

One-Way Repeated Measures ANOVA

*Dr. Tom Pierce
Department of Psychology
Radford University*

Differences between Between-subjects and within-subjects independent variables

Review of confounds and the control of confounds in between-subjects designs

I know you've heard all this before. It won't hurt you a bit to hear it again. Much.

Let's say that the design of your study goes like this: there are three levels of one independent variable and there are ten subjects assigned to each of these three levels. What kind of design is this? The comparison being made is between three separate groups of subjects. That makes it a **between-subjects design**. The independent variable in this case is an example of a **between-subjects factor**.

So what kinds of confounds should you be especially worried about when the independent variable is a between-subjects factor? In this design each level of the independent variable is comprised of different subjects. A confound, in general, is any reason, other than your independent variable, for why there are differences among the means for your independent variable. The independent variable is a difference between the conditions that the investigator put there. When (a) the statistical analysis shows that there are significant differences among the means and (b) the only possible explanation for differences between the means is the manipulation of the independent variable, the investigator is justified in inferring that differences in the means for the dependent variable must have been caused by the manipulation of the independent variable. If there is any other possible explanation for why the means ended up being different from each other you can't be sure that it was your I.V. that had the effect. If you can't tell whether the effect was caused by your I.V. or not, your data are officially worthless. You have a confound in your design. The confound is a competing explanation for why you got a significant effect.

Confounds in between-subjects designs usually take the form of individual difference variables. The independent variable being manipulated might be the dosage of caffeine subjects receive. But if you observe that subjects getting the high dosage are all in their 60s and 70s in terms of age, but the subjects getting the low dosages are all in their 20s and 30s, there is no way to tell if the significant effect that you observe is because of the independent variable you wanted to manipulate (dosage of caffeine) or because of the group difference in age. Age is a way in which subjects differ from each other. Age as an individual difference variable constitutes a confound in this particular example. Other individual difference variables that investigators often worry about are gender, race, socio-economic status, years of education, and employment history. There are probably hundreds or thousands of other ways in which subjects could differ from each other.

Every one of them is a potential reason for why the means in a given study are not all the same.

Investigators try to remove the threat of confounds in their designs through (a) random assignment to groups and (b) through matching. Random assignment to groups works to control confounds due to individual differences by making sure that every subject in the study has an equal chance of being assigned to every treatment condition. The only way that the subjects in one group could be significantly older than the subjects in another group would be by chance. The strategy of random assignment to groups is appealing because it automatically controls for every possible individual difference variable. The disadvantage to random assignment to groups is that it is always possible that there might be differences between the group on a particular variable just by chance. And if you never measured this variable (and there's no way to measure every possible confounding variable) you'd have to live with the possibility that a confound is present in the design just by chance.

Random assignment to groups leaves it up to chance as to whether or not there are confounds due to individual differences in the design. Matching allows the investigator the luxury of fixing it so that a particular variable will not be a confound in the design. Unfortunately, without extremely large sample sizes it is difficult to match on more than three or four possible confounding variables.

To sum this bit up, confounds in between-subjects designs are due to the unwanted influence of individual difference variables. Strategies for controlling for confounds in this case try to make sure that the only thing that makes the subjects in one group different from the subjects in another group is the one variable that the investigator has control over: the independent variable.

Controlling for confounds in within-subjects designs

In the search for ways to control for the confounding effects of individual difference variables, an alternative strategy is to simply use the same subjects in each treatment condition of the independent variable. Think about it. If you used the same subjects in Condition 1 that you did in Condition 2, Condition 3, and so on, could the subjects who contributed data to the different treatment conditions possibly be different from each other on age? No! because you're using the same subjects in each group. The mean age for the subjects giving you data for condition 1 is going to be exactly the same as the mean age of the subjects giving you data for Condition 2, Condition 3, and so on. You've automatically and perfectly controlled for the effects of age. You've automatically and perfectly controlled for the effects of gender because the ratio of male to female is going to be identical at every level of the I.V. Testing the same subjects under every level of the I.V. guarantee that confounds due to any individual difference variable will not be present in the design. That's the primary advantage of having every subjects receive every level of the independent variable. Because comparisons between the levels of the I.V are conducted within one group of subjects the I.V. is referred to as a

within-subjects factor. When the design of the study includes one independent variable that is within-subjects, the design is known as a **within-subjects design.**

So manipulating the I.V. within one group of subjects eliminates individual difference variables as potential confounds. Does this mean that there are no confounds associated with a within-subjects I.V.? No!

The subjects in the various treatment conditions are the same. Yet these subjects obviously could not have participated in these different conditions at the same time! Within-subjects factors are vulnerable to confounds due to changes in the testing conditions over time or in changes in the state of the subject over time. Changes in the testing conditions that are not related to the independent variable being manipulated are referred as the effects of time or history. Changes in the state of the subject that are not related to the independent variable being manipulated are referred to as practice effects. Let's deal with confounds due to historical factors first, then we'll look at practice effects.

Confounds due to the effects of historical events. Let's start with a concrete example. You're interested in the effects of age on cognitive function. You administer a battery of cognitive tests (perhaps from the WAIS and the Halstead-Reitan Neuropsychological Test Battery). One of these tests is the Digit Symbol subtest from the WAIS. The measure obtained from this subtest is the number of items completed in 90 seconds. You administer this battery of tests to a group of five subjects every five years (I know this is a pathetically low sample size, but it's just an example. Get off my back.). You now have data from subjects when they were 65 years old, 70 years old, and 75 years old. The data for this study are presented below.

Subject	65 Years Old -----	70 Years Old -----	75 Years Old -----
1	55	51	45
2	63	60	59
3	49	51	47
4	51	44	44
5	44	39	34

When we calculate the means for the three levels of Age we find that the mean level of performance on the Digit Symbol test seems go down as the subjects get older. But what if you knew that one year before the third time of testing, when the subjects were 74 years of age, there was a massive food shortage and pretty much everyone had to go without the recommended daily allowance of Vitamin Z. Even if we get to say that, statistically, the mean of the scores at time 3 are significantly lower than the scores at time 2, do we get to say that it was the increase of five years of age that caused the scores to go down? No, because it's also possible to argue that it was the lack of Vitamin Z that caused the decline in the scores, not the increase in age. The historical event, the food shortage, is a confound in this design. The food shortage produced a competing

explanation for why the scores for level three of the I.V. were different from the scores at the second level of the I.V.

Confounds due to practice effects. Let's think about a different study. You do an experiment in which you manipulate the amount of caffeine subjects get and observe the effects on reaction time. You ask subjects to come in at 8:00 in the morning to begin the experiment. You obtain their informed consent and then give them a pill that you've told them contains a high or a low dose of caffeine, or no caffeine at all. The subjects don't know what dosage of caffeine they're getting and the experimenter doing the testing doesn't know how much caffeine they're getting (that's what makes this a double-blind study). Twenty minutes after the subjects get the pill the experimenter has them do what's called a choice reaction time. On each trial the subject is asked to press one button as fast as they can if they see a digit appear on a computer screen and another button as fast as they can if they see a letter. There are 400 trials in the experiment. And it takes approximately a half an hour to complete the experiment.

After the subject has completed the reaction time task they are asked to wait a hour to give the caffeine a chance to wash out of their system. Then they go through the same procedure with a second dosage of caffeine (get the pill, wait 20 minutes, do the 400 trial RT task). After completing the task a second time they go through the identical procedure gain with a third dosage of caffeine. By the time the subject has ended their participation in the study they have taken all three dosages of caffeine and performed the same RT task three different times. Is there any potential for a confound here? What if you did the same excruciatingly boring RT task three times in the same day. Do you think you're performance the third time you did the task would be the same as doing the task for the first time? No way! What's potentially different about doing the task that third time? Well, for one thing you'd probably be pretty sick and tired of doing the task. It's a pretty good bet that you'd be much more tired and a lot more bored the last time you did the task that day than the first time you did it. That's a potential explanation for why performance at the third time of testing could be different than performance at the first time of testing! **Fatigue** and **boredom** have a potentially confounding effect in the design. If you find that performance in the high dose condition (third time of testing) is significantly different from performance in the no caffeine condition (first time of testing), there's no way to tell if it's the caffeine that's causing the effect or the onset of boredom or fatigue. In addition, it's possible that that performance at the third time of testing is better because of the extensive amount of **practice** the subjects have already received in doing the task.

Practice, fatigue, and boredom are all included under the heading of **practice effects**. Practice effects represent confounds in a within-subjects design that constitute changes in the state of the research subject across repeated administrations of the same procedures for getting a score on the dependent variable.

So how do experimenters control for the confounding effects of practice? The answer is that there is no way to eliminate practice effects. You simply cannot prevent people from gaining experience with a task, or from becoming bored or tired. However, it is possible

to make sure that, whatever effects there are of practice, fatigue, or boredom, these effects are spread out evenly over the various levels of the independent variable. This is accomplished by varying the order in which the levels of the independent variable are administered to subjects. For example, the table below presents a scheme for making sure that that, in a set of three subjects, every level of the independent variable is administered at every time of testing only once. Varying the order of administration of levels of the independent variable in order to control for practice effects is known as counterbalancing.

Subject	Time 1	Time 2	Time 3
-----	-----	-----	-----
1	No Caff.	Low Dose	High Dose
2	Low Dose	High Dose	No Caff.
3	High Dose	No Caff.	Low Dose

Counterbalancing works for control for practice effects by making sure that no one condition is any more influenced by practice, fatigue, or boredom than any other condition. After all, all three dosages of caffeine occur once at Time 3, when the effects of practice should be most evident. In this set of four subjects, the three levels of the independent variable occur one time each at each of the three times of testing. Counterbalancing doesn't eliminate practice effects. Counterbalancing spreads the effects of practice out evenly across every level of the within-subjects independent variable.

The error term for a one-way between-subjects ANOVA

In a one-way ANOVA in which the independent variable is a between-subjects factor, where does the error term come from? When you get a Mean Square Not Accounted-For in a between-subjects ANOVA, what makes this number "error"? Numerically, the number is based on how far each raw score is from the mean of their group. Because everyone in a particular group was treated exactly alike, with respect to the independent variable, differences between the raw scores and the means of their groups can't possibly be due to the effect of the independent variable. The sum of squared deviations between the raw scores and the group means is referred to as "error" because this is an amount of variability that the independent variable cannot explain. And in data analysis anything that is due to chance or that can't be explained is referred to as error. Again, it's not that anyone made a mistake. Error is just something that we don't have an explanation for. Error is variability that the independent variable should account for, but it can't account for.

The error for a between-subjects factor comes from the fact that there is variability in the scores within each group, but the independent variable can't explain why. What

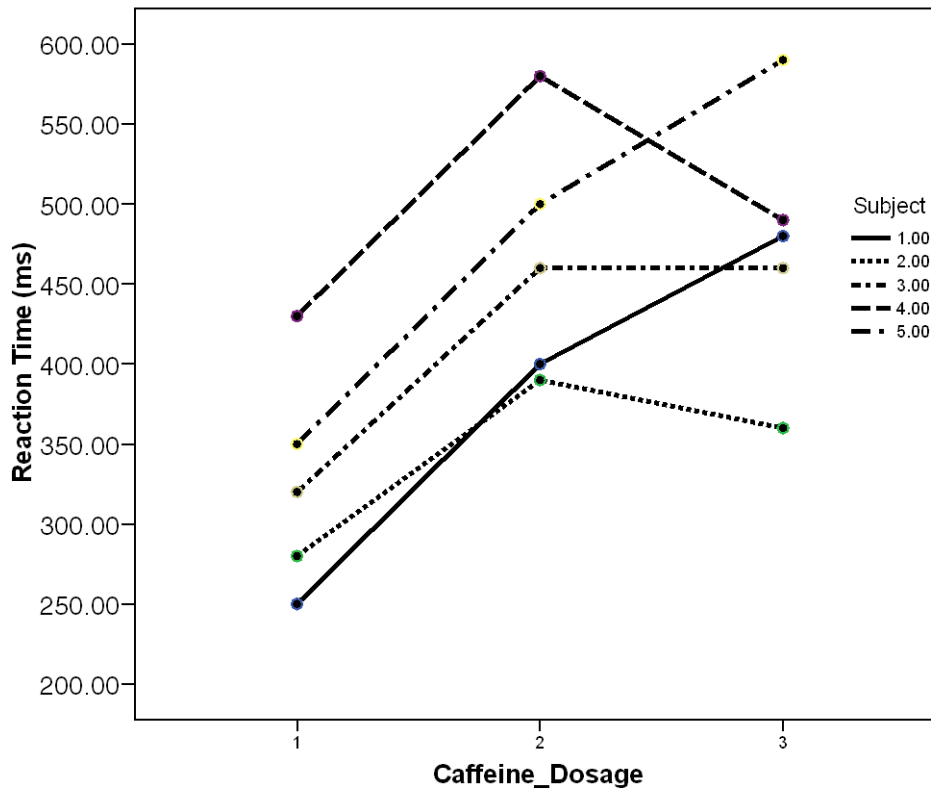
variability should a within-subjects I.V. be able to explain, but it can't explain? That amount of variability is the error term that we ought to be using.

Below is the data for the five subjects we mentioned earlier in the study on the effects of caffeine on reaction time.

Caffeine Dosage (A)				
Subject	a1 No Caff.	a2 Low Dose	a3 High Dose	
1	250	400	480	s1 = 410.0
2	280	390	360	s2 = 376.67
3	320	460	460	s3 = 446.67
4	430	580	490	s4 = 533.33
5	350	500	590	s5 = 513.33

	A1 = 326	A2 = 466	A3 = 476	

Below is a graph of the data presented above.



In the graph, what happens to subject 1 as they go from the No Caffeine Condition to the low dose condition to the high dose condition? They start out with a score of 250 and they go to 400 msec and then to 480 msec. What is the effect of changing the dosage on RT for this subject? The effect of the IV is the change in the scores that you can attribute to changing the conditions. For this subject, if you change the conditions from 0 mg to 4 mg you see an increase of 150 msec in RT. This 150 msec increase is the effect of giving the subject an additional 4 mg of caffeine. What's the effect of going from 4 mg of caffeine to 8 mg? Again, the effect is a further increase of 80 msec, going from 400 msec to 480 msec. That's the effect of Caffeine Dosage on subject one: a 150 msec increase going from no caffeine to a low dose and a 80 msec increase going from a low dose to a high dose.

What's the effect of changing the dosage on subject 2? For subject 2, going from no caffeine to a low dose resulted in an increase of 110 msec. Going from a low dose to a high dose was associated with a **decrease** of 30 msec. Is the effect of caffeine the same for subject 1 as it is for subject 2? No! Giving subject 1 more caffeine was associated with substantially greater increases in RT than for subject 2. Why?! Every subject was treated exactly alike with respect to the independent variable. Everyone was administered exactly the same treatment in exactly the same way. **So why didn't the two subjects respond in exactly the same way to the different dosages of caffeine? We don't know. We have no explanation for this.** If the I.V. was the only explanation for why scores on the dependent variable change over time, every subject in the study ought to display the identical pattern of change in their scores as you go from one dosage of caffeine to the next. But clearly they don't. The effect of the I.V. is not the same for the various subjects in the study. This is the error for a within-subjects factor: the degree to which the effect of the I.V. is not the same for every subject.

So how do we put a number on the error attributable to a within-subjects factor? It turns out that we've already talked about the basic idea of computing this source of variability. Think about the definition of the error for the within-subjects factor: the degree to which the effect of the I.V. is not the same for every subject. What does this remind you of? Let me say it again: **THE DEGREE TO WHICH THE EFFECTS OF THE WITHIN-SUBJECTS FACTOR ARE NOT THE SAME FOR EVERY SUBJECT.** This sounds like the definition for an interaction. That's because it is. But wait a minute. How can I have an interaction between two independent variables when there's only one independent variable? That's a fair question. The answer is that the interaction we're dealing with here is between the independent variable that we're interested in (Dosage of Caffeine) with another "variable" that we can treat as a second independent variable.

The mysterious second I.V. is nothing other than "Subject". Think about it for a second. Someone might theoretically be interested in seeing whether there are differences among the subjects when you average over the three dosages of caffeine. You can do this because we have data for every combination of a level of Caffeine Dosage and a level of Subject (because every subject went through every level of Caffeine Dosage). **The**

appropriate error term used to test the effect of Caffeine Dosage is the degree to which Caffeine Dosage interacts with Subject.

Look at the plot of the data for each subject as they provided data at each level of Caffeine Dosage. Are the lines parallel with each other? No, not really. Clearly, when you increase the dosage you don't get exactly the same effect for each of the five subjects. The job of the independent variable Caffeine Dosage is to explain why people's scores change as you go from one time of testing to another. If everyone displayed the same pattern of change, it would be fair to say that there would be no error, as far as the effect of Caffeine Dosage. If this were the case, the lines for all of the subject's would have to be perfectly parallel to each other. The degree to which these lines are not parallel to each other is variability that Caffeine Dosage cannot explain. Why is it that when every subject was treated exactly alike, the subjects didn't all display exactly the same pattern of change over time? Knowing how much caffeine people got at each time of testing can't help us to answer this question. That's what makes AXS the error term for a one-way within-subjects ANOVA.

The calculation of the sum of squares accounted-for is exactly the same as in the between-subjects ANOVA. It's based on the deviation between the treatment means (the means for each Caffeine Dosage) and the grand mean (the mean of all the scores in the data set).

The calculation of the error term AXS is the same idea as calculating the sum of squares for AXB that we just got done talking about. It helps if you think of this study as a design that has three levels of one I.V. (Caffeine Dosage) and five levels of the other I.V. (Subject). The only odd thing about the design is that there is only one subject at each combination of a level of Caffeine Dosage and a level of Subject.

Because the number of subjects in each cell (combination of a level of A and a level of S) is one, there is no variability of scores within cells. This means that there's no equivalent of S/AB. There's no variability of subjects within cells (you can't have variability when there's only one subject!). This means that there are only three sources of variability that contribute to the sum of squares total: (1) the main effect of Caffeine Dosage, (2) the main effect of Subject, and (3) the interaction between Caffeine Dosage and Subject. The interaction between Caffeine Dosage and Subject is the error term used to test the two "main effects". For the purposes of the researcher, the only effect that it really makes sense to test is the effect of Caffeine Dosage. You certainly could test the effect of Subject if you wanted to, but what would it tell you. You already know that people are different from each other. Big deal.

Below is the ANOVA table for the data discussed above. Notice that the degrees of freedom are calculated in exactly the same way as in the two-way between-subjects ANOVA.

Source	SS	df	MS	F(observed)	F(critical)
A	70333.33	2	35166.67	16.29	4.46
S	53293.33	4			
A X S	17266.67	8	2158.33		

The observed value for F for the effect of Caffeine Dosage is greater than the critical value for F, so we have a significant overall effect of Caffeine Dosage. What does this tell us? It tells us that there are differences among the means for the three levels of Caffeine Dosage. It doesn't tell us where these differences are. So what are we supposed to do? You guessed it. Comparisons among treatment means.

Comparisons among treatment means for a one-way within-subjects ANOVA

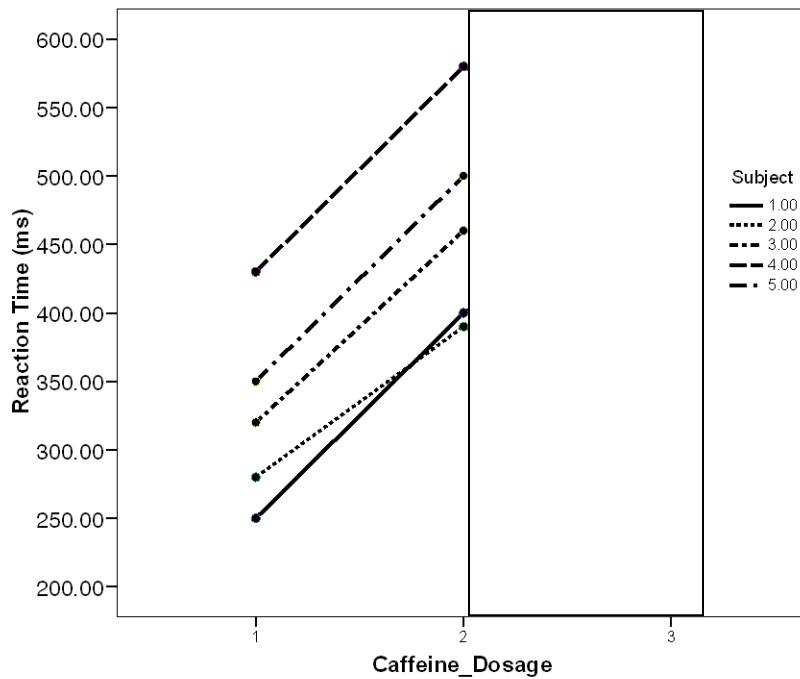
Back in the good old days, when life was simple, and the truth self-evident... when we had an independent variable that was between-subjects we always used the same error term. This error term might have been labeled S/A or S/AB. It didn't matter. It meant the same thing. It meant "the variability of subjects within groups". The "error" we had to measure was the degree to which subjects treated exactly alike with respect to the I.V. ended up having different scores.

The reason we got to keep using the same error term, no matter which scores were being compared to which other scores, was we **assumed that the variability of the scores in the different groups is always the same**. The mean square for S/AB is basically the average variance (S^2) for every group of subjects in the study. For a design with two levels of A and three levels of B (2 X 3 design) the MS S/AB is the mean variance for the six groups in the study. Because the values for S^2 are assumed to be the same for all six groups, it doesn't matter whether you're only looking at the subjects involved in the simple effect of B at a1. The average variance for the three groups involved in the simple effect will be the same number as the average variance of all six groups – because the group variances are all assumed to be the same! The assumption of homogeneity of variance saved us from having to calculate a new error term for every effect.

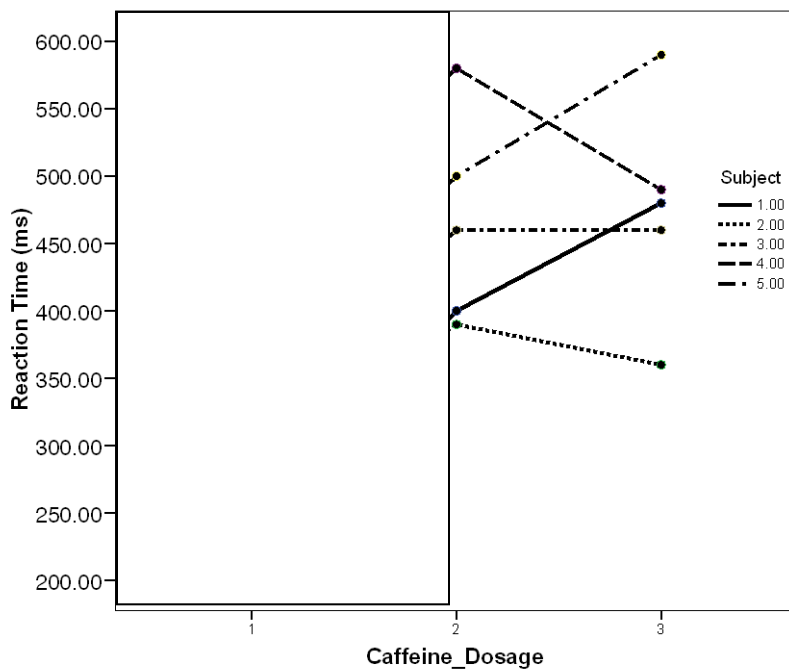
Unfortunately, the error we're dealing with in a within-subjects is not the variability of the scores within levels of the independent variable. It's the interaction between the independent variable and subjects. There is absolutely no guarantee that the sum of squares for the interaction between A and S for one comparison has to be the same as the sum of squares for A X S for another comparison. For example...

Let's say that we want to do two comparisons for this example. We want to test the prediction that subjects will perform less well with a low dose of caffeine than with no caffeine. Second, we want to test the prediction the subjects getting a high dose of caffeine will perform less well than subjects getting a low dose of caffeine. These comparisons are a1 vs a2 and a2 vs a3. Obviously, the only subjects that are relevant to

a1 vs a2 are the subjects in the no caffeine and low dose of caffeine conditions. Let's plot their data alone...



Let me ask you to rate the degree to which Caffeine Dosage interacts with Subjects for this comparison. How much do the lines deviate from being parallel? How about on a scale of one (no interaction at all) to ten (a perfect interaction)? You might reasonably give this a three. This means that this comparison deserves to have an error term that is a relatively small number. Now, let's plot the data for the second comparison, a2 vs a3.



Rate this interaction on the same scale of one to ten. What would you give this one? You might reasonably give this interaction an eight. The lines are clearly far from being parallel. The effect of caffeine dosage is great deal less consistent across subjects for this second comparison than for Comparison One. This means that the second comparison deserves to have an error term that is a relatively large number. If we kept using the error term to test the overall effect of Caffeine Dosage, we'd have used an error term that was quite a bit different for the number that was really appropriate. **If we used error term for the overall effect to test the first comparison we'd use a number that's a lot larger than it needs to be.** We'd end up with an observed value for F that's larger than it deserves to be, and we'd be less likely to reject the null hypothesis. **If we used the error term for the overall effect to test the second comparison we'd use a number that's a lot lower than it needs to be.** In this case we'd end up with an observed value for F that's larger than it really deserves to be.

Each comparison should be tested using an error term that reflects just the amount of error that exists among the scores that are relevant to that comparison. There is no assumption of homogeneity of variance to keep these error terms relatively close to each other. The AXS for one comparison can be whoppingly different from the A X S for another comparison. And it doesn't violate ANY assumption. That's just the way it is.

The implication of all this is that every effect involving the within-subjects factor has to have its own tailor-made error term to go with it. This means that the comparison of a1 vs a2 is essentially a one-way within-subjects ANOVA with only two levels of the independent variable. It's conducted only using the scores in levels one and two. It's a 2 (Caffeine Dosage) by 5 (Subjects) within-subjects ANOVA. The ANOVA table for this first comparison is presented below.

Source	SS	df	MS	F(observed)	F(critical)
A	49000	1	49000	326.67	5.32
S	43040	4			
A X S	600	4	150		

The second comparison (a2 vs a3) is a whole new within-subjects ANOVA with two levels of Caffeine Dosage. Again, it's conducted using only the scores in levels two and three. The ANOVA table for the second comparison is presented below. Notice that the MS for A X S for the two comparisons are wildly different from each other. And both are pretty different from the error term used to test the overall effect of Caffeine Dosage.

Source	SS	df	MS	F(observed)	F(critical)
A	250.0	1	250.0	0.087	5.32
S	39740.0	4			
A X S	11500.0	4	2875.0		