name did you give to the repeated factor? What was the default name that JMP placed in the text box where you could enter a name of your own?

   The name given to the repeated factor was *Session*. The default name that was present in the text box was, appropriately, *Time*.

7. Using the graph of means over responses created by the JMP Manova platform, eyeball the crossover point (*see* Figure 1, above). In what session does the crossover point occur?

   The crossover point appears to have occurred in session 11.

8. On a Windows machine, use Start Menu > Accessories > Snipping Tool (or PrtSc and Paint) to produce the graph here, including its legend. Feel free first to adjust the graph dimensions by dragging its borders.



9. What is each method's performance in the first session?

   In the first session, OPTI's speed was 16.25 WPM and QWERTY's speed was 27.87 WPM.

10. What is each method's average performance across all sessions?

    Over all sessions, OPTI's speed averaged 35.06 WPM and QWERTY's speed was 35.99 WPM. These are very similar overall values, but the graph clearly shows that these means arise from very different learning curves.

11. What was each method's peak performance averaged over subjects per session? Did each peak occur in the last session?

    The peak performance of OPTI was 44.25 WPM and occurred in session 19. The peak performance of QWERTY was 40.31 WPM and occurred in session 20, the last session.

12. State and interpret the finding for the *Method* factor. What does this result mean?

    Method exhibited a significant main effect on WPM ($F(1,8) = 66.23$, $p < .0001$), indicating that there was a significant difference between the speeds of OPTI and QWERTY. The OPTI keyboard averaged 35.06 WPM with a low of 16.25 WPM and a high of 44.25 WPM. The QWERTY keyboard averaged 35.99 WPM with a low of 27.87 WPM and a high of 40.31 WPM. OPTI was slower than QWERTY in the early sessions (1-6), comparable to QWERTY in the middle sessions (7-15), and faster than QWERTY in the late sessions (16-20).

13. State and interpret the finding for the *Session* factor. What does this result mean?

Session exhibited a significant main effect on WPM (F(19,152) = 473.83, *p* < .0001), indicating that text entry speeds in general changed over sessions. Session averages improved from 22.06 WPM in the first session to 42.18 WPM in the last session.

14. State and interpret the finding for the *Method*Session* interaction. What does this result mean?

    There was a significant *Session*Method* interaction for WPM (F(19,152) = 74.69, *p* < .0001), indicating that each method's speed changed at different rates over sessions. OPTI started more slowly than QWERTY in the first session but performance climbed more steeply over sessions, resulting in OPTI outperforming QWERTY for the latter sessions. QWERTY's performance over sessions was flatter, starting higher than OPTI but peaking lower than OPTI.

Part 2, Mixed-effects model. Now transform this data into a long-format table for the purposes of performing a mixed-effects model analysis of variance. To do this, go to Tables > Stack, place the 20 response columns in Stack Columns, enter "WPM" for Stacked Data Column, and "Session" for Source Label Column. Click OK. Note that the Session column will be a character data type (text), but you want it to be numeric, so right-click on the column, choose Column Info, and for Data Type, set Numeric. (For now, leave its modeling type as nominal.) You may wish to save this data table as *macbtwn.long.jmp*. Then answer the following questions.

15. How many rows *should* there be in the long-format table? How many rows do you have?

    In *macbtwn.jmp*, there were 10 rows each with 20 measures, resulting in 200 data points. A long-format table has one measure per row, so we need 200 rows to represent these 200 data points. The new long-format data table does indeed have 200 rows.

16. Prepare for a mixed-effects model analysis of variance by going to Analyze > Fit Model. What is your Y response? What are your model effects? Is nesting appropriate? If so, what is nested in what? If not, why not?

    The Y response variable is *WPM*. The model effects are *Method*, *Session*, *Method*Session*, and *Subject[Method]*. As usual, we make *Subject* a random effect so it appears in the model as "Subject[Method]& Random." Nesting *Subject* in *Method* is appropriate because each subject used only one method, and we want to reduce unwanted variance that would be introduced by pooling subjects together across methods.

17. What is the fit of this mixed-effects model? Does it fit well? (Hint: look for the R^2 of the model.) You can also judge the model's fit by the -2 LogLikelihood value in the REML Variance Component Estimates section. What is this value? Smaller numbers for -2 LogLikelihood are better model fits.

    The model fits very well, with R^2=.98. The -2 LogLikelihood is 502.20, which is not useful in isolation but will be useful for comparisons to different models.

18. What are the results of this analysis of variance? Do the results agree with those found in your repeated measures ANOVA (questions 12-14)?

    The results of this analysis of variance agree exactly with those from our traditional repeated measures ANOVA in questions 12-14. The degrees-of-freedom, F-ratios, and *p*-values all match.

19. Now change the modeling type of *Session* to ordinal. What are the results of this analysis? How do they compare to the results from when it was nominal? Which model fits better according to R^2? According to -2 LogLikelihood? How do the degrees-of-freedom compare?

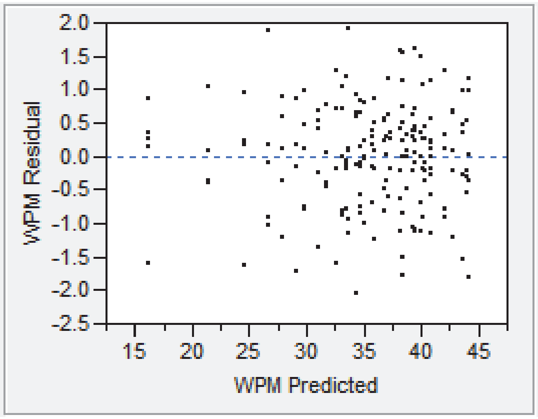
After changing the modeling type of *Session* to ordinal, the R^2 fit of the model is unchanged, but the -2 LogLikelihood is slightly improved from 502.20 to 490.21. The results for *Session* and *Method*Session* are unchanged; however, having turned *Session* to ordinal from nominal, the *Method* factor now has 160 denominator degrees-of-freedom instead of just 8, and the F-ratio increased to 479.08 from 66.23. The statistical result for *Method* is F(1,160) = 479.08, p < .0001. In looking at the Least Squares Means Table for *Method*, we can see that the Least Sq Mean values are the method speeds from the first session. By contrast, when *Session* was nominal, the Least Sq Mean values for *Method* were the averages over all sessions.

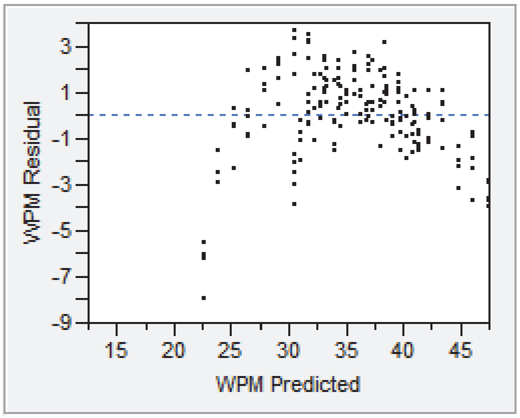
20. Now change the modeling type of *Session* to be continuous. What are the results of this analysis? How do they compare to the results from when it was nominal? ordinal? Which model fits better according to R^2? According to -2 LogLikelihood? How do the degrees-of-freedom compare?

After changing the modeling type of *Session* to continuous, the R^2 fit of the model drops from .98 to .88, and the -2 LogLikelihood is worsened from 502.20 and 490.21 to 827.54. The results for *Method* now agree with those from when *Session* was nominal (F(1,8) = 66.23, p < .0001). The results for *Session* now have only 1 numerator degree-of-freedom instead of 19, and 188 denominator-degrees of freedom instead of 152, and a much higher F-ratio (F(1,188) = 1588.30, p < .0001). The results for *Method*Session* also have these degrees-of-freedom, and a higher F-ratio than before (F(1,188) = 240.28, p < .0001).

21. Examine the plot of residuals. This plot should ideally seem as randomly scattered as possible; otherwise, a pattern is present in the data that is not being accounted for by the model. How do the residuals compare among the models with nominal, ordinal, and continuous modeling types for *Session*?

The residual plots for when *Session* was nominal or ordinal are identical, with points looking randomly scattered. The residual plot for when *Session* was continuous shows an arc pattern and do not look randomly scattered. These plots are shown below.





22. Which of the three modeling types seems most appropriate based on the above? Why? Which of the three modeling types *conceptually* seems most appropriate? Why?

On the one hand, the model fits (R^2 and -2 LogLikelihood) and the residual plots suggest that we make *Session* ordinal. On the other hand, when *Session* is ordinal, the least squares means for *Method*, and the resulting F-test, are based only on the performance of the methods' first sessions, not averaged over all sessions. Also, the results for *Method* agree when *Session* is nominal and continuous, lending support. All things considered, it seems most appropriate to model *Session* as a nominal variable. Conceptually, however, this seems to be the worst choice, as we know that *Session* is ordered in time, and should at least be ordinal if not continuous. Making it nominal suggests that the relationship of sessions 1 & 2 is equivalent to that of sessions 1 & 20. But making *Session* ordinal results in a somewhat dubious treatment of the *Method* factor, and making *Session* continuous reduces the model fit and causes patterns to emerge in the residuals.

23. Based on your exploration of *Session* modeling types, what *must* have been the implicit modeling type for *Session* in the traditional repeated measures ANOVA (questions 12-14)? Does your answer impact your final decision for the modeling type of *Session*? If so, how? If not, why not?

Our exploration of modeling types for *Session* shows that when *Session* is nominal, the results from our mixed-effects model analysis of variance agree exactly with the results from the traditional repeated measures ANOVA (*see* our answer to question 18). Thus, by using a wide-format table that encodes *Session* across 20 response columns and conducting a repeated measures ANOVA, we implicitly treat *Session* as a nominal factor. Given that the results for *Method* agree when *Session* is nominal or continuous, or when a repeated measures ANOVA is used, it seems appropriate to *not* model *Session* as ordinal, but either as nominal or continuous. Given the lower model fit and the patterned residuals for when *Session* is continuous, we affirm our choice to model *Session* as nominal, even when using a mixed-effects model. If conceptual considerations make this choice uncomfortable, the next best choice would be to model *Session* as continuous.

Part 3, Learning curves. We now will fit learning curves to this data. It is typical to fit curves to the average performance across subjects per method for each session. Therefore, we need to create an "average table" over subjects per session by method. Use Tables > Summary to create this from your long-format data table. After you have done so, be sure the table is sorted first by Method, then by Session number, since each method was used separately in each session. (If it is not sorted properly by session number, you may have neglected to set the data type of the Session column to numeric before answering question 15. Remedy this before proceeding.) Then answer:

24. What are the columns in your summary table?

The summary table has three meaningful columns, *Method*, *Session*, and *Mean(WPM)*. (By virtue of the Tables > Summary feature, an *N Rows* column is also produced, but we can delete or ignore this column.)

25. How many rows does your summary table have?

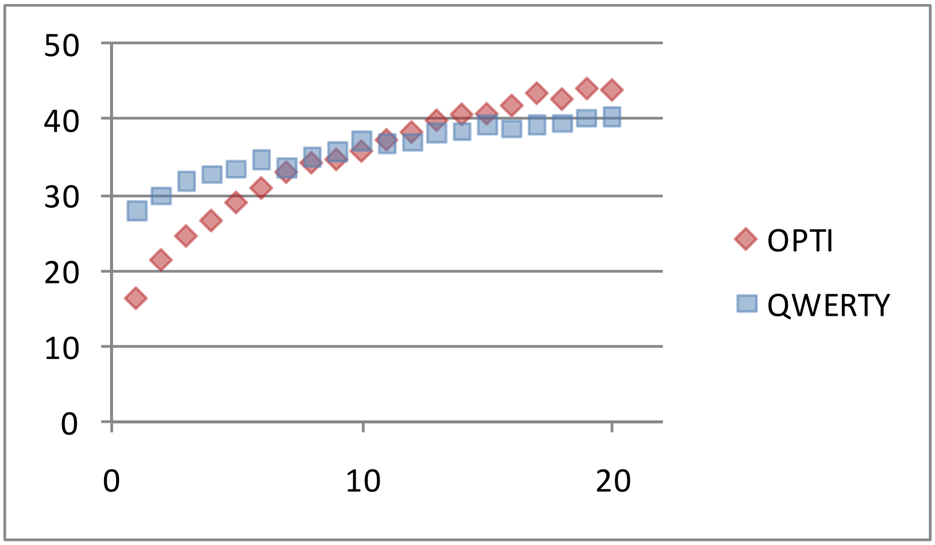
The summary table has 40 rows. This is correct for two methods that each have a speed result across 20 sessions.

26. What is the response column named?

The response column is *Mean(WPM)*, reflecting the average speed over subjects for that session within each method.

Now copy your summary data table into Microsoft Excel and save it as *macbtwn.xlsx*. (You can delete the N Rows column that JMP created.) Before proceeding, create a graph of WPM over sessions for both OPTI and QWERTY. Your graph should look similar to that created by JMP's Manova platform for question 8. However, do not connect the data points in each series.

27. Reproduce your Excel graph here, including its legend and marked axes.



Now we are going to use Microsoft Excel's Solver feature to fit a regression power curve for the Power Law of Learning. After the WPM column, add columns with the names *a*, *b*, Model WPM, RMSD, and Sum. As seeds, put the value 1.0 in row 2 under *a* and 0.1 in row 2 under *b*. These two columns, your regression coefficients, will only have one value in each, not a column full of numbers. Of course, you will need separate *a* and *b* values for the OPTI and QWERTY models, so halfway down, where QWERTY begins, put another 1.0 and 0.1 for QWERTY's *a* and *b*. (This should be in row 22 of the Microsoft Excel table.)

The Model WPM column contains model predictions based on the Power Law of Learning. At the start of this module you learned the equation for the Power Law, so you should be able to create the model predictions in this column. Fill up the column with these temporary seed predictions.

28. Do that now, and describe what you did here. Include the text of your Microsoft Excel formula.

The Power Law for this experiment is $WPM = a \times Session^{b}$. For OPTI, the *a* coefficient is in cell D2, the *b* coefficient is in cell E2, and the Session values are in cells B2 through B21. The model predictions are in cells F2 through F21. The Excel formula in cell F2 is =$D$2*B2^$E$2. In general, the formula in cell F*n* is =$D$2*B*n*^$E$2. Similarly, for QWERTY, the *a* coefficient is in cell D22, the *b* coefficient is in cell E22, and the Session values are in cells B22 through B41. The model predictions are in cells F22 through F41. The Excel formula in cell F22 is =$D$22*B22^$E$22. In general, the formula in cell F*n* is =$D$22*B*n*^$E$22.

The RMSD column stands for "root mean square deviation."[3] Each cell in this column contains a value for the difference between the observed response and the predicted response. Compute this column as the square root of the squared difference between observed WPM and model WPM. In symbols:

$$RMSD = \sqrt{\left(Y_{\text{observed}} - Y_{\text{model}}\right)^2} \, .$$

We need a way of knowing *overall* how far off our model predictions are from the observed data points. Do this by summing up RMSD in the Sum column, one sum for OPTI in row 2, and one sum for QWERTY in row 22.

29. What are the initial Sum values for OPTI and QWERTY?

    The initial sums for OPTI and QWERTY are 676.38 and 694.98, respectively.

Our goal now will be to *minimize Sum separately for each of the models*. Once we've done this, we'll have the best-fitting model possible. This requires an iterative improvement algorithm. Microsoft Excel provides this in its Solver feature, which for its first use must be installed. Go to the Office Button > Excel Options > Add-Ins and click Go. Check the Solver Add-In box, and click OK. Install the add-in when prompted.

30. The Solver can change the values of some cell(s) while minimizing or maximizing the value of another cell. In these terms, what should the solver do for fitting our power law curve for each of OPTI and QWERTY? *Record your answer before proceeding.*

    We want to minimize the difference between the observed WPM values and the model, or predicted, WPM values. These differences are captured in the RMSD column and summed up in the Sum column. Thus, we want the Solver to minimize the sum for OPTI and QWERTY by changing the values of the *a* and *b* coefficients in each model. Specifically, for OPTI, we want the Solver to minimize cell H2 while changing cells D2 and E2. For QWERTY, we want the Solver to minimize cell H22 while changing cells D22 and E22.

On the Data tab, in the Analysis section, you should now see the Solver. Click it. For fitting the OPTI curve, we want to set the target cell to the OPTI Sum in row 2. We want to minimize this by changing the values of OPTI's regression coefficients *a* and *b* in row 2. Then click Solve. Do this also for QWERTY's curve using the values of Sum, *a*, and *b* in row 22.

31. What are the final Sum, *a*, and *b* values for OPTI? What power law equation does this result in?

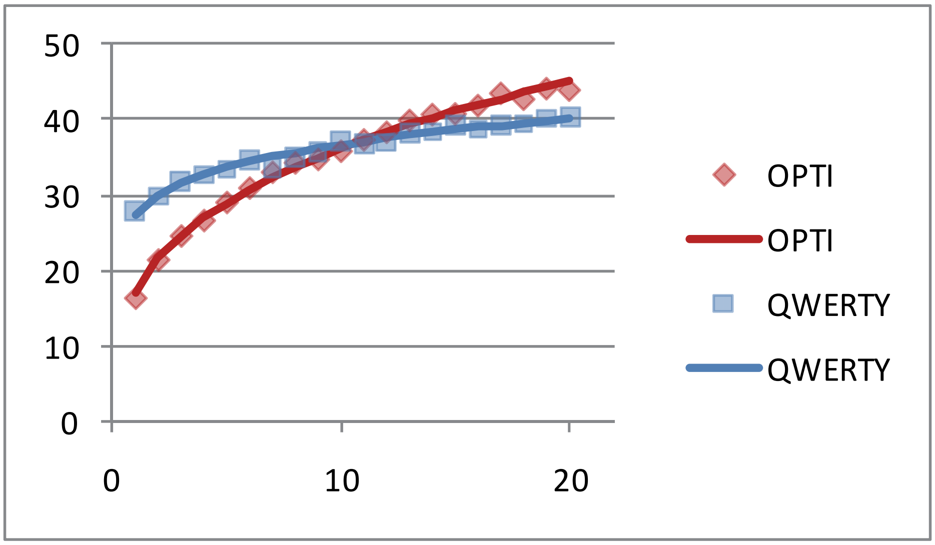
    The final Sum value is 8.45, leaving the *a* coefficient as 17.27 and the *b* coefficient as .3204. This results in a power law equation of $WPM = 17.27 \times Session^{.3204}$.

32. What are the final Sum, *a*, and *b* values for QWERTY? What power law equation does this result in?

    The final Sum value is 7.34, leaving the *a* coefficient as 27.44 and the *b* coefficient as .1263. This results in a power law equation of $WPM = 27.44 \times Session^{.1263}$.

[3] See http://en.wikipedia.org/wiki/Root_mean_square_deviation for more information.

33. Using your model predictions, plot the power curves on the graph you created previously. Do not plot points, just the power curves, so that we can easily see the original data points and the fitted curve over them. Display your graph here, including its legend.



We also want to know the goodness-of-fit of our power curves from R^2 values. Create a new column called R2 after Sum. For OPTI, in row 2, use the square of the CORREL or PEARSON Excel functions over the observed WPM and model WPM. For QWERTY, in row 22, do the same. (You can instead use the RSQ Excel function directly, but then do not square its results!)

34. What are the R^2 values for the OPTI and QWERTY models?

The R^2 fit for the OPTI model is .9953. For QWERTY, it is .9796.

35. Speculate as to why one of the models fits better than the other.

Although both models describe and predict performance very well, the OPTI model does so a bit better. This makes sense when we remember what the model captures—*learning*. There is probably more learning with OPTI than with QWERTY because OPTI is an unfamiliar keyboard layout, whereas QWERTY would presumably be familiar to most participants.

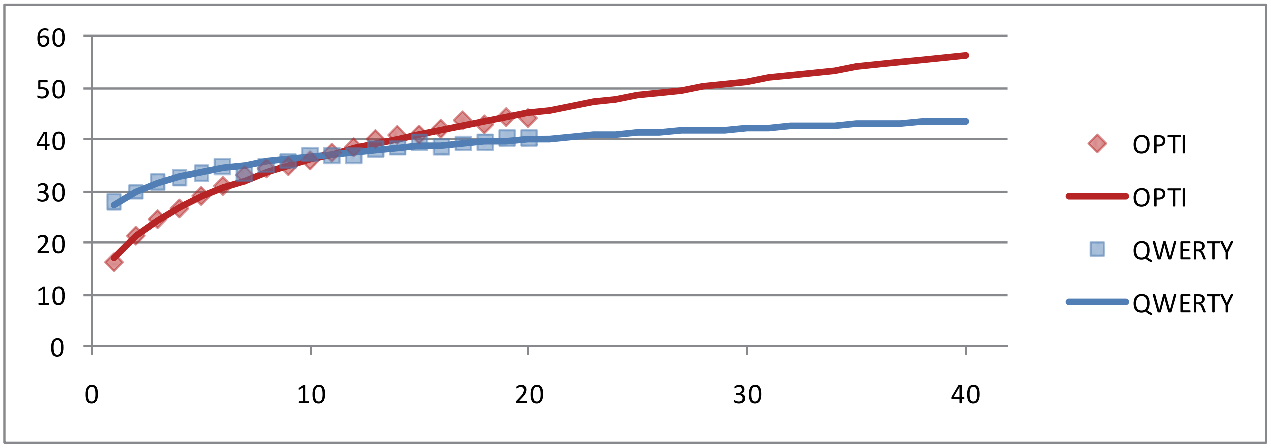
36. This data was estimated from the graph in MacKenzie & Zhang's (1999) original paper. In other words, it is "real human data." Does it surprise you that the Power Law of Learning models this data so well? Why or why not?

In the world of quantitative psychological models, the Power Law of Learning offers extremely robust model fits. Few behavioral science models regularly achieve such high R^2 values, and yet for innumerable human activities, this power law captures performance exceedingly well. So it does not come as a surprise that text entry speeds can be modeled in this way.

Models for data are useful not just for description, but also for *prediction*. Extend your predicted WPM values, and the curve you plotted, from session 20 to session 40 for each method.

37. What does the graph show as the predicted speed for OPTI and for QWERTY in the 40th session? Display your extended graph here, including its legend.

The predicted speed for OPTI in the 40th session is 56.32 WPM. The predicted speed for QWERTY in the 40th session is 43.73 WPM. The graph below shows the model predictions extended through 40 sessions.



Part 4, Covariance structures. In part 2, we analyzed the *macbtwn.jmp* data using a mixed-effects model analysis of variance. An important but hidden aspect of using mixed-effects models in JMP is that of specifying a *covariance structure*, which JMP does not expose but SPSS does. The covariance structure for repeated measures data indicates how measures "vary and vary together." For example, measures taken across sessions may vary differently with respect to each other than measures using different input methods in the same session. And yet, because the same subject was measured more than once, both are forms of repeated measures data.

As mentioned, SPSS allows us to pick the covariance structure when using mixed-effects models. Although there are numerous possible structures, the most common ones are Autoregressive, Compound Symmetry, Diagonal, Scaled Identity, and Unstructured. Although the underlying mathematical details are beyond the current scope, the intuition behind each one is explained below.

- **Autoregressive Order 1 (AR1).** Successive measures of a subject are more strongly correlated than distant measures. This will often be the case with measures of practice over time, *i.e.*, measurements at times 2 and 3 are more strongly correlated than measurements at times 1 and 3.
- **Compound Symmetry (CS).** This is the unchangeable default in JMP. The relationship between levels of a repeated factor all have constant variance and constant covariance; that is, measures n and m covary the same as measures n and k for $n \neq m \neq k$.
- **Diagonal (DIAG).** This is the changeable default in SPSS. Each level of the repeated factor has its own variance, but has zero covariance with every other level.
- **Scaled Identity (ID).** Each level of a repeated factor has unity (1) variance and zero (0) covariance with every other level.
- **Unstructured (UN).** The relationship between all $n \times m$ levels of the repeated factor is unstructured; that is, all variances and covariances are independent.

As stated, the unchangeable default in JMP is Compound Symmetry, while the changeable default in SPSS is Diagonal. To find out which is most appropriate for your data using SPSS, run each model and look at the *Information Criteria* output for each. Select the model that yields the *lowest* value for either the Akaike's Information Criterion (AIC) or Schwarz's Bayesian Criterion (BIC) values.

Paste your data from your long-format version of *macbtwn.jmp* into SPSS. After setting variable names, values, and types, perform a mixed-effects model analysis of variance. (Referring to questions 17-23, set the modeling type of *Session* to whatever you decided was most appropriate.) Go to Analyze > Mixed Models > Linear. On the first dialog, put *Subject* in Subjects and *Session* in Repeated. For Repeated Covariance Type, you will try the five covariance structures listed above. For now, start with Compound Symmetry to agree with JMP. Click Continue. Place *WPM* in Dependent Variable and *Method* and *Session* in Factor(s). Click Fixed and add *Method*, *Session*, and *Method*Session* in the Model. Click Continue. Click EM Means and transfer all factors to Display Means for. Click Continue. Click OK.

38. What modeling type did you choose for *Session*? Why?

Based on our reasoning in questions 22-23, we chose nominal for *Session*. (This exploration of covariance structures would also be worthwhile when modeling *Session* as continuous, but we leave that for future work.)

39. What are the AIC and BIC for Compound Symmetry? What are they for the other four covariance structures listed above? Make a table. Include -2 LogLikelihood for comparisons. Which covariance structure seems best suited to these data?

The table below shows the -2 LogLikelihood (-2LL), Akaike's Information Criterion (AIC), and Schwarz's Bayesian Criterion (BIC) for these data using the same mixed-effects model analysis of variance in SPSS.

	-2LL	***AIC***	***BIC***
Compound Symmetry	462.49	466.49	472.64
Autoregressive Order 1	461.67	465.67	471.82
Diagonal	445.42	485.42	546.92
Scaled Identity	462.51	464.51	467.59
Unstructured	Insufficient memory to estimate the model parameters.		

Regarding only -2LL, the Diagonal covariance structure is best, but it is the worst when regarding AIC or BIC. These latter two methods both indicate that Scaled Identity is best. With the exception of the Diagonal structure, however, the different covariance structures all give rise to similar information criteria.

40. How do your results using the best covariance structure compare to the results from using JMP (*see* questions 18-20)? Do the conclusions change from what you concluded previously based on using the best covariance structure?

The mixed-effects model analysis of variance in JMP using Compound Symmetry as our covariance structure with *Session* as a nominal variable resulted in significant effects for *Method* (F(1,8) = 66.23, *p* < .0001), *Session* (F(19,152) = 473.83, *p* < .0001), and *Method*Session* (F(19,152) = 74.69, *p* < .0001). Using Scaled Identity as our covariance structure in SPSS, we achieve very similar F-ratios, but with different denominator degrees-of-freedom: *Method* (F(1,160) = 61.37, *p* < .0001), *Session* (F(19,160) = 475.66, *p* < .0001), and *Method*Session*

(F(19,160) = 74.98, *p* < .0001). The conclusions of our analyses do not change based on switching from Compound Symmetry to Scaled Identity.

Part 5, Within-subjects. The original experiment from MacKenzie & Zhang (1999) had within-subjects data for *Method*, as each subject used either OPTI or QWERTY first in each session, followed by the other. Although MacKenzie & Zhang deterministically alternated the order of *Method* for each subject, let us pretend they chose the order randomly (*e.g.*, with a coin toss).

Open *macwthn.jmp*, which shows fictitious data similar to, but not the same as, that from the between-subjects data. Examine the data table and answer the following questions.

41. What are the columns in this data table?

    The columns in macwthn.jmp are *Subject*, *Handedness*, *Session*, *Method*, *Order*, *Trial*, and *WPM*.

42. How many rows are in the table? What does each row represent?

    There are 2000 rows in this data table. Each row represents one trial, which seems to be a single text entry phrase.

43. How many distinct subjects took part in this study? How many rows are there per subject? Explain how this number arises.

    There are five distinct subjects in this study. There are 400 rows per subject, which represent 400 trials. In each of 20 sessions, subjects used two methods entering 10 phrases with each. Thus, per subject, we have 20×2×10 = 400. With five subjects, we have 400×5 = 2000.

44. Is this data table in wide-format or long-format? How can you tell?

    This data table is in long-format because each row in the table corresponds to only one measure and subjects are (heavily) repeated across rows.

45. What are the factors and their levels? For each factor, say whether it is between-subjects or within-subjects.

    Session is a within-subjects factor with 20 levels (1-20). *Method* is a within-subjects factor with 2 levels, OPTI and QWERTY. *Trial* is a within-subjects factor with 10 levels (1-10).

46. Are there any covariates? If so, what are they? Are they between- or within-subjects?

    Handedness is a between-subjects covariate with two values, right and left. *Order* is a within-subjects covariate with two values, 1 and 2 (equivalent to "first" and "second").

47. Explain the meaning of the *Order* column. It indicates the order of what, exactly?

    The *Order* column indicates which method a subject used first or second in a given session. It allows us to test whether counterbalancing worked or whether the order of presentation had a significant effect on *WPM*.

48. Explain the meaning of the *Trial* column. What does this tell us about how the experiment was run?

The presence of the *Trial* column tells us that 10 individual trials were administered per method per session per subject. Although we do not know for sure, we can speculate that a trial consisted of the entry of a single text phrase.

49. What is the formal multiplicative notation for describing this experiment?

The experiment is a 2×20×10 within-subjects design with repeated factors *Method*, *Session*, and *Trial*.

Now, prepare to conduct an analysis of variance using a mixed-effects model. *Before* you do so, answer the following questions.

50. What should be the Y response for the model?

The Y response for the model should be words per minute, or *WPM*.

51. Ignoring *Handedness* and *Order* for now, what should be the model effects? (Hint: think carefully about whether any effects should be nested, or even doubly nested.)

The model effects should be *Session*, *Method*, *Session*Method*, and *Trial[Session, Method]*. We justify the nesting of *Trial* in *Session* by reasoning that we do not want all trial *n*'s to be pooled together, as trial *n* performed in session *m* will be quite different than trial *n* performed in session *k* for *k ≠ m*. We justify the nesting of *Trial* in *Method* by reasoning that we do not want trial *n* with OPTI to be pooled with trial *n* with QWERTY; these should be modeled separately. Thus, we end up with *Trial* modeled separately for each method in each session. Finally, we account for repeated measures on subjects by including *Subject* as a random effect, written as "Subject& Random" in the model. *Subject* is not nested as there are no between-subjects factors.

52. Did you encounter the following issue? If a model contains *A[B]*, then it is not allowed to contain the interaction *B*A[B]*. Why does such an interaction make no sense?

We would encounter this issue if we attempted to include interactions between *Trial* and *Session* or *Trial* and *Method*. The reason these interactions make no sense in light of the nesting of *Trial[Session, Method]* is that this nesting means trials are modeled *within* each level of method and session, but an interaction requires a crossing of terms, *i.e.*, the combination of levels of the crossed factors. With nesting, these levels are never combined, only contained.

Only after recording your answers to the above questions should you proceed to run your analysis. (Be patient; complex models take a while to execute!)

53. What was the final model you ran? What were any nested and/or random effects?

The final model contained *Session*, *Method*, *Session*Method*, *Trial[Session, Method]*, and *Subject* as a random effect, written as "Subject& Random."

54. Now open the saved analysis called "WPM." Examine the model effects. How do they differ from the model you used?

The saved model matches the model used.

55. Why is *Trial[Session, Method]* in the model? What does this effect mean? (Hint: you may want to look in the JMP help documentation for nested effects.)

    Our answer to question 51 contains the rationale for nesting *Trial* in both *Session* and *Method*.

56. Why are there no interactions for *Trial* in the model?

    By nesting *Trial* in *Session* and *Method*, we remove the possibility of interactions between these factors and *Trial* (*see* question 52). This is satisfactory because such interactions are not particularly meaningful anyway.

57. Judging by R^2, how well does this model fit the data?

    The linear mixed-effects model fits the data very well at R^2 = .986.

58. Report and interpret the results of this analysis.

    In a longitudinal study of five subjects, each subject participated in 20 consecutive sessions in which he or she used two soft keyboards to enter 10 text phrases on a mobile device. The soft keyboards were OPTI and QWERTY, and their order in each session was determined by a coin toss. The experiment design was a 2×20×10 within-subjects design with factors for *Method* (OPTI, QWERTY), *Session* (1-20), and *Trial* (1-10). A mixed-effects model analysis of variance shows that *Session* exhibited a significant main effect on *WPM* (F(19,1596) = 4990.23, *p* < .0001), indicating unsurprisingly that *WPM* changed over sessions. Average performance began at 21.89 WPM (*SD* = 5.78) in the first session and climbed to 42.22 WPM (*SD* = 2.17) by the last session. *Method* also exhibited a significant main effect on *WPM* (F(1,1596) = 442.40, *p* < .0001), indicating that overall, the two methods performed differently. Overall, OPTI's 35.16 WPM (*SD* = 7.89) was slightly slower than QWERTY's 35.93 WPM (*SD* = 3.53). However, this main effect must be interpreted cautiously in light of the significant *Method*Session* interaction (F(19,1596) = 776.29, *p* < .0001), indicating that over sessions, each method's speed changed differently. Whereas OPTI started slower than QWERTY (16.19 *vs.* 27.58 WPM), OPTI finished faster than QWERTY (44.25 *vs.* 40.20 WPM). Lastly, *Trial[Session, Method]* did not exhibit an effect on *WPM* (F(360,1596) = 1.01, *n.s.*), indicating that within a session for a given method, *WPM* did not change significantly over trials.

59. How do these findings compare to those for *macbtwn.jmp* in questions 12-14?

    Owing to the vast increase in data from *macbtwn.jmp* to *macwthn.jmp*, and to the increased power from using a within-subjects analysis, the F-ratios are generally much stronger in our current analysis. However, the statistical conclusions are the same as for the between-subjects data, and the session and method means are very similar.

Part 6, Covariates. Now let us consider the covariate for handedness. It may be that handedness makes no difference, but since we bothered to record it for each participant, we might as well check. There is a reasonable hypothesis that handedness matters for stylus keyboard use, so it is a justifiable exploration. Conceptually, we are not interested in *Handedness*Session* or *Handedness*Trial[Session, Method]*. It is unclear what such findings would really mean. We are, however, interested in a main effect of handedness or in an interaction between handedness and methods. To examine these, run the saved model "Handedness."

60. How many subjects were right-handed? How many were left-handed?

    There were three right-handed subjects and two left-handed subjects.

61. What is a rationale for any nested and/or random effects?

    The saved "Handedness" analysis contains the term "Subject[Handedness]& Random," meaning that *Subject* is nested within *Handedness* and that *Subject* is a random effect. Nesting is desirable because we want to remove any unwanted variance that may arise by pooling right- and left-handed subjects. *Subject* is a random effect because subjects were sampled randomly from a larger population about which we wish to draw conclusions, and that we do not care about the differences among our levels of *Subject* itself. (For more on random effects, *see* modules 3-5.)

62. What would it mean to have a *Handedness* main effect? Give an example that would give rise to such an effect.

    A main effect for the covariate *Handedness* would indicate that right- and left-handed subjects entered text at significantly different speeds. For example, if left-handed subjects entered text faster than right-handed subjects, this would produce a *Handedness* main effect.

63. What would it mean to have a *Handedness*Method* interaction? Give an example that would give rise to such an interaction.

    A *Handedness*Method* interaction would indicate that the two text entry methods' speeds were affected disproportionately by right- and left-handed subjects. For example, if one method enjoyed an advantage with left-handed subjects that the other method did not enjoy.

64. Write up the conclusions regarding the effects of handedness on each method's performance.

    Handedness did not exhibit a main effect on *WPM* (F(1,3) = 0.38, *n.s.*), indicating that whether a subject was right- or left-handed did not, in itself, cause a difference in entry speed. Neither was there a significant *Handedness*Method* interaction (F(1,1993) = 0.06, *n.s.*), indicating that neither method was affected disproportionately by right- or left-handed subjects.

Part 7, Order effects. We also care about whether or not the coin toss for counterbalancing method presentation in a given session "worked." The *Order* column allows us to check for this. Conceptually, we are interested in a main effect of *Order* and an *Order*Method* interaction. Run the saved model "Order" before answering the following questions.

65. Over the entire experiment, how many times did OPTI go first? How many times did QWERTY go first? Describe in step-by-step detail how you arrived at these answers.

    With five subjects and 20 sessions, there is a total of 100 occasions to tally for whether a method went first or second. OPTI went first 57 times and QWERTY went first 43 times. This was determined by first using Tables > Summary and placing Session, Method, and Order in Group, with nothing in Statistics. On the resulting summary table, Rows > Row Selection > Select Where (CTRL+SHIFT+W) was used to select all rows for which Order equals 2; then Rows > Delete Rows was used to remove them. Tables > Sort was used to sort the table by Method. A row's N Rows column divided by 10 gives the number of subjects who went first with that row's method in that row's session. These values were totaled in Microsoft Excel for each method and then divided by 10.

66. Conceptually, what would an *Order* main effect mean? Give an example that would give rise to such an effect.

    A main effect for *Order* would indicate that *Method*'s order of presentation significantly affected text entry speeds. For example, if due to fatigue, the method used first in each session were significantly faster than the method used second.

67. Conceptually, what would an *Order*Method* interaction mean? Give an example that would give rise to such an interaction.

    An *Order*Method* interaction would indicate that the two text entry methods' speeds were affected asymmetrically by their order of presentation. For example, if when OPTI was presented first, it did better than when it was presented second, but this pattern did *not* hold for QWERTY, we would see an *Order*Method* interaction.

68. What are your conclusions regarding order effects? Did the coin-toss counterbalancing work? How do you know?

    Order did not exhibit a main effect on *WPM* (F(1,1992) = 0.20, *n.s.*), indicating that order of presentation alone did not affect text entry speeds. Also, a *Method*Order* interaction was not significant (F(1,90.8) = 2.40, *n.s.*), indicating that neither method was given an advantage over the other by being presented first. (Note that the fractional denominator degrees-of-freedom is due to the imbalance of *Order* across *Method*, with 57% of the time OPTI being presented first and only 43% of the time QWERTY being presented first.) We can conclude that coin-toss counterbalancing, although imperfect in the resulting proportions, did manage to work.

69. What could you argue to save the integrity of your experiment if you ever found an *Order* main effect but not an *Order*Method* interaction? How can this be okay?

    Even if an *Order* main effect is present, provided there is no significant *Order*Method* interaction, *and* provided that full counterbalancing was used, the statistical conclusions still can be trusted. The fact that a method being presented first made it perform faster, for example, is not worrisome provided that (1) this pattern occurred for both methods, and that (2) the methods were presented first an equal number of times. Issue #1 can be established with a nonsignificant *Method*Order* interaction. Issue #2 can be established by showing an equal number of occasions in which each method was presented first. Full counterbalancing guarantees this.

Part 8, Repeated measures ANOVA. For questions 53-59, we used a mixed-effects model analysis of variance on *macwthn.jmp*. For comparisons, let's also do this analysis (once again, ignoring *Handedness* and *Order*, having examined those already) using a traditional repeated measures ANOVA. Recall that JMP can only handle up to two repeated factors, but for *macwthn.jmp*, we have three: *Session*, *Method*, and *Trial*. One approach would be to "average down" over trials to create one measure per subject per session per method, as we had in *macbtwn.jmp*. But this costs us statistical power. If we wish to retain *Trial*, we'll need to use SPSS, which as you will recall, allows for an arbitrary number of repeated factors. (See module 5, part 7.) And we'll need a (very!) wide table.

Using Tables > Split, transform *macwthn.jmp* into a wide-format table. (Hint: you will need to split three times, once for each within-subjects factor.) Run SPSS and paste this data into it, naming your data table *macwthn.wide.sav*. You don't want to manually name the 400 data columns you'll have in SPSS; instead, first paste the column names from JMP into Microsoft Excel. Then use the Excel command SUBSTITUTE to replace any spaces with underscores, as SPSS does not allow spaces in its column names. Also, SPSS won't allow numbers to be the lead characters in column names, so if you split such that the session number is the leading character, use CONCATENATE to prepend an "s" to the front of each column name, *e.g.*, s1_OPTI_5 would be session 1, OPTI keyboard, trial 5. Once your column names are prepared, copy them with CTRL+C in Microsoft Excel; in SPSS, drag your mouse down the Name column in the Variable View, and paste the names all at once using CTRL+V.

Once the SPSS data table is ready, conduct a repeated measures ANOVA by going to Analyze > General Linear Model > Repeated Measures. Fill out the dialog box and execute the analysis. Remember that unlike with our mixed-effects models, explicit nesting is not applicable here, so *Trial[Session, Method]* is not modeled, but instead *Trial* is an unnested model term just like *Session* and *Method*.

70. What are the results? Write up a publishable paragraph of findings similar to the one you crafted for question 58.

    In a longitudinal study of five subjects, each subject participated in 20 consecutive sessions in which they used two soft keyboards to enter 10 text phrases on a mobile device. The soft keyboards were OPTI and QWERTY, and their order in each session was determined by a coin toss. The experiment design was a 2×20×10 within-subjects design with factors for *Method* (OPTI, QWERTY), *Session* (1-20), and *Trial* (1-10). A repeated measures ANOVA shows that *Session* exhibited a significant main effect on *WPM* (F(19,76) = 4416.96, *p* < .0001), indicating that *WPM* changed over sessions. Average performance began at 21.89 WPM (*SD* = 5.78) in the first session and climbed to 42.22 WPM (*SD* = 2.17) by the last session. *Method* also exhibited a significant main effect on *WPM* (F(1,4) = 182.74, *p* < .0001), indicating that overall, the two methods performed differently. Overall, OPTI's 35.16 WPM (*SD* = 7.89) was slightly slower than QWERTY's 35.93 WPM (*SD* = 3.53). However, this main effect must be interpreted cautiously in light of the significant *Method*Session* interaction (F(19,76) = 568.00, *p* < .0001), indicating that over sessions, each method's speed changed differently. Whereas OPTI started slower than QWERTY (16.19 *vs.* 27.58 WPM), OPTI finished faster than QWERTY (44.25 *vs.* 40.20 WPM). *Trial* did not exhibit a significant main effect (F(9,36) = 0.62, *n.s.*), and neither did any interactions with *Trial*: *Session*Trial* (F(171,684) = 1.14, *n.s.*), *Method*Trial* (F(9,36) = 1.09, *n.s.*), *Session*Method*Trial* (F(171,684) = 0.97, *n.s.*).

71. How do these results compare to those from your mixed-effects model analysis of variance (*see* question 58)?

    The results from this traditional repeated measures ANOVA are similar to those from the mixed-effects model analysis of variance. Although the degrees-of-freedom and F-ratios are different, the significance conclusions are the same. The table below compares the two sets of results.

	repeated measures ANOVA				***linear mixed-effects model***			
	df$_{num}$	*df*$_{den}$	*F*	*p*	*df*$_{num}$	*df*$_{den}$	*F*	*p*
Session	19	76	4416.96	.000	19	1596	4990.23	<.0001
Method	1	4	182.74	.000	1	1596	442.40	<.0001
Trial	9	36	0.62	.773				
Trial[Session, Method]					360	1596	1.01	.454
*Session*Method*	19	76	568.00	.000	19	1596	776.29	<.0001
*Session*Trial*	171	684	1.14	.139				
*Method*Trial*	9	36	1.09	.393				
*Session*Method*Trial*	171	684	0.97	.607				

Part 9, Learning curves. Referring to the work you did in part 3 graphing performance over sessions and fitting learning curves, return to JMP to prepare your long-format *macwthn.jmp* data table for the same treatment.

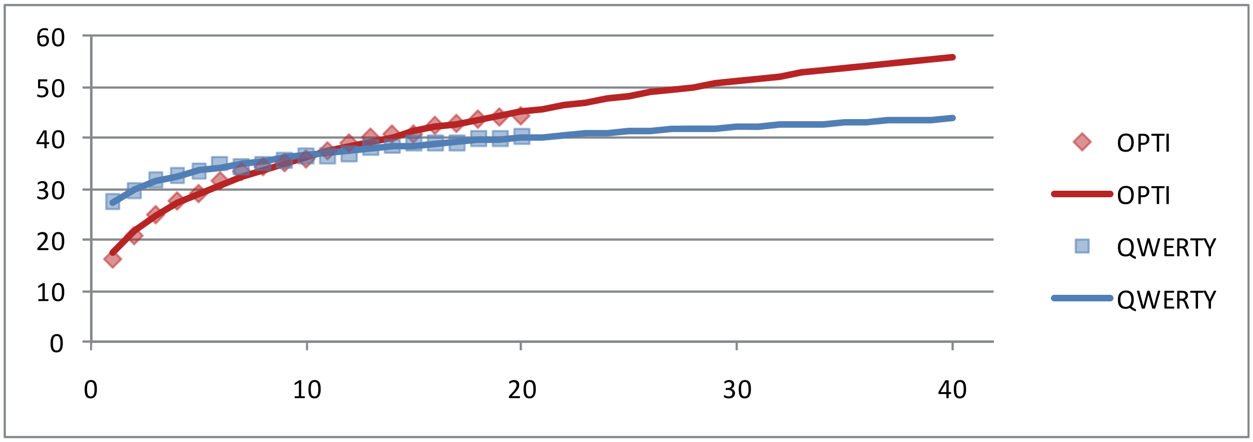
72. Describe what procedure you followed to prepare your data for fitting learning curves.

We used Tables > Summary on *macwthn.jmp*, putting Mean(WPM) in Statistics, and Session and Method in Group. This resulted in 40 rows. Then we used Tables > Sort and included Method and Session in that order in the By box. Finally, we moved the Method column to be first and deleted the N Rows column.

73. Paste your summary data into Microsoft Excel and follow the same procedure as in part 3 to plot the *macwthn.jmp* data, fit a Power Law of Learning curve, compute R^2 values, and predict performance to 40 sessions. Save this data table as *macwthn.xlsx*. Here, give the power law equations for OPTI and QWERTY, their R^2 fits, and their predicted performance in the 40th session.

The power curve equation for OPTI is WPM = 17.64 × $Session^{.3133}$ with R^2 = .9946. The power curve equation for QWERTY is WPM = 27.28 × $Session^{.1281}$ with R^2 = .9882. Predicted performance in the 40th session for OPTI is 56.03 WPM. For QWERTY, it is 43.76 WPM.

74. Place your finished graph with performance extended to 40 sessions below.



Part 10, Learning stabilization. It is often important to know when learning has stabilized during a longitudinal study. In casual terms, we might say that the learning curve has "flattened out." Of course, humans will keep improving a skill for a very long time, even if that improvement becomes almost imperceptible. Therefore, we cannot expect a learning curve to truly *flatten*. But we might like, for example, to compare sessions only *after* learning has stabilized by some quantifiable definition. By doing so, we can find a principled way to remove sessions where learning was still prominent, retaining sessions for analysis in which learning was stable. To avoid bias, the method of determining learning stabilization should be determined *before* analyzing the data.

Multiple comparisons. One approach to determining when learning has stabilized is to use the Tukey-Kramer HSD ("honestly significant difference") test for multiple comparisons across sessions comparing each session to the one before it. At least three criteria for determining "stabilized" are possible from such a test. From least to most conservative (modulo chosen thresholds), the point of stabilization would be:

- *Criterion 1.* The first point at which a predetermined number *N* of consecutive sessions cease to be significantly different. *N* may be just two, or more, or a percentage (*e.g.*, 20% of conducted sessions).
- *Criterion 2.* The first point at which the number of consecutive sessions that are not significantly different never decreases (ignoring the end boundary), perhaps with a minimum threshold (*e.g.*, of *N* = 3).
- *Criterion 3.* The first point at which all remaining consecutive sessions up through the last session are not significantly different.

We will use your long-format version of *macbtwn.jmp*, although our procedure is not specific to it. (You may have saved it as *macbtwn.long.jmp*.) After you have this data in long-format, go to Analyze > Fit Y by X and place *WPM* as your Y, Response, *Session* as your X, Factor, and *Method* in the By box. Click OK. You now have plotted the data points separately for each method over sessions. Use the red-arrow menu on the OPTI report and choose Compare Means > All Pairs, Tukey HSD. (Unlike the Student's *t*-test, the Tukey test will adjust alpha (α) for the number of comparisons made.) After executing Tukey, use the newly created red-arrow menu to uncheck everything except the Connecting Letters Report. Then right-click anywhere inside the report and choose Make into Data Table. On the data table, right-click the Level column and choose Column Info. Set the Data Type to Numeric. Then sort the table using Tables > Sort by descending Level order. Repeat this procedure for the QWERTY results. Then answer the following questions.

75. Carefully study the connecting letters report from the Tukey-Kramer HSD test. Describe what this report shows. What is in the leftmost column? What do you notice about it? What is the rightmost column? How is the table sorted? Why did you have to make your own tables sorted by Level?

    The connecting letters report indicates which levels of the X variable in the Fit Y by X platform are significantly different according to a Tukey-Kramer HSD test. Levels connected by the same letter are not significantly different; levels not sharing the same letter *are* significantly different. The leftmost column gives the levels of the X variable, in our case, *Session*. It is important to notice that it is not sorted in descending numeric order. The rightmost column is the mean value of the X variable at the given level. The table is sorted by descending mean value, which is why the Level column is not the sort column. We had to make our own tables sorted by Level so that we could more easily judge which sessions were *consecutive*.

76. Using the first of the three criterion to determining stabilization above, and using $N = 4$ (or 20% of conducted sessions), at which session would we say learning stabilized for OPTI? for QWERTY?

Using the first criterion and *N* = 4, we can say learning stabilized for OPTI by the 13th session. It is here that four consecutive sessions' performance is not significantly different, namely sessions 13-16. For QWERTY, we can say learning stabilized by the 5th session, as performance in sessions 5-8 is not significantly different.

77. Using the second criterion and a threshold of $N = 3$, at which session would we say learning stabilized for OPTI? for QWERTY? (Hint: make an extra column in your data tables and in it, place the number of consecutive sessions that are not significantly different; a value of "1" indicates that a session is only not different from itself. Then find the first number $\geq N$ after which the numbers never decrease, ignoring the inevitable decrease at the end of sessions.)

Using the second criterion and a threshold of *N* = 3, we can say learning stabilized for OPTI by the 16th session. For QWERTY, we can say learning stabilized by the 12th session. It is only after these points that the number of nonsignificantly different consecutive sessions never decreases.

78. Using the third criterion, at which session would we say learning stabilized for OPTI? for QWERTY?

Using the third criterion, we can say learning stabilized for OPTI by the 17th session. For QWERTY, we can say learning stabilized by the 15th session. It is only at these points that all remaining consecutive sessions are nonsignificantly different.

79. Return to your long-format data table of 200 rows and, using your answer for each criterion from questions 76-78, exclude all sessions for which learning was not yet stabilized. (You may wish to use Rows > Row Selection > Select Where (CTRL+SHIFT+W) and then Rows > Exclude (CTRL+E).) For each criterion, how many sessions are considered "stable?" How many total rows are thereby excluded?

The table below shows for each criterion and for the two methods the sessions included and the table rows excluded.

	Criterion 1		***Criterion 2***		***Criterion 3***	
	Stable Sessions	*Rows Excluded*	*Stable Sessions*	*Rows Excluded*	*Stable Sessions*	*Rows Excluded*
OPTI	13-20	60/100	16-20	75/100	17-20	80/100
QWERTY	5-20	20/100	12-20	55/100	15-20	70/100
		80/200		**130/200**		**150/200**

80. Using each criterion to exclude the proper rows, run a mixed-effects model analysis of variance on *macbtwn.jmp* to compare the methods' stable performance. (Hint: because you are now unbalanced, you must make *Session* continuous and you will see fractional degrees-of-freedom.)

Report the model fit (R^2 and -2 LogLikelihood) and your F-test results for each of the three criterion. You may wish to make one or more tables to convey your results. Then reflect on any trends you see in your results based on the conservativeness of the stabilization criterion.

	Criterion 1		R^2 = .9211	-2LL = 329.27
	df_{num}	df_{den}	F	p
Session	1	108.30	274.82	<.0001*
Method	1	34.50	118.25	<.0001*
*Session*Method*	1	108.30	6.70	.0110*
	Criterion 2		R^2 = .8400	-2LL = 186.07
	df_{num}	df_{den}	F	p
Session	1	59.72	36.22	<.0001*
Method	1	53.72	231.03	<.0001*
*Session*Method*	1	59.72	0.78	.3819
	Criterion 3		R^2 = .8795	-2LL = 130.01
	df_{num}	df_{den}	F	p
Session	1	37.65	8.49	.0060*
Method	1	10.88	193.05	<.0001*
*Session*Method*	1	37.65	0.00	.9794

In comparing these three mixed-effects model analyses of variance, we can see that as the criterion by which sessions are deemed "stable" grows more conservative, and therefore more sessions are excluded, the strength of the *Session* and *Session*Method* findings weakens as increasingly more stable sessions are analyzed. Similarly, the *Method* F-ratio grows stronger, although using the most conservative criterion (3), power is reduced enough that the F-ratio grows slightly weaker compared to criterion 2. In the end, however, the only statistical finding that differs among the three criteria is that of the *Session*Method* interaction, which is significant only under the most inclusive criterion (1).

Successive linear regression. Another approach to determining learning stabilization is to first create a summary table of session means (per method), and then fit regression lines through the session means, removing points from the early sessions successively until the regression is not significantly different from horizontal. When a regression line is fitted in JMP, an ANOVA result is produced, which indicates whether the regression fit is significantly better than a horizontal line set at the mean value of the data.

Compute a summary of *macbtwn.jmp* containing session means for each method. After doing so, turn *Session* into a continuous variable for regression purposes. Use Analyze > Fit Y by X with *WPM* in Y, Response, *Session* in X, Factor, and *Method* in By. On each graph, use the red-arrow menu to select Fit Line. Examine the Analysis of Variance table and the *p*-value there. As long as it remains $p < .05$, exclude points, each in turn, from the starting sessions using Rows > Exclude (Ctrl+E). After each exclusion, repeat this process until the linear fit is not significantly different from horizontal. This can be used as an indication for where learning stabilized.

81. Having followed this process above for OPTI and QWERTY, indicate at which session each method stabilizes under this scheme. Give the nonsignificant regression ANOVA results for each method.

Following the successive linear regression scheme, OPTI stabilized by the 16th session (F(1,3) = 5.53, *p* = .100), while QWERTY stabilized by the 18th session (F(1,1) = 6.00, *p* = .247).

82. Repeat your mixed-model analysis of variance on the main long-format data table of 200 rows after excluding sessions for each method according to your answer to question 81. How many rows are excluded for each method? What are the results of your analysis?

Excluding rows according to our answer to question 81 results in 75/100 rows excluded for OPTI and 85/100 rows excluded for QWERTY, for 160/200 excluded rows in all. A mixed-effects model analysis of variance results in a model with R^2 = .7678 and -2LL = 110.95. *Session* exhibited a main effect on *WPM* (F(1,28.05) = 7.16, *p* = .0123), as did *Method* (F(1,19.84) = 127.06, *p* < .0001). There was no significant *Session*Method* interaction (F(1,28.05) = 0.05, *p* = .8197).

83. How does the successive linear regression approach compare to the answers obtained for each of the three criteria based on multiple comparisons? Is it more or less conservative?

Successive linear regression regarded OPTI's stable performance as consisting of sessions 16-20, which matches Criterion 2 from the multiple comparisons approach. However, successive linear regression regarded QWERTY's stable performance as consisting only of sessions 18-20, which is more conservative than any of the multiple comparisons criteria. The statistical conclusions, however, still match those from all three criteria except for the significant interaction under Criterion 1 that is nonsignificant under Criteria 2-3 and successive linear regression.

84. Of the four approaches, which do you think is conceptually most sound? least sound? Why?

Conceptually, it seems that either multiple comparisons' Criterion 2 or 3 are most sound. In the case of Criterion 2, it makes sense that we would regard stabilization as a point where enough successive sessions agree (our threshold), and from that point, this level of agreement never lessens unto the end of sessions. Criterion 3 is stricter and the most conceptually "pure," requiring for stabilization that all sessions agree to the end. The least sound approach seems to be Criterion 1 because all it requires is a stretch of agreement regardless of any disagreement that may come later. However, Criterion 1 is highly dependent upon its chosen threshold, and the greater it is, the less valid this criticism becomes.

85. Of the four approaches, which do you think is easiest to employ? hardest to employ? Why?

Mechanically, it is easiest to employ either multiple comparisons' Criterion 1 or 3, as both only require a quick visual inspection of a sorted-by-session connecting letters report. It is more difficult to employ Criterion 2 because of the need to judge how many sessions agree from each session. Successive linear regression is conceptually easy but can be tedious, since multiple regression fits must be made until the regression becomes nonsignificant. The good news evident in this exercise is that all four approaches result in very similar statistical conclusions, with the significant *Session*Method* interaction under Criterion 1 being the only "outlier" conclusion among the approaches. If we think about it, after learning has stabilized, *Session*Method* shouldn't be significant any longer!

09. Nonparametric tests

nonparam.xlsx

Until now, our modules have mostly used parametric tests. Although deeply technical definitions exist for "parametric" and "nonparametric," intuitively, they can be thought of as, in the former, assuming an underlying normal distribution, and in the latter, being free of assumptions about distribution. By "underlying normal distribution," we mean that the measure in question taken over the whole population would exhibit a Gaussian curve. While this is true for many measures (*e.g.*, height, weight, age, certain abilities), it is not true for all (*e.g.*, household income, prison sentence durations). Some measures follow non-normal distributions; common alternatives are *lognormal*, *exponential*, *binomial* and *Poisson*. Still other data exhibits no discernable distribution at all. You encountered these concepts briefly in module 6.

You ran a nonparametric test in module 4, part 3, when you learned how to use the JMP and SPSS statistical tools. Some common analyses are available in only one or the other of these tools. Still others are not widely available at all.

An important feature of many nonparametric tests is that they use *ranks* on the data being analyzed, rather than the raw data points themselves. By ranking data, and using average ranks (or "midranks") for ties, distributions that were present in the data are destroyed.

In module 6, you learned how to examine the distributions of data and test for normality. Because we cannot know the distribution of the whole population, we must use our sample as an approximation. Thus, if our sample's measure is normally distributed, we can assume that the population's measure (also known as the "population parameter") is also normally distributed. This is but one of the assumptions required for parametric *t*- and *F*-tests.

Part 1, ANOVA assumptions. There are classically *three* assumptions required for one-way ANOVA. Discover these and list them below. Beside each item, give *your own* definition that indicates you grasp the *intuition* behind the assumption. Then say how you would check each assumption using a tool like Microsoft Excel, JMP, or SPSS.

1. ***Independence of cases.*** Responses must be independent from one another. This is not mathematically tested but is confirmed from the experiment design.
2. ***Normality.*** Responses form a normal distribution. This can be tested with a Shapiro-Wilk *W* test in JMP using Analyze > Distribution > Fit Distribution > Normal > Goodness of Fit. (Technically, this assumption is actually about the normality of residuals, which are the differences between model and observed values.)
3. ***Equal variances.*** Called "homoscedasticity," the variance (or standard deviation) among different experimental groups should be about the same. Levene's test for homogeneity scrutinizes this. If homoscedasticity is violated, a Welch ANOVA can be used, which does not assume equal variances. In JMP, these tests are both available in Fit Y by X > UnEqual Variances.

Part 2, Parallel tests. Study the table on the next page, which lists the parametric tests based on number of factors and levels, and whether or not those factors are between- or within-subjects. By exploring the features and help available in JMP or SPSS or on the Web, find a way to enter the proper tests for the semiparametric/nonparametric column on the right. The tests to properly place, mixed with tests you should *not* place, are:

> Student's *t*, paired-samples *t*, ANOVA, repeated measures ANOVA, MANOVA, Tukey-Kramer HSD, Chi-Square (χ^2), Fisher's exact, Mann-Whitney *U*, Kruskal-Wallis, Wilcoxon signed-rank, Friedman, Pearson correlation, Spearman correlation, R^2, Kolmogorov-Smirnov *D*, Shapiro-Wilk *W*, Aligned Rank Transform (ART), General Linear Models (GLMs), Generalized Linear Models (GZLMs), Generalized Linear Mixed Models (GLMMs), Generalized Estimating Equations (GEEs).

No. Factors	**No. Levels**	**between-subjects or within-subjects**	**Parametric Test**	**Semiparametric or Nonparametric Equivalent**
1	2	between	independent-samples *t*	4) Mann-Whitney *U*
1	2	within	paired-samples *t*	5) Wilcoxon signed-rank
1	3+	between	one-way ANOVA	6) Kruskal-Wallis
1	3+	within	repeated measures ANOVA	7) Friedman
2+	2+ ea.	between *only* (cannot do within)	*n*-way ANOVA	8) GZLMs
2+	2+ ea.	within (can also do between)	repeated measures ANOVA	9) ART or, 10) GLMMs or, 11) GEEs.

STOP! Complete this table before proceeding to the next page.

No. Factors	**No. Levels**	**between-subjects or within-subjects**	**Parametric Test**	**Semiparametric or Nonparametric Equivalent**
1	2	between	independent-samples *t*	4) Mann-Whitney *U*
1	2	within	paired-samples *t*	5) Wilcoxon signed-rank
1	3+	between	one-way ANOVA	6) Kruskal-Wallis
1	3+	within	repeated measures ANOVA	7) Friedman
2+	2+ ea.	between *only* (cannot do within)	*n*-way ANOVA	8) GZLMs
2+	2+ ea.	within (can also do between)	repeated measures ANOVA	9) ART or, 10) GLMMs or, 11) GEEs.

Part 3, Discrepancies. The above table contains the answers to part 2. Examine it closely.

12. Describe and explain any discrepancies from what you had in your table. Where did the confusions arise and why?

    There were no discrepancies between our table and the answers above.

Part 4, Touring the tests. In this part, the mechanics for performing each of the above nonparametric tests are given for JMP and SPSS. Afterwards, a collection of generic unlabeled data tables are given to you in *nonparam.xlsx*. Your job will be to examine the tables' structures, perform tests for normality (and report the results), and determine what nonparametric tests, and their parametric equivalents, should be conducted on the data, writing up both results and comparing them. All data tables are given as separate worksheets in *nonparam.xlsx*. It is up to you to create JMP or SPSS data tables as necessary.

• Mann-Whitney *U* test • The Mann-Whitney *U* test is the nonparametric equivalent of the independent-samples *t*-test. It analyzes one between-subjects factor with two levels.

JMP can analyze two independent nonparametric samples, but a Mann-Whitney *U* test is not performed; rather, a Wilcoxon rank sums test is used. (This is not to be confused with a Wilcoxon signed-rank test.) SPSS is therefore preferred for the Mann-Whitney *U* test. However, the results from JMP will agree. In JMP, go to Analyze > Fit Y by X, and fill out the Y, Response and X, Factor fields. On the Oneway Analysis report, use the red-arrow menu to select Nonparametric > Wilcoxon Test. The result of the Wilcoxon rank sums test appears in the 2-Sample Test report, where the *S*, *Z*, and *p*-values are of interest.

In SPSS, go to Analyze > Nonparametric Tests > Legacy Dialogs > 2 Independent Samples. Put the response variable in Test Variable List. Put the factor with two levels in Grouping Variable and click Define Groups to indicate each level. Ensure only Mann-Whitney U is checked and click OK. The results are in the Test Statistics report, where the *U*, *Z*, and Asymp. Sig. (2-tailed) values are of interest.

• Wilcoxon signed-rank test • The Wilcoxon signed-rank test is the nonparametric equivalent of the paired-samples *t*-test. It analyzes one within-subjects factor with two levels.

In JMP, go to Analyze > Matched Pairs. Place the two matched response columns in Y, Paired Responses. Click OK. (A paired-samples *t*-test is shown by default.) On the Matched Pairs report, use the red-arrow menu to choose Wilcoxon Signed Rank. The Test Statistic is the *Z* value and the two-sided *p*-value is Prob > |z|.

In SPSS, go to Analyze > Nonparametric Tests > Legacy Dialogs > 2 Related Samples. Put a response in each of the variable slots for Pair 1. Be sure only Wilcoxon is checked and click OK. The *Z* value and *p*-value are given.

• Kruskal-Wallis test • The Kruskal-Wallis test is the nonparametric equivalent of a one-way ANOVA. It analyzes one between-subjects factor with two or more levels.

In JMP, go to Analyze > Fit Y by X. Fill out the dialog and click OK. On the Oneway Analysis report, use the red-arrow menu to select Nonparametric > Wilcoxon Test. A Kruskal-Wallis test is given in the 1-way Test report. It is reported as a Chi-Square (χ^2).

In SPSS, go to Analyze > Nonparametric Tests > Legacy Dialogs > K Independent Samples. Place the response in the Test Variable List, and the factor in the Grouping Variable box. Click Define Range to define the minimum and maximum levels of the factor. Ensure that only Kruskal-Wallis H is checked and click OK. The Test Statistics report shows the results, reported as a Chi-Square (χ^2).

• Friedman test • The Friedman test is the nonparametric equivalent of a repeated measures ANOVA on one factor. It analyzes one within-subjects factor with two or more levels.

Unfortunately, JMP does not provide the Friedman test.

In SPSS, go to Analyze > Nonparametric Tests > Legacy Dialogs > K Related Samples. Transfer the response columns into Test Variables. Ensure that Friedman is checked and click OK. In the Test Statistics report, all rows are of interest. The test is reported as a Chi-Square (χ^2).

• Generalized Linear Models (GZLMs) • Before discussing Generalized Linear Models, let's briefly discuss General Linear Models (GLMs). GLMs are what you have used many times already to perform ANOVAs. In fact, you have used SPSS's Analyze > General Linear Model menu to do so. A GLM is a parametric model that assumes a normal distribution in a continuous response variable. In essence, a GLM is the machinery you've come to regard as simply "an ANOVA," whether repeated measures or not.

Generalized Linear Models (GZLMs) advance GLMs by no longer assuming that a continuous response variable must be normally distributed. Instead, the response may be Poisson, exponential, binomial, and so on. GZLMs are useful for Likert data, event-count data, binary data, interval-censored survival data, and many other types of data.

GZLMs require their user to specify a distribution and a link function. Certain distributions and link functions commonly go together. JMP sets the most common link function as the default whenever you pick a distribution. SPSS enables you to simply choose a model type to set both the distribution and link function at once. For example, a GZLM using the Normal distribution and Identity link function is equivalent to a GLM with only between-subjects factors.

A major limitation of GZLMs is that they assume each observation is independent. In other words, GZLMs can only be used for between-subjects factors, as within-subjects factors are correlated within experimental units (*i.e.*, subjects). Also, GZLMs are not "distribution free" like rank-based methods. Rather, GZLMs simply have the power and flexibility to not require a normal distribution. If a rank-based method is desired instead, use the Aligned Rank Transform (ART) procedure, covered next.

GZLMs can be referred to using certain shorthands depending on their configuration. “Poisson regression” is used on event-count responses and refers to GZLMs with Poisson distribution and Log link functions. “Binomial logistic regression” is used on binary responses and refers to GZLMs with Binomial distributions and Logit link functions. “Nominal logistic regression” is used on nominal responses and refers to GZLMs with Multinomial distributions and Logit link functions.[4] “Ordinal logistic regression” is used on ordinal responses such as Likert scale data and refers to GZLMs with Multinomial distributions and Cumulative Logit link functions. The SPSS 19 help documentation for Generalized Linear Models has a “Model Types” section that explains common distributions and their link functions.

In JMP, Analyze > Fit Model provides access to GZLMs. Nominal logistic regression and ordinal logistic regression are set as the Personality automatically when a nominal or ordinal response is placed into Y. Personality can also be manually set to Generalized Linear Model, which allows access to distributions and link functions other than for nominal and ordinal logistic regression. As usual, you place your model effects into the Construct Model Effects box. Remember: GZLMs cannot handle non-independent observations, so you cannot execute an analysis containing random effects. After running the model, if the Lack of Fit test is significant, it is likely that more model terms are necessary. (You want the highest possible RSquare (U) model fit, which is the R^2 value of the model.) The significance results are shown in the Effect Likelihood Ratio Tests report as a Chi-Square (χ^2).

In SPSS, Analyze > Generalized Linear Models > Generalized Linear Models provides access to GZLMs. The Type of Model tab provides convenient shortcuts for selecting both a distribution and a link function at once. Custom combinations can also be set at the bottom of the tab. For example, to analyze Likert scale data, Choose Ordinal logistic in the Ordinal Response section. On the Response tab, place the response in the Dependent Variable box. On the Predictors tab, place all factors in the Factors box. On the Model tab, create the full factorial model in the Model box. On the Estimation tab, set the Method to Newton-Raphson. On the Statistics tab, choose Likelihood ratio in the Chi-square Statistics box and Profile likelihood in the Confidence Interval Type. Click OK. The Tests of Model Effects box will be of greatest interest.

• Aligned Rank Transform • The Aligned Rank Transform (ART) is a procedure for preparing data in such a way that a parametric ANOVA can be used but the procedure can still be considered nonparametric in nature. The ART is therefore the nonparametric equivalent of either an *n*-way ANOVA or a repeated measures ANOVA on any number of factors. It can analyze between-subjects or within-subjects factors with two or more levels.

Although the regular rank transform (RT) has been studied since the early 1980s, it is known to produce unreliable interaction effects, making it unsuitable for factorial designs. The ART procedure produces accurate interaction effects by first *aligning* the data for each effect, main or interaction, and then ranking it. Data aligning is a procedure by which all potential effects other than the one of interest are subtracted from a continuous response by using marginal means, leaving only the effect of interest present, if there is one.

[4] In SPSS, the Type of Model tab on the dialog available from Analyze > Generalized Linear Models > Generalized Linear Models does not permit using a Multinomial distribution with a Logit link function. Such a combination *is* available in SPSS 19 for Generalized Linear Mixed Models (GLMMs) from Analyze > Mixed Models > Generalized Linear. Owing to its simplicity, JMP is preferred when analyzing nominal responses.

Rather than produce a lengthy explanation of the ART procedure here, you are required to read the following Note from CHI 2011:

> Wobbrock, J.O., Findlater, L., Gergle, D. and Higgins, J.J. (2011). The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11).* Vancouver, British Columbia (May 7-12, 2011). New York: ACM Press, pp. 143-146.

Data aligning is tedious, so the above publication provided a tool to do it for you. Please read the write-up for the ARTOOL program posted at this URL: http://depts.washington.edu/aimgroup/proj/art/. Next, download ARTOOL and save your data to a comma separated file (*.csv). (You can do this in Microsoft Excel.) Run your data through ARTOOL and ensure you understand the output file and its columns. Once the output is created, you can paste it into a statistics package like JMP or SPSS for performing your ANOVA.

As the write-up for ARTOOL says, it is good practice to first perform an appropriate analysis of variance on the "aligned" columns to ensure that all effects except the one for which the column was aligned are indeed "stripped out." Then perform an analysis of variance as appropriate on the "ART" columns, interpreting the result only for the effect matching each column.

• Generalized Linear Mixed Models (GLMMs) • Generalized Linear Mixed Models (GLMMs) are like the linear mixed-effects models you have used in previous modules to carry out analyses of variance. Like those models, GLMMs can contain both fixed and random effects. However, as *generalized* models, they are like GZLMs in not requiring the response variable to be normally distributed. GLMMs can use a variety of distributions and link functions, just like GZLMs, discussed above.

Because GLMMs can have random effects, they remedy the major limitation of GZLMs requiring independent observations and only between-subjects factors. GLMMs can handle correlated measurements and can therefore accommodate within-subjects factors, just as linear mixed models have done for us all along. GLMMs are therefore extremely powerful, and only recently has the computing power necessary to implement them become widely available.

Unfortunately, JMP does not provide access to GLMMs.

With version 19, GLMMs became available for the first time in SPSS. To access GLMMs, go to Analyze > Mixed Models > Generalized Linear.[5] The Data Structure tab enables you to use drag-and-drop to specify how your data table is laid out. Drop the name of your experimental unit (*e.g.*, "Subject") on the Subjects rectangle on the Canvas. Drop any within-subjects factors on the Repeated Measures rectangle on the Canvas. Do this in such a way that the picture developed on the Canvas resembles your tabulated data.

On the Fields & Effects tab, with Target selected at left, choose the Target, which is your response measure. Depending on its modeling type, different radio buttons will be enabled in the box labeled Target Distribution and Relationship (Link) with the Linear Model. Choosing "Linear model" here is

[5] If you are using a GLMM to analyze repeated measures data, the response variable must have its Measure set to Scale or Nominal on the Variable View pane, even if it is ordinal in nature. This limits the utility of GLMMs for ordinal response data, but GLMMs remain useful for analyzing other types of nonparametric data (*e.g.*, event-count data using Poisson regression). For repeated measures on ordinal responses, use Generalized Estimating Equations (GEEs) or the Aligned Rank Transform (ART) procedure.

equivalent to running a mixed-effects model analysis of variance as you have done in previous modules. SPSS provides a description of the currently selected option in the Description box, which can be useful. If you are uncertain as to which option to choose, view the distribution of your response measure to gain an intuition. An easy way to do this is in JMP using the Analyze > Distribution command.

On the Fields & Effects tab, with Fixed Effects selected at left, highlight and drag each fixed effect separately onto the rectangle labeled Main in the Effect builder. Similarly, highlight combinations of factors simultaneously and drop them on the 2-way, 3-way, or * boxes for fixed interaction effects.

On the Fields & Effects tab, with Random Effects selected at left, use the Add Block button to add random effects to the model. In general, intercepts should be used with the Include intercept checkbox. If the experimental unit is the only random effect, choose it from the Subject combination dropdown, include an intercept, and leave the Effect builder otherwise empty.

With these settings in place, click Run. Double-click the Model Summary graphic in the output viewer and maximize the Model Viewer window that opens. Find the fixed effects thumbnail on the left and click it. Set the Style dropdown to Table at the bottom of the window to see the analysis of variance results.

• Generalized Estimating Equations (GEEs) • Like GLMMs, Generalized Estimating Equations (GEEs) extend GZLMs to accommodate correlation among measurements, and therefore can handle within-subjects factors from studies with repeated measures. GEEs allow for correlation but do not require the structure of that correlation to be identified, and instead use estimators to derive the structure from the data. Also, unlike GLMMs, GEEs can accommodate repeated measures on ordinal response variables.

Unfortunately, JMP does not provide access to GEEs.

In SPSS, go to Analyze > Generalized Linear Models > Generalized Estimating Equations. On the Repeated tab, place all columns necessary to uniquely identify the experimental units in the Subject variables box. Place any within-subjects variables in their eponymous box. Set the Structure in the Working Correlation Matrix to reflect the covariance structure across measurements. (The default is "Independent." "Exchangeable" is synonymous with Compound Symmetry; *see* module 8, part 4.) On the Type of Model tab, choose the model to use as you did for GZLMs. On the Response tab, fill in Dependent Variable. On the Predictors tab, add all Factors, which must be nominal, and any Covariates, which must be scalar. On the Model tab, construct the full factorial model in the Model box. On the Estimation tab, change the Method to Newton-Raphson. Click OK to run the GEE analysis. In the output, the Tests of Model Effects table is of most interest. Results are reported as a Wald Chi-Square (χ^2).

For *each* of the six data sets A-F in *nonparam.xlsx*, answer the following questions. You should first duplicate these questions for each data set A-F. Also, since each method named in questions 4-11 will be performed, you may wish to first label *all* of the data tables with the proper test(s), and then begin performing the tests themselves. For the data set requiring the analyses in the last row of the table (ART, GLMM, GEE), perform all three tests on it.

A. Kruskal-Wallis

13. What type of experimental design appears to have been used? Factors? Levels? Within- or between-subjects? Responses? Number of subjects? Is the table in wide- or long-format?

The experiment is a one-way between-subjects design. *Factor* is a between-subjects factor with three levels (1-3). The response is *Y*. There are 30 subjects. The table qualifies as being in both wide-format and long-format. It is in wide-format because all responses for a given subject are in a single row. It is in long-format because there is only one response per row.

14. Using Analyze > Distribution in JMP, perform a test for normality and report the result. Do the data seem to violate the normality assumption?

A significant Shapiro-Wilk *W* test for normality on *Y* indicates that the data are non-normal ($W = .85$, $p < .001$). The data seem to violate the normality assumption.

15. Using Fit Y by X > UnEqual Variances in JMP, perform Levene's test for homogeneity of variances. (Hint: if you have multiple factors, you will have to test for each factor separately.) Report the result of each test. Do the data seem to violate the homoscedasticity assumption?

A significant Levene's test for homogeneity of variances on *Y* by *Factor* indicates that the data seem to violate homoscedasticity ($F(2,27) = 23.83$, $p < .0001$).

16. What *nonparametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

We have a single between-subjects factor with three levels, so a Kruskal-Wallis test should be used. Using JMP, this test results in a significant effect of *Factor* on *Y* ($\chi^2(2,N=30) = 20.47$, $p < .0001$).

17. What *parametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

We have a single between-subjects factor with three levels, so an ANOVA should be used. Using JMP, this test results in a significant effect of *Factor* on *Y* ($F(2,27) = 17.81$, $p < .0001$).

18. How do the results and conclusions compare between your nonparametric and parametric tests for these data?

Both tests agree with strong *p*-values ($p < .0001$) that *Factor* had a significant effect on *Y*.

B. Generalized Linear Model (GZLM)

19. What type of experimental design appears to have been used? Factors? Levels? Within- or between-subjects? Responses? Number of subjects? Is the table in wide- or long-format?

The experiment is a 2×2 between-subjects design. *Factor1* is a between-subjects factor with two levels (1-2). *Factor2* is a between-subjects factor with two levels (1-2). The response is *Likert*. There are 20 subjects. The table qualifies as being in both wide-format and long-format. It is in wide-format because all responses for a given subject are in a single row. It is in long-format because there is only one response per row.

20. Using Analyze > Distribution in JMP, perform a test for normality and report the result. Do the data seem to violate the normality assumption?

A nonsignificant Shapiro-Wilk *W* test for normality on *Likert* indicates that the data are normal ($W = .92$, $p = .09$). The data do not seem to violate the normality assumption.

21. Using Fit Y by X > UnEqual Variances in JMP, perform Levene's test for homogeneity of variances. (Hint: if you have multiple factors, you will have to test for each factor separately.) Report the result of each test. Do the data seem to violate the homoscedasticity assumption?

    A nonsignificant Levene's test for homogeneity of variances on *Likert* by *Factor1* indicates that the data do not seem to violate homoscedasticity (F(1,18) = 0.41, *p* = .53). A nonsignificant Levene's test for homogeneity of variances on *Response* by *Factor2* indicates that the data do not seem to violate homoscedasticity (F(1,18) = 0.04, *p* = .85).

22. What *nonparametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

    We have two between-subjects factors each with two levels, so a Generalized Linear Model (GZLM) using ordinal logistic regression can be used. Using JMP, ordinal logistic regression results in a significant effect of *Factor1* on *Likert* (χ^2(1,*N*=20) = 4.12, *p* < .05), a nonsignificant effect of *Factor2* on *Likert* (χ^2(1,*N*=20) = 0.97, *n.s.*), and a nonsignificant *Factor1*Factor2* interaction (χ^2(1,*N*=20) = 1.91, *n.s.*).

23. What *parametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

    We have two between-subjects factors each with two levels, so an ANOVA should be used. Using JMP, this test results in a marginal result for *Factor1* on *Likert* (F(1,16) = 4.17, *p* = .06), a nonsignificant effect of *Factor2* on *Likert* (F(1,16) = 0.86, *n.s.*), and a nonsignificant *Factor1*Factor2* interaction (F(1,16) = 1.69, *n.s.*).

24. How do the results and conclusions compare between your nonparametric and parametric tests for these data?

    Ordinal logistic regression finds that *Factor1* has a significant effect on *Likert*, although the *p*-value, at .0425, is not much less than .05. The parametric ANOVA found *Factor1* to be nonsignificant with a marginal *p*-value just above .05. The other results agree.

C. Wilcoxon Signed-Rank

25. What type of experimental design appears to have been used? Factors? Levels? Within- or between-subjects? Responses? Number of subjects? Is the table in wide- or long-format?

    The experiment is a one-way within-subjects design. *X* is a within-subjects factor with two levels (1-2). The response name is unknown. There are 20 subjects. The table is in wide-format because all responses for a given subject are in a single row. It is not in long-format because there is more than one response per row.

26. Using Analyze > Distribution in JMP, perform a test for normality and report the result. Do the data seem to violate the normality assumption?

    The two responses were combined in a single column using Tables > Stack before their distribution was analyzed. A significant Shapiro-Wilk *W* test for normality indicates that the data are non-normal (*W* = .78, *p* < .0001). The data seem to violate the normality assumption.

27. Using Fit Y by X > UnEqual Variances in JMP, perform Levene's test for homogeneity of variances. (Hint: if you have multiple factors, you will have to test for each factor separately.) Report the result of each test. Do the data seem to violate the homoscedasticity assumption?

    The two responses were combined in a single column using Tables > Stack before using Fit Y by X > UnEqual Variances. A significant Levene's test for homogeneity of variances indicates that the data seem to violate homoscedasticity (F(1,38) = 14.32, *p* < .001).

28. What *nonparametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

    We have one within-subjects factor with two levels, so a Wilcoxon signed-rank test should be used. Using JMP, this test results in a significant effect of *X* on the response (*Z* = 53.00, *p* < .05).

29. What *parametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

    We have one within-subjects factor with two levels, so a paired-samples *t*-test should be used. Using JMP, this test results in a significant effect of *X* on the response (*t*(19) = 2.32, *p* < .05).

30. How do the results and conclusions compare between your nonparametric and parametric tests for these data?

    Both tests agree with similar *p*-values (*p* < .05) that *X* had a significant effect on the response.

D. Friedman

31. What type of experimental design appears to have been used? Factors? Levels? Within- or between-subjects? Responses? Number of subjects? Is the table in wide- or long-format?

    The experiment is a one-way within-subjects design. *T* is a within-subjects factor with four levels (1-4). The response name is unknown. There are 16 subjects. The table is in wide-format because all responses for a given subject are in a single row. It is not in long-format because there is more than one response per row.

32. Using Analyze > Distribution in JMP, perform a test for normality and report the result. Do the data seem to violate the normality assumption?

    The four responses were combined in a single column using Tables > Stack before their distribution was analyzed. A significant Shapiro-Wilk *W* test for normality indicates that the data are non-normal (*W* = .84, *p* < .0001). The data seem to violate the normality assumption.

33. Using Fit Y by X > UnEqual Variances in JMP, perform Levene's test for homogeneity of variances. (Hint: if you have multiple factors, you will have to test for each factor separately.) Report the result of each test. Do the data seem to violate the homoscedasticity assumption?

    The four responses were combined in a single column using Tables > Stack before using Fit Y by X > UnEqual Variances. A significant Levene's test for homogeneity of variances indicates that the data seem to violate homoscedasticity (F(3,60) = 6.85, *p* < .001).

34. What *nonparametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

    We have one within-subjects factor with four levels, so a Friedman test should be used. Using SPSS, this test results in a significant effect of *T* on the response ($\chi^2(3,N=16) = 30.38$, $p < .001$).

35. What *parametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

    We have one within-subjects factor with four levels, so a repeated measures ANOVA should be used. Using JMP's Analyze > Fit Model > Manova personality, the repeated measures ANOVA results in a sphericity violation ($\chi^2(5,N=16) = 22.22$, $p < .001$) requiring a Greenhouse-Geisser correction. With this correction, the repeated measures ANOVA results in a significant effect of *T* on the response (F(2.1,31.2) = 17.08, $p < .0001$). For completeness, we also used Tables > Stack to produce a long-format table and then ran a mixed-effects model analysis of variance on the data, which is not subject to sphericity concerns, and which also results in a significant effect of *T* on the response (F(3,45) = 17.08, $p < .0001$).

36. How do the results and conclusions compare between your nonparametric and parametric tests for these data?

    All tests agree with $p < .001$ that *T* had a significant effect on the response.

E. Aligned Rank Transform (ART), Generalized Linear Mixed Model (GLMM), Generalized Estimating Equation (GEE)

37. What type of experimental design appears to have been used? Factors? Levels? Within- or between-subjects? Responses? Number of subjects? Is the table in wide- or long-format?

    The experiment is a 2×3 mixed design. *Interface* is a between-subjects factor with two levels (1-2). *Technique* is a within-subjects factor with three levels (1-3). The response is *Errors*. There are 24 subjects. The table is in long-format because there is only one response per row. It is not in wide-format because responses for a given subject occur on multiple rows.

38. Using Analyze > Distribution in JMP, perform a test for normality and report the result. Do the data seem to violate the normality assumption?

    A significant Shapiro-Wilk *W* test for normality on *Errors* indicates that the data are non-normal ($W = .97$, $p = .046$). The data seem to violate the normality assumption.

39. Using Fit Y by X > UnEqual Variances in JMP, perform Levene's test for homogeneity of variances. (Hint: if you have multiple factors, you will have to test for each factor separately.) Report the result of each test. Do the data seem to violate the homoscedasticity assumption?

    A nonsignificant Levene's test for homogeneity of variances on *Errors* by *Interface* indicates that the data do not seem to violate homoscedasticity (F(1,70) = 0.01, $p = .94$). A nonsignificant Levene's test for homogeneity of variances on *Errors* by *Technique* indicates that the data do not seem to violate homoscedasticity (F(2,69) = 0.09, $p = .91$).

40. What *nonparametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

    We have two factors, one of which is within-subjects. First, the Aligned Rank Transform (ART) method can be used. Using JMP on the ART data produced by ARTool, we see a significant effect of *Interface* on *Errors* (F(1,22) = 9.69, *p* < .01), a nonsignificant effect of *Technique* on *Errors* (F(2,44) = 0.58, *n.s.*), and a nonsignificant *Interface**Technique* interaction (F(2,44) = 0.54, *n.s.*).

    Second, a Generalized Linear Mixed Model (GLMM) can be used. The Analyze > Distribution > Fit Distribution > Poisson > Goodness of Fit process in JMP was used to show the data can be considered as coming from a Poisson distribution (Pearson $\chi^2(1,N=72) = 5.07$, $p = .99$). Therefore, the GLMM was modeled in SPSS using Poisson regression with a Log link function and Errors as a scalar measure. Results show a trend for *Interface* on *Errors* (F(1,66) = 3.25, *p* = .08), a nonsignificant effect of *Technique* on *Errors* (F(2,66) = 1.04, *n.s.*), and a nonsignificant *Interface**Technique* interaction (F(2,66) = 1.00, *n.s.*).

    Third, a Generalized Estimating Equation (GEE) can be used in SPSS with an ordinal logistic model, Errors as an ordinal measure, and an Exchangeable covariance structure. Results show a significant effect of *Interface* on *Errors* (Wald $\chi^2(1,N=72) = 4.97$, $p < .05$), a nonsignificant effect of *Technique* on *Errors* (Wald $\chi^2(2,N=72) = 0.75$, *n.s.*), and a nonsignificant *Interface**Technique* interaction (Wald $\chi^2(2,N=72) = 1.76$, *n.s.*).

41. What *parametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

    We have two factors, one of which is within-subjects, so a repeated measures ANOVA should be used. Using JMP, we first create a wide-format table using Tables > Split, putting *Technique* in Split By, *Errors* in Split Columns, and *Subject* and *Interface* in Group. Using JMP's Analyze > Fit Model > Manova personality, the repeated measures ANOVA results in a significant effect of *Interface* on *Errors* (F(1,22) = 6.09, *p* < .05), a nonsignificant effect of *Technique* on *Errors* (F(2,44) = 0.77, *n.s.*), and a nonsignificant *Interface**Technique* interaction (F(2,44) = 0.90, *n.s.*). For completeness, we also ran a mixed-effects model analysis of variance on the original long-format table, which results in exactly the same statistical findings.

42. How do the results and conclusions compare between your nonparametric and parametric tests for these data?

    All tests show a significant effect of *Interface* on *Errors* except the GLMM, which resulted in a trend (*p* = .08). All tests agree on the nonsignificance of *Technique* on *Errors* and *Interface**Technique* on *Errors*.

F. Mann-Whitney *U*

43. What type of experimental design appears to have been used? Factors? Levels? Within- or between-subjects? Responses? Number of subjects? Is the table in wide- or long-format?

    The experiment is a one-way between-subjects design. *Factor* is a between-subjects factor with two levels (1-2). The response is *Response*. There are 20 subjects. The table qualifies as being in both wide-format and long-format. It is in wide-format because all responses for a given subject are in a single row. It is in long-format because there is only one response per row.

44. Using Analyze > Distribution in JMP, perform a test for normality and report the result. Do the data seem to violate the normality assumption?

A significant Shapiro-Wilk *W* test for normality on *Response* indicates that the data are non-normal (*W* = .85, *p* < .01). The data seem to violate the normality assumption.

45. Using Fit Y by X > UnEqual Variances in JMP, perform Levene's test for homogeneity of variances. (Hint: if you have multiple factors, you will have to test for each factor separately.) Report the result of each test. Do the data seem to violate the homoscedasticity assumption?

A significant Levene's test for homogeneity of variances on *Response* by *Factor* indicates that the data seem to violate homoscedasticity (F(1,18) = 15.64, *p* < .001).

46. What *nonparametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

We have a single between-subjects factor with two levels, so a Mann-Whitney *U* test should be used. Using SPSS, this test results in a nonsignificant effect of *Factor* on *Response* (*U* = 34.00, *N* = 20, *n.s.*).

47. What *parametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.

We have a single between-subjects factor with two levels, so an independent-samples *t*-test should be used. Using JMP, this test results in a significant effect of *Factor* on *Response* (*t*(18) = 2.29, *p* < .05).

48. How do the results and conclusions compare between your nonparametric and parametric tests for these data?

The nonparametric test did not reject the null hypothesis of no effect, but the parametric *t*-test did reject the null hypothesis to conclude *Factor* had a significant effect on *Response*. The nonparametric *p*-value was .226, which is quite a bit higher than the *t* test *p*-value of .035. We note that because Levene's test indicated a violation of homoscedasticity, the *t*-test we should use would *not* assume equal variances, but even an unequal-variance *t*-test results in a significant *p*-value of .045. These results demonstrate the common situation where nonparametric tests are not as powerful as their parametric counterparts. Of course, this fact does not warrant the use of a parametric test, especially in light of normality and homoscedasticity violations for these data.

10. Categories, counts and proportions

prefs2., prefs3.*, liked.jmp, liked.freq.jmp, likedwthn.sav, comm.jmp, ptgbtwn.jmp, ptgmixed.jmp*

For some studies in human-computer interaction, especially field studies where behavior is logged, *count* measures are common. Take, for example, the deployment of a novel text entry method to a selection of participants "in the wild" for a period of 10 weeks. From raw field data, it would be very difficult to measure words per minute or error rates, because a researcher cannot know the intention of the participant. What were they *trying* to write? Did someone interrupt the participant? Did the phone ring mid-sentence? Did the participant change his mind mid-sentence? However, log files could show *counts* of events, such as letters, numbers, spaces and backspaces, use of cursor keys, and so on. If, say, a word prediction system was installed, we could count how many characters were entered via this feature, and thus approximate the character savings from word prediction.

Counts also appear in data where *events* are being coded from videos or interviews. If participants are videotaped for later coding, that coding often entails counting the number of times a certain behavior occurred.

Counts become *rates* when measured per unit time. In our example of fielding a text entry method, we could compute certain rates, like the number of characters entered per day. Words per minute is also a rate, but it has a formal definition which requires a laboratory setting.

A count can sometimes be turned into a *proportion*. If we can answer, "how many letters were entered?" we can sometimes also answer, "of all characters entered, what was the proportion of letters?" By definition, to have a proportion, we must be able to define a "denominator," *i.e.*, the maximum possible value in which our count occurs.

Counts, events, rates, and proportions are often not suitable to ANOVA. You learned in module 6 about normality tests and module 9 about nonparametric tests. This module continues your nonparametric training but with methods specifically designed to evaluate counts that fall into various response categories. Some of these methods are new, and some are those from these prior modules.

Part 1, One-sample Chi-Square (χ^2) test of proportions. You have seen the Chi-Square (χ^2) statistic before, *e.g.*, when reporting the results of a sphericity test, a Friedman test, or ordinal logistic regression. However, there are also Chi-Square tests of proportions. A *one-sample* test compares counts occurring among *C* unordered categories for whether those counts differ from hypothesized values, often the values set by *chance* (*i.e.*, equal probabilities for each category).

Open *prefs2.jmp*, which contains subjective preferences from 30 subjects who participated in a within-subjects study of the EdgeWrite versus Graffiti text entry methods. Each subject performed with both text entry methods, and at the end of the study, chose which method they preferred. Note that although the study itself had repeated measures, there is only one response in *prefs2.jmp* from each subject, making the data suitable to analysis with a Chi-Square test.

Although the *Subject* column is often omitted in such tables, it is a good idea to keep it for clarity. The question of interest can be framed as whether the proportion of subjects who liked one method or the other was significantly different than chance. To answer this question, go to Analyze > Distribution. Put *Preference* in Y, Columns and click OK. Using the Preference red-arrow menu, choose Test Probabilities. To test with equal weighting, one only needs to enter equal values into the two Hypoth Prob boxes. (If the entered values do not sum to 1.00, JMP will automatically rescale them according to their relative proportion.) For ease, simply enter "1" into each box. Be sure the first radio button option is selected to perform a two-sided test, and click Done.

1. What happened to the Hypoth Prob values after you clicked Done? Why did this happen?

   The two Hypoth Prob values changed from 1 to 0.5. This happened because the hypothesized probabilities must add to 1.00, and since each value had been given equal weight, they retain this equal relationship when being made to sum to 1.00.

2. Report the result from the Pearson one-sample Chi-Square test of proportions. What can you conclude?

   A one-sample Pearson Chi-Square test of proportions shows that EdgeWrite was significantly preferred to Graffiti ($\chi^2(1,N=30) = 4.80$, $p < .05$). EdgeWrite was preferred by 70% of respondents to Graffiti's 30%.

The *prefs2.jmp* file is in long-format, with one row for each subject's response. Another common table format for Chi-Square tests is a "count table," "frequency table," or "weighted cases table." To make this table, go to Tables > Summary, highlight *Preference* on the left, and click Group. Then click OK. The resulting table will show, for each preference, how many rows matched it, giving you the count for that response.

Now go to Analyze > Distribution again, putting *Preference* in Y, Columns and *N Rows* in Freq. Click OK and answer the following questions.

3. How does the resulting report compare to the one you viewed for questions 1-2?

   The resulting report is identical to the one for questions 1-2.

4. Conduct the one-sample Chi-Square test of proportions, report the result, and compare it to that from question 2.

   A one-sample Pearson Chi-Square test of proportions shows that EdgeWrite was significantly preferred to Graffiti ($\chi^2(1,N=30) = 4.80$, $p < .05$). This is the same result as for question 2.

Open *prefs2.sav* in SPSS to conduct a one-sample Chi-Square analysis on the long-format table. Go to Analyze > Nonparametric Tests > Legacy Dialogs > Chi-square. Transfer *Preference* into Test Variable List. Ensure All categories equal is selected in Expected Values. (This setting causes SPSS to do what you achieved by typing "1" into all Hypoth Prob cells in JMP.) Click OK.

5. Report the result for this one-sample Chi-Square test of proportions.

   A one-sample Chi-Square test of proportions shows that EdgeWrite was significantly preferred to Graffiti ($\chi^2(1,N=30) = 4.80$, $p < .05$).

6. Comparing your results to those from question 2, determine whether SPSS is using a Likelihood Ratio Chi-Square result or a Pearson Chi-Square result. How do you know?

   SPSS appears to be using the Pearson Chi-Square result because the *p*-value of .028 agrees with the Pearson *p*-value in JMP. Meanwhile, the Likelihood Ratio Chi-Square result in JMP is about .026.

Open *prefs3.jmp* in JMP and see fictitious results for a within-subjects study of three text entry methods, EdgeWrite and Graffiti, as before, and now also ShapeWriter. The counts of responses can now fall into one of three categories instead of just two. Following the same procedures as above, answer the following:

7. Using the long-format table, determine whether the proportions of preferences among the three text entry methods are significantly different from chance.

   The proportions of responses of subjects preferring EdgeWrite and ShapeWriter were the same at about 43%, while only about 13% of subjects preferred Graffiti. A one-sample Likelihood Ratio Chi-Square test of proportions shows that these proportions were significantly different than chance ($\chi^2(2,N=30) = 6.31$, $p < .05$), indicating subjects significantly preferred EdgeWrite and ShapeWriter to Graffiti. However, a one-sample Pearson Chi-Square test of proportions exhibits only a trend ($\chi^2(2,N=30) = 5.40$, $p = .067$), casting some doubt on the trustworthiness of this finding.

8. Transform the long-format table into a weighted cases table. Describe what this table looks like.

   Using Tables > Summary, placing *Preference* in Group, and clicking OK results in a weighted cases table of two columns, *Preference* and *N Rows*, and three rows, one each for each text entry method. The table shows that in the long-format table, EdgeWrite had 13 rows, Graffiti had 4 rows, and ShapeWriter had 13 rows.

9. Using the weighted cases table, perform the Chi-Square test again and report your results. How do they compare to those from question 7?

   Using Analyze > Distribution with *Preference* in Y, Columns and *N Rows* in Freq, and then the Test Probabilities red-arrow menu option with equal weighting, results in the same findings as in question 7. Specifically, the one-sample Likelihood Ratio Chi-Square test of proportions is significant ($\chi^2(2,N=30) = 6.31$, $p < .05$) while the one-sample Pearson Chi-Square test of proportions shows a trend ($\chi^2(2,N=30) = 5.40$, $p = .067$).

Open *prefs3.sav* in SPSS.

10. Perform a one-sample Chi-Square test of proportions on the long-format table. Report your results. How do they compare to those from question 7?

    Using Analyze > Nonparametric Tests > Legacy Dialogs > Chi-square with *Preference* in Test Variable List results in a one-sample Pearson Chi-Square test of proportions that shows a nonsignificant trend ($\chi^2(2,N=30) = 5.40$, $p = .067$). These results are identical to the Pearson Chi-Square results from question 7.

Part 2, Chi-Square (χ^2) test of proportions with independent groups. A *two-sample* Chi-Square test of proportions compares counts, but now coming from two independent groups. A *three-sample* test compares counts coming from three independent groups. And so on. Note that Chi-Square tests always presume the independence of responses, so they are not appropriate for multiple responses obtained from one subject.

In general, a Chi-Square test using independent groups compares whether *C* counts spread among *R* unordered categories are significantly different. The variables *R* and *C* are chosen because such data forms a *R*ow × *C*olumn table called a "contingency table." And again, an assumption of the test is that responses are independent, *i.e.*, each subject (or other experimental unit) can be counted only once.[6]

Open *liked.jmp*, which is a long-format table containing data from a 7-point Likert scale from *disliked* (1) to *liked* (7). (Note that these are ordered categories, but to begin with, we will simply regard them as seven separate unordered categories.) Two text entry methods have been judged after use, each by 32 different subjects who this time only used one text entry method or the other, constituting independent groups.

Now open *liked.freq.jmp*, a weighted cases table. Note the *Count* column indicating how many subjects marked the corresponding Likert value (1-7) for each text entry method.

11. Describe how both *liked.jmp* and *liked.freq.jmp* show the same data. (Do they? Are you sure? How can you verify this?)

    The data are the same; the frequencies of Likert scores from 1-7 are the same in both tables. We can verify this by creating a formula column called *Mult* in *liked.freq.jmp* defined as *Disliked-Liked*Count* and then using Tables > Summary to sum *Mult* by *Method*. On *liked.jmp*, we can use Tables > Summary to sum *Disliked-Liked* by *Method*. The sum for EdgeWrite is 151. The sum for Graffiti is 120. The one type of information in *liked.jmp* lost in *liked.freq.jmp* is *which* subject marked each Likert score for each method.

12. Verify that the mean Likert response in both tables is the same for EdgeWrite and Graffiti. What are these means? How did you calculate the means for each method in *liked.jmp*? How did you do so in *liked.freq.jmp*? (Hint: the long-format table method means can be calculated in one step; it will require multiple steps, perhaps involving the creation of formula columns, to do so for the weighted cases table.)

    The mean response for EdgeWrite is 4.72. The mean response for Graffiti is 3.75. The mean responses for each method in *liked.jmp* can be calculated using Tables > Summary, placing Mean(*Disliked-Liked*) in Statistics and *Method* in Group. The mean responses for each method in *liked.freq.jmp* can be calculated by first creating a new column called *Mult* defined as *Disliked-Liked*Count*. Then Tables > Summary is used to sum *Mult* by *Method* and to sum *Count* by *Method*. In the new summary table, a new formula column is created defined as *Sum(Mult) / Sum(Count)*, which gives the average responses per method.

Now perform a Chi-Square test of proportions on *liked.freq.jmp*. We have only one factor, *Method*, so we can first use Analyze > Fit Y by X. Put *Disliked-Liked* in Y, Response; *Method* in X, Factor; and *Count* in Freq. The Tests report shows the *N*, *df*, χ^2, and *p*-value necessary to report your Chi-Square result. (Here we will use the Likelihood Ratio result, but the Pearson result is fine in general, too.)

[6] If correlated (non-independent) measures are taken, McNemar's test may be used, but it is only applicable to 2×2 tables, and thus has limited utility. McNemar's test is not covered in these modules. See http://en.wikipedia.org/wiki/McNemar%27s_test.

13. What is the result of this test? What does the result mean? Report the result and interpret the findings.

    A Chi-Square test on how liked or disliked EdgeWrite and Graffiti were exhibits no significant difference ($\chi^2(6,N=64) = 6.01$, *n.s.*), indicating that the proportions of responses along a 7-point Likert scale were not significantly different.

We can also perform the same test using Analyze > Fit Model, which would allow for more than one factor if we required it. However, whereas the Fit Y by X platform regards a nominal or ordinal response variable as equivalently categorical (unordered), the Fit Model platform changes its analysis based on this modeling type. Therefore, for now, flip the modeling type of *Disliked-Liked* to nominal. Then use Analyze > Fit Model with Nominal Logistic for Personality.

14. Examine the output and report the results. Do they match the results from question 13?

    A Chi-Square test on how liked or disliked EdgeWrite and Graffiti were exhibits no significant difference ($\chi^2(6,N=64) = 6.01$, *n.s.*), indicating that the proportions of responses along a 7-point Likert scale were not significantly different. These results exactly match those from question 13.

Now turn to *liked.jmp*. As before, flip *Disliked-Liked* to nominal for the sake of performing a Chi-Square test. Use both Analyze > Fit Y by X and Analyze > Fit Model on the long-format table. Note that there is no Freq parameter required when using the long-format table.

15. Examine the output and report the results from Fit Y by X. Do they match the results from questions 13-14?

    A Chi-Square test on how liked or disliked EdgeWrite and Graffiti were exhibits no significant difference ($\chi^2(6,N=64) = 6.01$, *n.s.*), indicating that the proportions of responses along a 7-point Likert scale were not significantly different. These results exactly match those from questions 13-14.

16. Examine the output and report the results from Fit Model. Do they match the results from questions 13-15?

    A Chi-Square test on how liked or disliked EdgeWrite and Graffiti were exhibits no significant difference ($\chi^2(6,N=64) = 6.01$, *n.s.*), indicating that the proportions of responses along a 7-point Likert scale were not significantly different. These results exactly match those from questions 13-15.

Note when using Analyze > Fit Y by X, two warnings are produced due to low cell counts. A rule of thumb for using the Likelihood Ratio Chi-Square is that the average cell count is 5 or more. A rule of thumb for using the Pearson Chi-Square is that 80% of cells have a count of 5 or more. When this is not the case *and when 2×2 designs are analyzed*, Fisher's Exact test can be used instead. If we had only two Likert values (say, "disliked" and "liked") along with two methods (EdgeWrite and Graffiti), then JMP would automatically perform Fisher's Exact Test for us. We'll see this in part 3, below.

Because semantically, we know that these data represent ordered Likert scores, we can do better than a mere Chi-Square test of proportions of unordered categories. Return to *liked.freq.jmp* and flip *Disliked-Liked* back to ordinal. Use Analyze > Fit Model to run ordinal logistic regression on this ordinal response.

17. Examine the output and report the results. How do they compare to those for questions 13-16?

    Ordinal logistic regression on how liked or disliked EdgeWrite and Graffiti were exhibits a significant difference ($\chi^2(1,N=64) = 5.00$, $p < .05$), indicating that EdgeWrite was liked significantly more than Graffiti (4.72 *vs.* 3.75 on a 1-7 scale, with 7 as most liked). These results for ordinal responses do *not* match those from question 13-16, where no significant difference in unordered proportions was found.

Similarly, on *liked.jmp*, flip *Disliked-Liked* back to ordinal, and use Analyze > Fit Model to analyze this long-format data using ordinal logistic regression.

18. Examine the output and report the results. How do they compare to those for questions 13-17?

    Ordinal logistic regression on how liked or disliked EdgeWrite and Graffiti were exhibits a significant difference ($\chi^2(1,N=64) = 5.00$, $p < .05$), indicating that EdgeWrite was liked significantly more than Graffiti (4.72 *vs.* 3.75 on a 1-7 scale, with 7 as most liked). These results do not match those from question 13-16, where no significant difference in unordered proportions was found, but do match those from question 17 exactly.

Now we will perform a Chi-Square test of proportions in SPSS using a weighted cases table. Open a blank data table in SPSS and replicate the data from *liked.freq.jmp*. Provide appropriate column names and values, with "EdgeWrite" as 1 and "Graffiti" as 2. (Be sure View > Value Labels is turned on.) For your variable modeling types, *Method* should be nominal, *Disliked2Liked* should be ordinal, and *Count* should be scalar. You may wish to save this table as *liked.freq.sav*.

Now perform a Chi-Square test of proportions. First go to Data > Weight Cases, choose Weight cases by, and set *Count* as the Frequency Variable. Then use Analyze > Descriptive Statistics > Crosstabs and for Row(s) put *Method*, for Column(s) put *Disliked2Liked*. Click Statistics and check Chi-square. Click Continue. Click OK.

19. Examine the output from the Chi-Square Tests report. Report the result and conclusion of this Chi-Square test.

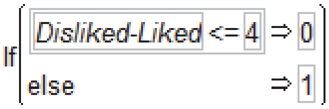
    A Chi-Square test on how liked or disliked EdgeWrite and Graffiti were exhibits no significant difference ($\chi^2(6,N=64) = 6.01$, *n.s.*), indicating that the proportions of responses along a 7-point Likert scale were not significantly different. These results exactly match those from questions 13-16.

20. In light of the results you saw from JMP when *Disliked-Liked* was nominal (questions 13-16) and when it was ordinal (questions 17-18), what can you conclude about how SPSS treated *Disliked2Liked*, which you modeled as ordinal, in this procedure?

    Despite modeling *Disliked2Liked* as ordinal, the Analyze > Descriptive Statistics > Crosstabs procedure in SPSS, like the Analyze > Fit Y by X procedure in JMP, treats ordinal variables as nominal, *i.e.*, as having unordered categories, at least when the Chi-square option is checked. This is adequate for responses that fall into unordered categories, but may not be adequate when analyzing responses that fall into ordered categories.

Part 3, Fisher's Exact test. As noted above, when cell counts are low, the Chi-Square test is suspect. In the case of 2×2 designs, Fisher's Exact test can be used. JMP produces this test automatically in its Analyze > Fit Y by X platform when it is warranted.

First, we need to turn our *Disliked-Liked* ordinal outcome into a dichotomous one in *liked.jmp*. Define a new column called *Liked?*[7] and set it equal to the following formula:



```
If ( Disliked-Liked <= 4  ⇒ 0
     else                 ⇒ 1 )
```

.

21. Semantically, why is the value "4" the correct value in the above formula for defining *Liked?*.

   Setting *Liked?* to be false for values 1-4 makes sense because the original *Disliked-Liked* Likert scale had disliking from 1-3 and neutral at 4. Since feeling neutral cannot be said to be liking, 4 should be included in the range for *Liked?*=false. This leaves *Liked?*=true to be *Disliked-Liked* values 5-7, which were the values for actual liking. Note that if our new column was *Disliked?*, then the neutral value 4 should be included in *Disliked?*=false, since neutral is not actual disliking any more than it is actual liking.

Before we perform our tests, we need to define the value ordering for the *Liked?* column. The first value will determine how the odds ratio is constructed. Think of this as determining whether we ask about the relative odds of having 0/1 *vs*. 1/0. Semantically, we'd like to define the question as 1/0 because we are naming the column *Liked?*. Therefore, for the *Liked?* column, go to Column Info > Column Properties > Value Ordering and move the "1" above the "0". Then click OK.

Having defined the value order, use Analyze > Fit Y by X on *Liked?* and examine the results.

22. What is the Likelihood Ratio Chi-Square result for this test? Report and interpret the finding.

   A Chi-Square test on whether EdgeWrite and Graffiti were liked or disliked exhibits a significant difference ($\chi^2(1,N=64) = 4.06$, $p < .05$), indicating that EdgeWrite was liked more than Graffiti (19/32 *vs.* 11/32 responses).

23. How does this result compare to the findings from using the *Disliked-Liked* response as a nominal outcome (questions 13-16, 19-20)? as an ordinal outcome (questions 17-18)?

   This result showed a significant difference in favor of EdgeWrite, whereas the results for the unordered *Disliked-Liked* response categories did not (questions 13-16, 19-20). However, the results for *Disliked-Liked* as an ordinal outcome *did* show a significant difference in favor of EdgeWrite (questions 17-18), and agree with this finding here.

24. What is the Fisher's Exact test result? Report and interpret the finding.

   Fisher's Exact test on whether EdgeWrite and Graffiti were liked or disliked exhibits a marginal result (p = .08), indicating a degree of inconclusiveness as to whether or not EdgeWrite was liked more than Graffiti (19/32 *vs.* 11/32 responses).

25. Is a Fisher's Exact test warranted in this case? Why or why not?

[7] In JMP, I prefer to use a question mark (?) at the end of dichotomous variables, whether they are independent factors or dependent measures. This serves as a visual reminder that the variable is dichotomous. SPSS does not allow this symbol in column names.

Fisher's Exact test is warranted in lieu of a Likelihood Ratio Chi-Square test when the average cell count is less than 5. The average cell count in this case is (19+13+11+21) / 4 = 16, so a Fisher's Exact test is not necessary.

Also, analyze *Liked?* using nominal logistic regression with Analyze > Fit Model. (Recall that you would have no option of using Fit Y by X if you had a factorial design.)

26. Report the result of the Chi-Square test. Does it agree with what you found in question 22?

Nominal logistic regression for whether EdgeWrite and Graffiti were liked or disliked exhibits a significant difference ($\chi^2(1,N=64) = 4.06$, $p < .05$), indicating that EdgeWrite was liked more than Graffiti (19/32 *vs.* 11/32 responses). This result exactly matches the one from question 22.

Part 4, Binomial logistic regression. Recall from module 9, part 4 that binomial logistic regression is a Generalized Linear Model (GZLM) with a binomial distribution and logit link function designed to handle dichotomous (0/1) outcomes. Recall that GZLMs cannot handle within-subjects factors, but Generalized Linear Mixed Models (GLMMs) and Generalized Estimating Equations (GEEs) can, effectively extending and remedying GZLMs in this way.

For practice, we'll use two new tables to perform binomial logistic regression on the data in *liked.jmp*. (In reality, you already performed binomial logistic regression for question 26 on the long-format table because *Liked?* is dichotomous. These are two new table formats.)

The first table we need is a weighted cases table. Create this table from *liked.jmp* such that you have the following columns: *Method*, *Liked?*, and *Count*. Define the value ordering on *Liked?* such that 1 is first.

27. How did you create the new table? How many rows are in the table?

The new table was created by using Tables > Summary on our *liked.jmp*, placing *Method* and *Liked?* in Group and nothing in Statistics. The *N Rows* column is then renamed *Count*. On the *Liked?* column, Column Info > Column Properties > Value Ordering is used to change the order to 1, 0. The number of rows in this table is just four.

28. What are the modeling types of each variable?

Method is a nominal variable, *Liked?* is a nominal (dichotomous) variable, and *Count* is a continuous variable.

29. What are the counts for the four combinations of *Method* = {EdgeWrite, Graffiti} and *Liked?* = {1, 0}?

The counts are: (EdgeWrite, 1) = 19, (EdgeWrite, 0) = 13, (Graffiti, 1) = 11, (Graffiti, 0) = 21.

You may wish to save this table as *likedornot.jmp*. Now perform your analysis using Analyze > Fit Model. You should by now understand how to fill out the model dialog. Don't forget to place your *Count* column in Freq.

30. What is the outcome of this test? Report and analyze your finding.

    Binomial logistic regression on whether EdgeWrite and Graffiti were liked or disliked exhibits a significant difference ($\chi^2(1,N=64) = 4.06$, $p < .05$), indicating that EdgeWrite was liked more than Graffiti (19/32 *vs.* 11/32 responses).

31. How does it compare to the results from questions 22 and 26?

    This result exactly matches the one from questions 22 and 26.

The second type of table, which we'll call *likedornot.gzlm.jmp*, allows us to explicitly choose a Generalized Linear Model for binomial logistic regression. This table will be similar to *likedornot.jmp* but has just two rows and an additional column named *Trials*, indicating the total trials. So its columns should be *Method*, *Liked?*, *Count*, and *Trials*. The *Liked?* column is superfluous at this point, and should just have "1" in each row. (We'll keep it for clarity.) *Count* should now have the total number of 1's for each method. *Trials* has the total number of trials. In essence, in moving from *likedornot.jmp* to *likedornot.gzlm.jmp*, we've removed the two rows that contained *Liked?* = 0, and replaced them with the *Trials* column that identifies the total number of responses.

32. How did you create the new table? What are the *Count* and *Trials* values for the EdgeWrite row? for the Graffiti row?

    The new table was created by using Tables > Summary on our *liked.jmp* table, placing *Method* and *Liked?* in Group and nothing in Statistics. Then we rename *N Rows* to *Count* and create a new column named *Trials*. Finally, we delete the row for (*Method* = EdgeWrite, *Liked?* = 0) and add its *Count* value to the *Count* value for (*Method* = EdgeWrite, *Liked?* = 1), placing this value, 32, in *Trials*. Similarly, we delete the row for (*Method* = Graffiti, *Liked?* = 0) and add its *Count* value to the *Count* value for (*Method* = Graffiti, *Liked?* = 1), placing this value, 32, in *Trials*. Thus, the *Count* value for EdgeWrite is 19 and the *Trials* value is 32. The *Count* value for Graffiti is 11 and the *Trials* value is 32.

When your table is ready, go to Analyze > Fit Model. For Personality, select Generalized Linear Model. For Distribution, select Binomial. (Leave the link function as Logit.) Add *Method* to your model effects box. Place both *Count* and *Trials* in the Y box. The first value here identifies the number of 1's, while the second value identifies the total number of trials. Click Run Model.

33. Examine the Effect Test box. Report the result, being careful to report the correct *N* in the Chi-Square result.

    Binomial logistic regression on whether EdgeWrite and Graffiti were liked or disliked exhibits a significant difference ($\chi^2(1,N=64) = 4.06$, $p < .05$), indicating that EdgeWrite was liked more than Graffiti (19/32 *vs.* 11/32 responses).

34. How does the result compare to those from questions 22, 26, and 30?

    This result exactly matches the one from questions 22, 26, and 30.

It turns out we can also use our *liked.jmp* data to perform explicit binomial logistic regression. Return to that table and choose Analyze > Fit Model, setting Personality to Generalized Linear Model, Distribution as Binomial, Y as *Liked?*, and *Method* as the model effect. Click Run Model.

35. Report the result of this analysis.

    Binomial logistic regression on whether EdgeWrite and Graffiti were liked or disliked exhibits a significant difference ($\chi^2(1,N=64) = 4.06$, $p < .05$), indicating that EdgeWrite was liked more than Graffiti (19/32 *vs.* 11/32 responses).

36. How does the result compare to questions 22, 26, 30, and 33?

    This result exactly matches the one from questions 22, 26, 30, and 33.

As one last exploration, we'll also analyze *liked.jmp* in SPSS. Open a blank spreadsheet in SPSS and paste the data from *liked.jmp* into it. On the Variable View, set variable names to *Subject* (Nominal), *Method* (Numeric, Nominal), *Disliked2Liked* (Ordinal), and *LikedP*[8] (Nominal). For *Method*, define Values as 1 = "EdgeWrite" and 2 = "Graffiti". Then enter these values in your spreadsheet (*Method* rows 1-32 get "1", 33-64 get "2"). (Be sure View > Value Labels is turned on.)

When your table is ready, go to Analyze > Generalized Linear Models > Generalized Linear Models. On Type of Model, select Binary logistic. On Response, place *LikedP* in the Dependent Variable box. On Predictors, place *Method* in the Factors box. On Model, place *Method* in the Model box. On Estimation, set Method to be Newton-Raphson. On Statistics, choose Likelihood ratio for Chi-Square Statistics and Profile likelihood for Confidence Interval Type. Then click OK.

37. Report the results. How do the findings compare to those from questions 22, 26, 30, 33, and 35?

    Binomial logistic regression on whether EdgeWrite and Graffiti were liked or disliked exhibits a significant difference ($\chi^2(1,N=64) = 4.06$, $p < .05$), indicating that EdgeWrite was liked more than Graffiti (19/32 *vs.* 11/32 responses). This result exactly matches the one from questions 22, 26, 30, 33, and 35.

Part 5, Logistic regression with repeated measures. In parts 2 and 4, above, you analyzed Likert scale data and dichotomous data with Generalized Linear Models (GZLMs). Recall that GZLMs cannot handle repeated measures, *i.e.*, all factors must be between-subjects, not within-subjects. Generalized Linear Mixed Models (GLMMs) and Generalized Estimating Equations (GEEs) remedy these limitations. Both of these analyses are unavailable in JMP but are available in SPSS as of version 19.

In SPSS, open *likedwthn.sav*, which contains the same data as *liked.jmp* except that now there are only 32 subjects, not 64.

38. Is the data table *likedwthn.sav* in long format or wide format? How do you know?

    This data table is in long format. We know because there exists only one measure per subject per row (per dependent variable—the fact that two response columns, *Disliked2Liked* and *LikedP*, exist is irrelevant; the issue is whether multiple measures per subject are in one row for different levels of the *same dependent*

[8] SPSS does not allow question marks (?) in variable names. The use of suffix "P" harkens back to Lisp programming, where Boolean variables, by convention, are given this suffix. Nick Parlante (http://www-cs-faculty.stanford.edu/~nick/), computer science lecturer at Stanford University, tells a story about querying his roommates for their pizza orders by asking, "Pizza-P?"

variable, and that is not the case here). Further, we know the table is in long format because subjects are repeated across multiple rows.

39. For each factor in *likedwthn.sav*, identify whether it is between-subjects or within-subjects and say how you know.

    Method is the one fixed independent factor in the table, and it is a within-subjects factor, giving rise to repeated measures data.

40. How many dependent measures are shown in *likedwthn.sav*? What are their names and types?

    There are two dependent measures shown in the table, *Disliked2Liked*, an ordinal response, and *LikedP*, a nominal dichotomous response, which just codes the former into a binary outcome.

Recall that liked linear mixed-effects models, Generalized Linear Mixed Models (GLMMs) must have a scalar outcome for repeated measures analyses. While we could turn *Disliked2Liked* from "Ordinal" to "Scale" in the Variable View, we can leave it as an ordinal measure if we use Generalized Estimating Equations (GEEs) to carry out our analysis. You may wish to refer back to GEEs from module 9, part 4 before continuing.

41. Run a GEE on *Disliked2Liked* as an ordinal measure in *likedwthn.sav*. What are the results, and how do they compare to the between-subjects analyses of the ordinal response using GZLMs from questions 17-18?

    Using a Generalized Estimating Equations (GEE) with an Independent covariance structure shows *Method* had a marginal effect on the *Disliked2Liked* Likert responses (Wald $\chi^2(1,N=64) = 3.70$, $p = .055$). These results are actually "weaker" *p*-value-wise than when *Method* was analyzed as a between-subjects factor using a Generalized Linear Model (GZLM) in questions 17-18, where it was significant ($\chi^2(1,N=64) = 5.00$, $p = .025$).

For question 37, above, you analyzed *LikedP* as a dichotomous response using binomial logistic regression, *i.e.*, using Generalized Linear Models (GZLMs) and the Binary logistic model type. Now use Generalized Linear Mixed Models (GLMMs) to analyze *LikedP* in *likedwthn.sav*. You may wish to refer back to GLMMs from module 9, part 4 before continuing.

42. Run a GLMM on *LikedP* as a nominal dichotomous measure in *likedwthn.sav*. What are the results, and how do they compare to the between-subjects analyses using GZLMs from questions 22, 26, 30, 33, 35, and 37?

    Using a Generalized Linear Mixed Models (GLMM) with Binary logistic regression as its model shows *Method* had a marginal effect on the *LikedP* dichotomous response ($F(1,62) = 3.81$, $p = .056$). These results are actually "weaker" *p*-value-wise than when *Method* was analyzed as a between-subjects factor using a Generalized Linear Model (GZLM) in questions 22, 26, 30, 33, 35, and 37, where it was significant ($\chi^2(1,N=64) = 4.06$, $p = .044$).

Part 6, Nominal logistic regression. You've already seen nominal logistic regression at work; it is the name for the analysis JMP's Fit Model platform uses when the response is nominal. In SPSS, nominal logistic regression is indicated by a Multinomial distribution and Logit link function. It can be considered the generalization of binomial regression with *N* categorical outcomes, not just two.

We will use this opportunity to examine a between-subjects factorial design. The file *comm.jmp* has fictitious data about the communication preferences of teenagers and young adults. Thirty-six people, 18 of each gender, were surveyed as to whether they prefer the new communication media of Facebook or texting to email. (Phrased this way, it requires email to be our last category; examine *Pref*'s Column Info > Column Properties > Value Ordering, and see how email has been placed last. The default would have been alphabetical order.)

43. What are the factors and levels in this data table? (Hint: if we want to be picky, there are no factors here, but other things.) Are these between-subjects or within-subjects variables?

    There are two covariates in this table: *Female?* has two levels (1, 0); *Age* is a continuous variable and thus does not have levels, but has a range of 13-22. These are both between-subjects terms.

44. What is the response variable?

    The response variable is *Pref*, which can have one of three values: Facebook, email, or texting.

45. Form three hypotheses about this data, one for each variable and one for their interaction.

    It seems unlikely that gender should have an effect on communication preferences, but it does seem likely that age has such an effect, perhaps with younger subjects preferring newer forms of communication (Facebook or texting) to the older form (email). It seems unlikely that age would affect each gender differently, *i.e.*, unlikely that there would be a *Female?*Age* interaction.

46. Use Analyze > Distribution to examine all three hypotheses. Report what you see and revise your hypotheses if necessary.

    First, *Pref*'s distribution by *Female?* was examined. Males' preference for email was quite low, whereas females' was much higher. This difference should cause an effect of gender on communication preferences, so this hypothesis is revised. Second, *Age*'s distribution by *Pref* was examined (because *Age* is continuous, it's best to put *Age* in Y, Columns and *Pref* in By). Higher ages seemed to prefer email, while lower ages seemed to prefer texting. A spread of ages preferred Facebook. So the hypothesis about an effect of *Age* is maintained. Third, analyzing *Pref*'s distribution by *Female?* and *Age* showed differences across genders and ages. These differences should result in a significant *Female?*Age* interaction, so this hypothesis is revised.

Use Analyze > Fit Model and run the factorial nominal logistic regression.

47. What is the effect of gender on communication preference? Report and interpret the finding.

    Nominal logistic regression indicates that gender did not have a significant effect on communication preference ($\chi^2(2,N=36) = 4.11$, $p = .13$). Out of 18 males, 7 preferred Facebook, 8 preferred texting, and 3 preferred email. Out of 18 females, 6 preferred Facebook, 5 preferred texting, and 7 preferred email.

48. What is the effect of age on communication preference? Report and interpret the finding.

    Nominal logistic regression indicates that age had a significant effect on communication preference ($\chi^2(2,N=36) = 8.92$, $p < .05$). The average age preferring texting was 16.15 (*SD* = 2.70), preferring Facebook was 17.31 (*SD* = 2.25), and preferring email was 18.70 (*SD* = 1.77).

49. Is there an interaction between gender and age on communication preferences? Report and interpret the finding. (Hint: remember that *p*-values between .05 and .10 are "trends" and deserve some extra interpretive attention.)

    Nominal logistic regression shows a marginal gender by age interaction ($\chi^2(2,N=36) = 5.58$, $p = .06$). A distribution analysis of *Pref* by *Age* and *Female?* shows some discrepancies by gender for the same ages. For example, all three male 15 year-olds preferred texting, while both female 15 year-olds preferred Facebook. Both male 18 year-olds preferred Facebook, while only one of four female 18 year-olds preferred Facebook; of the other three, one preferred texting and two preferred email. These types of preference differences between genders for the same ages give rise to the marginal interaction.

Ignore gender by using Analyze > Fit Y by X to explore the relationship between *Age* and *Pref*. You should produce a graph called "Logistic Fit of Pref By Age." This graph is very useful once you know how to read it. Consult the JMP help documentation's section on "Nominal Logistic Regression," particularly the subsection called "The Cumulative Logistic Probability Plot." Google Books has the same text online if you do not have access to the JMP help documentation. (Direct your browser here: http://books.google.com/books?q="The Cumulative Logistic Probability Plot".)

50. Explain how to read the odds of a preference being Facebook for a given age from the plot. Then do the same for texting and email.

    In general, the odds of a given preference for a given age are given by the *y*-axis height of the relevant curve for the *x*-axis value of age, minus the height of the curve directly below it. For the n^{th} preference without one of the *n*-1 curves, the odds are measured from the top of the graph minus the height of the curve directly below it. Thus, the odds of preferring Facebook at age *x* can be found by first finding *x* on the *x*-axis, then going straight up until the first curve, which corresponds to Facebook, is reached. There, the odds can be read from the left-hand *y*-axis. For the odds of preferring texting at age *x*, the same procedure applies except we travel up to the second curve, which corresponds to texting, and then subtract the odds at that point from the odds of the Facebook curve directly below it. For the odds of preferring email at age *x*, we travel up from *x* to the top of the graph, which corresponds to 1.00, and then subtract the cumulative odds of the texting curve at *x* directly below.

51. Judging from the graph only, what age is most likely to prefer Facebook? texting? email? For each of these preferences, say how many subjects out of all subjects this age made that choice.

    The age most likely to prefer Facebook is about 19 (at this age, 4/7 people prefer Facebook). The age most likely to prefer texting is 13 (at this age, 2/2 people prefer texting). The age most likely to prefer email is 20 (at this age, 2/2 people prefer email).

Part 7, Poisson regression. As you read in module 9, part 4, Poisson regression is a Generalized Linear Model (GZLM) that is useful when analyzing event-count data. Poisson regression is especially suited to rare event data, meaning small counts are common and large counts are exceedingly rare. Consider a study of the number of suicides per school district for a state in a given decade. We should expect, and hope, that most districts in the state have counts of zero; but some will have one, fewer will have two, and still fewer, three. In human-computer interaction, recognition errors for an accurate gesture recognizer might be considered rare events, or other situations when errors are infrequent. Importantly, Poisson regression requires all response data to be non-negative integers. If you have error rates expressed as decimals, you will need to multiply them by 100 and regard the rates as *counts per hundred*.

Open *ptgbtwn.jmp* and examine this data table. Then answer the following questions.

52. What is an educated guess as to what this experiment shows? (After you answer, read the Column Info > Notes for the *Errors* column to discover what this fictitious experiment was about.)

    In this study, each subject was given one input device and one cursor type. Subjects performed 100 random pointing trials as quickly and accurately as they could. This column shows the total number of errors out of those 100 trials.

53. Express this study design in multiplicative notation identifying the factors and levels and whether they are within- or between-subjects.

    This study was a 3×3 between-subjects design. *Device* is a between-subjects factor with 3 levels: mouse, trackball, and touchpad. *Cursor* is a between-subjects factor with 3 levels: point, area, and bubble.

54. Use Analyze > Distribution to examine how *Errors* are distributed for each level of *Device*. Do this again for each level of *Cursor*. Show six test results for examining whether the response is Poisson-distributed within each factor. You may wish to make a table.

    The following table shows the results of a Kolmogorov's *D* test for a Poisson distribution of *Errors* by levels of *Device* and levels of *Cursor*. The Poisson distribution is characterized by one parameter, lambda (λ). Nonsignificant *p*-values indicate that the data are from the Poisson distribution.

***Errors* for…**	**Lambda (λ)**	**Kolmogorov's *D***	***p*-value**
mouse	6.417	.218	.275
touchpad	16.667	.264	.190
trackball	18.083	.179	.648
area	18.000	.318	.052
bubble	7.000	.327	.051
point	16.167	.156	.708

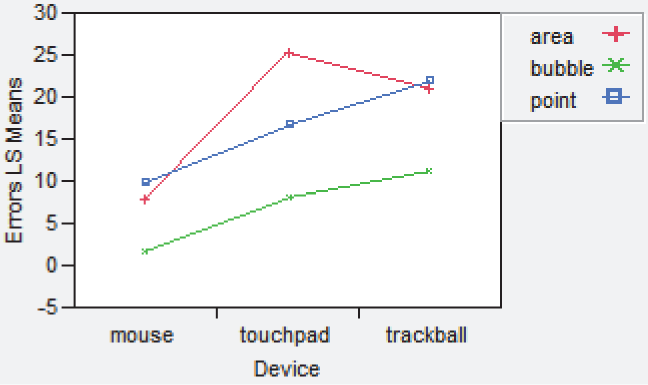
    As the table shows, all levels of each factor result in *Errors* data that can be considered coming from a Poisson distribution. Admittedly, the area and bubble cursors are quite close to significant, meaning they are quite close to being considered non-Poisson by the Kolmogorov's *D* test.

55. Now analyze the distribution of *Errors* as a whole. Is it Poisson-distributed? How can this be in light of what you found in question 54?

    According to a Pearson Chi-Square test, *Errors* overall is not Poisson-distributed ($\chi^2(1,N=36) = 28{,}215.58$, $p < .0001$). Even though each level of both factors is Poisson-distributed, there is no guarantee that when all distributions are combined, the overall distribution remains Poisson. (For example, two combined normal distributions can create a bimodal distribution that is no longer normal at all.) The lambda parameters give an indication of the "center" of the distribution, and as we can see, these lambda values differ. If the lambda parameters were all close, then perhaps the resulting distribution would retain its Poisson shape.

56. Now to go to Analyze > Fit Model and run a Generalized Linear Model with Poisson distribution, Log link, and overdispersion checked. Report and interpret the conclusions from the Effect Tests table.

    Poisson regression shows a significant main effect of *Device* on *Errors* (χ^2(2,*N*=36) = 62.00, *p* < .0001). The mouse committed the fewest errors with an average of 6.42 (*SD* = 4.14). The touchpad and trackball had similar errors with averages of 16.67 (*SD* = 8.02) and 18.08 (*SD* = 6.93), respectively. Poisson regression also shows a significant main effect of *Cursor* on *Errors* (χ^2(2,*N*=36) = 50.67, *p* < .0001). The bubble cursor committed the fewest errors with an average of 7.00 (*SD*=5.72). The point and area cursors had similar errors with averages of 16.17 (*SD* = 6.00) and 18.00 (*SD* = 8.57), respectively. Finally, Poisson regression shows a significant *Device*Cursor* interaction (χ^2(4,*N*=36) = 9.82, *p* < .05). The graph below shows that the area cursor with the touchpad is mostly responsible for this significant interaction.



57. Also run an analysis of variance on these data for comparisons. Report and interpret the conclusions. How do the results compare to those from the Poisson regression in question 56?

    An ANOVA shows a significant main effect of *Device* on *Errors* (F(2,27) = 32.15, *p* < .0001). The mouse committed the fewest errors with an average of 6.42 (*SD* = 4.14). The touchpad and trackball had similar errors with averages of 16.67 (*SD* = 8.02) and 18.08 (*SD* = 6.93), respectively. An ANOVA also shows a significant main effect of *Cursor* on *Errors* (F(2,27) = 27.55, *p* < .0001). The bubble cursor committed the fewest errors with an average of 7.00 (*SD* = 5.72). The point and area cursors had similar errors with averages of 16.17 (*SD* = 6.00) and 18.00 (*SD* = 8.57), respectively. Finally, an ANOVA shows a significant *Device*Cursor* interaction (F(4,27) = 3.01, *p* < .05).

    These conclusions are identical to those using Poisson regression in question 56. The *p*-value for the interaction from Poisson regression was .0435, while the *p*-value from the ANOVA is .0354, suggesting the ANOVA's power is a little stronger but similar in this case.

Find the Device*Cursor report and use the red-arrow menu to select LSMeans Plot. The default places the *Cursor* factor on the *x*-axis and the *Device* factor as the line types. Now remove this plot and once again use the red-arrow menu to select LSMeans Plot, *but before you click to open the menu, hold down SHIFT*. This will "slice" the graph the other way, placing *Device* on the *x*-axis and *Cursor* as the line types.

58. In looking at this graph, does the reason for the significant *Device*Cursor* interaction become clear? What is it? Explain your reasoning.

The graph created is the one shown in the answer to question 56. Nonparallel lines indicate a significant interaction because a disproportionate effect is at work by one level of a factor on another level of a different factor. In the graph, the effect of *Device* as we move from left-to-right along the *x*-axis is proportional on the point and bubble cursors, and proportional for the mouse and trackball on the area cursor. But the touchpad with the area cursor is way out-of-parallel, and is the main culprit for the significant interaction.

Part 8, Contrast tests. Continuing with your analysis of variance output from questions 57-58, run a contrast test by using the red-arrow menu on the Device*Cursor report to select LSMeans Contrast. Use the + and – buttons to select the levels you want to compare. From the graph, it appears possible that the point cursor and area cursor each do not significantly differ between the mouse and trackball.

59. Perform a contrast to see if the point cursor and area cursor differ within the mouse. Report and interpret your finding.

    A contrast test shows that for the mouse, the point cursor and area cursor do not significantly differ on *Errors* (F(1,27) = 0.53, *p* = .47). On average, the point cursor with the mouse committed 9.75 errors (*SD* = 2.75), while the area cursor with the mouse committed 7.75 errors (*SD* = 2.50).

60. Perform a contrast to see if the point cursor and area cursor differ within the trackball. Report and interpret your finding.

    A contrast test shows that for the trackball, the point cursor and area cursor do not significantly differ on *Errors* (F(1,27) = 0.13, *p* = .72). On average, the point cursor with the trackball committed 22.00 errors (*SD* = 3.56), while the area cursor with the trackball committed 21.00 errors (*SD* = 5.60).

61. In light of the findings from questions 59 and 60, it may be that the point and area cursors do not differ at all. If they do, what level of *Device* must this be due to? Perform the relevant contrast test to confirm or refute your hypothesis.

    If there is a significant difference in *Errors* between the point and area cursors, it must be due to the touchpad. A contrast test shows that for the touchpad, the point cursor and area cursor differ significantly on *Errors* (F(1,27) = 9.55, *p* < .01). On average, the point cursor with the touchpad committed 16.75 errors (*SD* = 3.30), while the area cursor with the touchpad committed 25.25 errors (*SD* = 3.10).

62. Investigate whether the point and area cursors are significantly different *overall* by performing a contrast on the Cursor report. Report and interpret your finding.

    A contrast test on *Errors* between the point and area cursors shows no significant difference (F(1,27) = 1.33, *p* = .26). On average, the point cursor had 16.17 errors (*SD* = 6.00), while the area cursor had 18.00 errors (*SD* = 8.57).

63. Now investigate the same thing *but using a contrast on the Device*Cursor report*. (Hint: you will be setting three positive weights against three negative weights.) Do your results agree with question 62? Explain why or why not.

    A contrast test on *Errors* between the point and area cursors shows no significant difference (F(1,27) = 1.33, *p* = .26). On average, the point cursor had 16.17 errors (*SD* = 6.00), while the area cursor had 18.00 errors (*SD* = 8.57). Yes, these results agree exactly with those of question 62. They should agree because a contrast test on levels of *Cursor* averages over all levels of the other factor, in this case, *Device*. When we perform the same contrast on *Device*Cursor*, we are using our weights to do the same thing manually, *i.e.*, to average over all levels of the other factor.

64. Finally, return to your Poisson regression output and use the red-arrow menu on Generalized Linear Model Fit to select Contrast. Determine whether the point and area cursors significantly differ. What contrast did you perform? Report and interpret your results. Are these cursors closer or further from being significantly different using Poisson regression than ANOVA?

Using the red-arrow menu on Generalized Linear Model Fit, we selected Contrast. In Select Contrast Effect, we chose Cursor and clicked Go. In the Contrast Specification, we set area to +1 and point to -1. Then we clicked Done.

This Poisson regression contrast test shows no significant difference between point and area cursors (χ^2(1,*N*=24) = 0.13, *p* = .72).On average, the point cursor had 16.17 errors (*SD* = 6.00), while the area cursor had 18.00 errors (*SD* = 8.57). Using Poisson regression, the contrast between point and area cursors is, judging by the *p*-value, further from significance than using an ANOVA (*p*-values of .72 *vs.* .26).

Part 9, Counts from a mixed factorial design. The event-count data has thus far been between-subjects, *i.e.*, each subject has only contributed one count. What if we have factorial event-count data? Recall from module 9 that the Wilcoxon signed-rank and Friedman tests handle non-independent samples for only one factor. If a factorial design is used, the Aligned Rank Transform (ART) procedure, Generalized Linear Mixed Models (GLMMs), and Generalized Estimating Equations (GEEs) can be used. Each of these analyses can be used for event-count data, which is what we'll do here.

Open *ptgmixed.jmp*, which shows the same pointing error data as *ptgbtwn.jmp*, but this time, with only 12 subjects instead of 36. Each subject used one cursor type with three input devices, resulting in a mixed factorial design. The between-subjects factor is *Cursor*, and the within-subjects factor is *Device*. Note the column order has been changed to reflect this relationship clearly in the table.

65. Before conducting the nonparametric procedures, first analyze this data set using a (linear) mixed-effects model analysis of variance. Report and interpret your findings.

A mixed-effects model analysis of variance exhibits a significant main effect of *Cursor* on *Errors* (F(2,9) = 29.23, *p* = .0001). The bubble cursor committed the fewest errors with an average of 7.00 (*SD* = 5.72). The point and area cursors had similar errors with averages of 16.17 (*SD* = 6.00) and 18.00 (*SD* = 8.57), respectively. A mixed-effects model analysis of variance also exhibits a significant main effect of *Device* on *Errors* (F(2,18) = 31.25, *p* < .0001). The mouse committed the fewest errors with an average of 6.42 (*SD* = 4.14). The touchpad and trackball had similar errors with averages of 16.67 (*SD* = 8.02) and 18.08 (*SD* = 6.93), respectively. Finally, there exists a significant *Cursor*Device* interaction (F(4,18) = 2.93, *p* < .05).

66. How do your results for the mixed factorial design compare to those from the between-subjects design (question 57)? Make a table comparing your degrees-of-freedom, F-ratios, and *p*-values.

The statistical reports for these two analyses are shown in the table below:

	between-subjects design (q. 57)				***mixed factorial design (q. 65)***			
	df_{num}	df_{den}	*F*	*p*	df_{num}	df_{den}	*F*	*p*
Cursor	2	27	27.55	<.0001	2	9	29.23	.0001
Device	2	27	32.15	<.0001	2	18	31.25	<.0001
*Cursor*Device*	4	27	3.01	.0354	4	18	2.93	.0499

Now copy your data from JMP to Microsoft Excel, including column names, and then save it as an MS-DOS *.csv file. Run the ARTOOL program available at http://depts.washington.edu/aimgroup/proj/art/ to align and rank the data. Then open the aligned *.csv file in Microsoft Excel, and paste the data into a new JMP table. Paste in the column names as well. Save the new JMP table as *ptgmixed.art.jmp* to reflect the *.csv filename created for you by ARTOOL.

67. Use Tables > Summary to sum the aligned columns. Aligned columns in a long-format table should always sum to zero. Do they?

    The sums of columns named aligned(Errors) for Cursor, aligned(Errors) for Device, and aligned(Errors) for Cursor*Device are, ignoring precision, equal to zero.

68. Run a (linear) mixed-effects model analysis of variance on the "aligned" columns. Are all effects for which the response was *not* aligned properly "stripped out?"

    The analysis of variance for *aligned(Errors) for Cursor* shows that the *Cursor* effect is intact but the *Device* and *Cursor*Device* effects are properly stripped out (F = 0.00, *p* = 1.00). The analysis of variance for *aligned(Errors) for Device* shows that the *Device* effect is intact but the *Cursor* and *Cursor*Device* effects are properly stripped out (F = 0.00, *p* = 1.00). The analysis of variance for *aligned(Errors) for Cursor*Device* shows that the *Cursor*Device* interaction is intact but the *Cursor* and *Device* effects are properly stripped out (F = 0.00, *p* = 1.00).

69. Now run the model on the "ART" columns. Report and interpret your findings for *Device*, *Cursor*, and their interaction.

    Using the Aligned Rank Transform (ART) procedure, we see that *Cursor* exhibits a significant effect on *Errors* (F(2,9) = 19.94, *p* < .001). The bubble cursor committed the fewest errors with an average of 7.00 (*SD* = 5.72). The point and area cursors had similar errors with averages of 16.17 (*SD* = 6.00) and 18.00 (*SD* = 8.57), respectively. *Device* also exhibits a significant effect on *Errors* (F(2,18) = 27.82, *p* < .0001). The mouse committed the fewest errors with an average of 6.42 (*SD* = 4.14). The touchpad and trackball had similar errors with averages of 16.67 (*SD* = 8.02) and 18.08 (*SD* = 6.93), respectively. Finally, the ART procedure also gives a significant *Cursor*Device* interaction (F(4,18) = 3.09, *p* < .05).

70. How do each of these findings compare to those from the parametric analysis (question 65)? Reuse the mixed factorial design results from your table in question 66, and fill in the other half with the results from question 69 to facilitate comparisons.

    The statistical reports for these two analyses are similar, as shown in the table below:

	mixed factorial design (q. 65)				***Aligned Rank Transform (q. 69)***			
	df_{num}	df_{den}	*F*	*p*	df_{num}	df_{den}	*F*	*p*
Cursor	2	9	29.23	.0001	2	9	19.94	.0005
Device	2	18	31.25	<.0001	2	18	27.82	<.0001
*Cursor*Device*	4	18	2.93	.0499	4	18	3.09	.0424

In general, it seems that the Aligned Rank Transform procedure produces slightly weaker main effect F-ratios than the parametric mixed-effects model analysis of variance, but a slightly stronger interaction F-ratio. In general, the results are similar.

Now we will employ Generalized Linear Mixed Models (GLMMs) and Generalized Estimating Equations (GEEs) to analyze *Errors* for the *ptgmixed.jmp* data. As these two analyses are unavailable in JMP, recreate the *ptgmixed.jmp* data in SPSS 19 and save the new table as *ptgmixed.sav*. Then answer the following questions.

71. Run a GLMM on *Errors* as a scalar response, *Cursor* as a between-subjects nominal factor, *Device* as a within-subjects nominal factor, and *Subject* as a random effect. Use Poisson regression, accessed via choosing Loglinear as the model type. On the Random Effects pane, be sure to add *Subject* as a random effect by adding a block, choosing *Subject* in the Subject combination dropdown and clicking Include intercept. What are the results of the statistical analyses?

    Using a Generalized Linear Mixed Model (GLMM), we see that *Cursor* had a significant effect on *Errors* (F(2,27) = 19.64, *p* < .001). *Device* also had a significant effect on *Errors* (F(2,27) = 23.87, *p* < .001). Finally, there was a marginal *Cursor*Device* interaction (F(4,27) = 2.56, *p* = .061).

72. Run a GEE on *Errors* as a scalar response, *Cursor* as a between-subjects nominal factor, *Device* as a within-subjects nominal factor, and *Subject* as a Subject variable. Use a Poisson loglinear model. What are the results of the statistical analyses?

    Using a Generalized Estimating Equation (GEE), we see that *Cursor* had a significant effect on *Errors* (χ^2(2,*N*=36) = 89.13, *p* < .001). *Device* also had a significant effect on *Errors* (χ^2(2,*N*=36) = 42.68, *p* < .001). Finally, there was a significant *Cursor*Device* interaction (χ^2(4,*N*=36) = 21.75, *p* < .001).

73. Create a table to compare the results from your ART, GLMM, and GEE analyses of the *ptgmixed.** data. Which analysis seems most powerful for this data set?

    The table below compares our three nonparametric analyses of the *ptgmixed.** data. The ART and GLMM give similar results, while the GEE gives “stronger” results, at least judging by the highly significant interaction, which was near .05 in the other two analyses.

	ART (q. 69)				***GLMM (q. 71)***				***GEE (q. 72)***		
	df_{num}	df_{den}	*F*	*p*	df_{num}	df_{den}	*F*	*p*	*df*	χ^2	*p*
Cursor	2	9	19.94	.0005	2	27	19.64	.000	2	89.13	.000
Device	2	18	27.82	<.0001	2	27	23.87	.000	2	42.68	.000
*Cursor*Device*	4	18	3.09	.0424	4	27	2.56	.061	4	21.75	.000

Bibliography

Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution*. Cambridge, England: Cambridge University Press.

Akritas, M. G. and Brunner, E. (1997). A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference 61* (2), pp. 249-277.

Akritas, M. G. and Osgood, D. W. (2002). Guest editors' introduction to the special issue on nonparametric models. *Sociological Methods and Research 30* (3), pp. 303-308.

Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics 23* (2), pp. 193-212.

Anderson, T.W. and Darling, D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association 49* (268), pp. 765-769.

Bartlett, M. S. (1947). The use of transformations. *Biometrics 3* (1), pp. 39-52.

Bennington, C. C. and Thayne, W. V. (1994). Use and misuse of mixed model analysis of variance in ecological studies. *Ecology 75* (3), pp. 717-722.

Berry, D. A. (1987). Logarithmic transformations in ANOVA. *Biometrics 43* (2), pp. 439-456.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B 26* (2), pp. 211-252.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88* (421), pp. 9-25.

Card, S. K., Moran, T. P. and Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, New Jersey: Lawrence Erlbaum.

Conover, W.J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association 67* (339), pp. 591-596.

Conover, W. J. and Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics *The American Statistician 35* (3), pp. 124-129.

D’Agostino, R.B. (1986). Tests for the normal distribution. In *Goodness-of-Fit Techniques*, R. B. D'Agostino and M. A. Stephens (eds.). New York: Marcel Dekker, pp. 367-420.

Fawcett, R. F. and Salter, K. C. (1984). A Monte Carlo study of the F test and three tests based on ranks of treatment effects in randomized block designs. *Communications in Statistics: Simulation and Computation 13* (2), pp. 213-225.

Findlater, L., Moffatt, K., McGrenere, J. and Dawson, J. (2009). Ephemeral adaptation: The use of gradual onset to improve menu selection performance. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '09)*. Boston, Massachusetts (April 4-9, 2009). New York: ACM Press, pp. 1655-1664.

Frederick, B. N. (1999). Fixed-, random-, and mixed-effects ANOVA models: A user-friendly guide for increasing the generalizability of ANOVA results. In *Advances in Social Science Methodology*, B. Thompson (ed). Stamford, Connecticut: JAI Press, pp. 111-122.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32* (200), pp. 675-701.

Geisser, S. and Greenhouse, S. W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics 29* (3), pp. 885-891.

Greenhouse, S. W. and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika 24* (2), pp. 95-112.

Hanley, J. A., Negassa, A., Edwardes, M. D. deB. and Forrester, J.E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology 157* (4), pp. 364-375.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Mathematical Statistics 12*, pp. 61-75.

Higgins, J. J., Blair, R. C. and Tashtoush, S. (1990). The aligned rank transform procedure. *Proceedings of the Conference on Applied Statistics in Agriculture*. Manhattan, Kansas: Kansas State University, pp. 185-195.

Higgins, J. J. and Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World 1* (2), pp. 201-211.

Higgins, J. J. (2004). *Introduction to Modern Nonparametric Statistics*. Pacific Grove, California: Duxbury Press.

Hodges, J.L. and Lehmann, E.L. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics 33* (2), pp. 482-497.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6* (2), pp. 65-70.

Hoyle, M. H. (1973). Transformations: An introduction and a bibliography. *International Statistical Review 41* (2), pp. 203-223.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effect in mixed linear models. *Journal of the American Statistical Association 79* (388), pp. 853-862.

Kaiser, J. F. and Reed, W. A. (1977). Data smoothing using low-pass digital filters. *Review of Scientific Instruments 48* (11), pp. 1447-1457.

Kaptein, M., Nass, C. and Markopoulos, P. (2010). Powerful and consistent analysis of Likert-type rating scales. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '10).* Atlanta, Georgia (April 10-15, 2010). New York: ACM Press, pp. 2391-2394.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics 53* (3), pp. 983-997.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics 12*, pp. 309-310.

Kristensson, P.-O. and Zhai, S. (2004). SHARK^2: A large vocabulary shorthand writing system for pen-based computers. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '04)*. Santa Fe, New Mexico (October 24-27, 2004). New York: ACM Press, pp. 43-52.

Krueger, C. and Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing 6* (2), pp. 151-157.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association 47* (260), pp. 583-621.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics 38* (4), pp. 963-974.

Lehmann, E. L. (2006). *Nonparametrics: Statistical Methods Based on Ranks*. New York: Springer.

Levene, H. (1960), Robust tests for the equality of variances. In I. Olkin (ed.), *Contributions to Probability and Statistics*. Stanford, California: Stanford University Press.

Limpert, E., Stahel, W. A. and Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience 51* (5), pp. 341-352.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics 46* (3), pp. 673-687.

Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996). Analysis of repeated measures data. In *SAS System for Mixed Models*. Cary, North Carolina: SAS Institute, Inc., pp. 87-134.

Littell, R. C., Henry, P. R. and Ammerman, C. B. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science 76* (4), pp. 1216-1231.

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology 25* (12), pp. 1032-1037.

MacKenzie, I. S. and Zhang, S. X. (1999). The design and evaluation of a high-performance soft keyboard. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '99)*. Pittsburgh, Pennsylvania (May 15-20, 1999). New York: ACM Press, pp. 25-31.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics 18* (1), pp. 50-60.

Mansouri, H. (1999). Aligned rank transform tests in linear models. *Journal of Statistical Planning and Inference 79* (1), pp. 141-155.

Mansouri, H. (1999). Multifactor analysis of variance based on the aligned rank transform technique. *Computational Statistics and Data Analysis 29* (2), pp. 177-189.

Mansouri, H., Paige, R. L. and Surles, J. G. (2004). Aligned rank transform techniques for analysis of variance and multiple comparisons. *Communications in Statistics: Theory and Methods 33* (9), pp. 2217-2232.

Massey, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association 46* (253), pp. 68-78.

Milliken, G. A. and Johnson, D. E. (1984). Analysis of Messy Data vol. I: Designed Experiments. New York: Van Nostrand Reinhold Company.

Petocz, P. (2005). An upper bound on standard deviation as a function of range. *Teaching Statistics 27* (2), pp. 42-44.

Richter, S. J. (1999). Nearly exact tests in factorial experiments using the aligned rank transform. *Journal of Applied Statistics 26* (2), pp. 203-217.

Salter, K. C. and Fawcett, R. F. (1985). A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics: Simulation and Computation 14* (4), pp. 807-828.

Salter, K. C. and Fawcett, R. F. (1993). The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation 22* (1), pp. 137-153.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin 2* (6), pp. 110-114.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research 60* (1), pp. 91-126.

Schuster, C. and von Eye, A. (2001). The relationship of ANOVA models with random effects and repeated measurement designs. *Journal of Adolescent Research 16* (2), pp. 205-220.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology 46* (1), pp. 561-584.

Shah, A. K. (1985). A simpler approximation for areas under the standard normal curve. *The American Statistician 39* (1), p. 80.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika 52* (3 & 4), pp. 591-611.

Stephens, M.A. (1972). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association 69* (347), pp. 730-737.

Tukey, J. (1953). A problem of multiple comparisons. Dittoed manuscript of 396 pages, Princeton University.

Vermunt, J. K. (1997). *Log-linear Models for Event Histories*. Thousand Oaks, California: Sage Publications.

Ware, J.H. (1985). Linear models for the analysis of serial measurements in longitudinal studies. *American Statistician 39* (2), pp. 95-101.

Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika 38* (3 & 4), pp. 330-336.

West, B.T. (2009). Analyzing longitudinal data with the linear mixed models procedure in SPSS. *Evaluation and the Health Professions 32* (3), pp. 207-228.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1* (6), pp. 80-83.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review 49* (4), pp. 512-525.

Wobbrock, J. O., Chau, D. H. and Myers, B. A. (2007). An alternative to push, press, and tap-tap-tap: Gesturing on an isometric joystick for mobile phone text entry. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '07)*. San Jose, California (April 28-May 3, 2007). New York: ACM Press, pp. 667-676.

Wobbrock, J.O., Findlater, L., Gergle, D. and Higgins, J.J. (2011). The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI ’11)*. Vancouver, British Columbia (May 7-12, 2011). New York: ACM Press, pp. 143-146.

Wobbrock, J. O., Myers, B. A. and Kembel, J. A. (2003). EdgeWrite: A stylus-based text entry method designed for high accuracy and stability of motion. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '03)*. Vancouver, British Columbia (November 2-5, 2003). New York: ACM Press, pp. 61-70.

Wobbrock, J. O. and Myers, B. A. (2006). Trackball text entry for people with motor impairments. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '06)*. Montréal, Québec (April 22-27, 2006). New York: ACM Press, pp. 479-488.

Wobbrock, J. O., Rubinstein, J., Sawyer, M. W. and Duchowski, A. T. (2008). Longitudinal evaluation of discrete consecutive gaze gestures for text entry. *Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA '08)*. Savannah, Georgia (March 26-28, 2008). New York: ACM Press, pp. 11-18.

Wobbrock, J. O., Wilson, A. D. and Li, Y. (2007). Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '07)*. Newport, Rhode Island (October 7-10, 2007). New York: ACM Press, pp. 159-168.

Zhai, S. and Kristensson, P. (2003). Shorthand writing on stylus keyboard. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '03)*. Ft. Lauderdale, Florida (April 5-10, 2003). New York: ACM Press, pp. 97-104.

Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics 44* (4), pp. 1049-1060.