Keppel, G. & Wickens, T. D. *Design and Analysis*
Chapter 3: Variance Estimates and the *F* Ratio

3.1 Completing the Analysis

• Computation of the *F*-ratio requires that you first compute two variances (called mean squares in ANOVA), whose ratio forms the *F*-ratio. The mean squares are determined by dividing the *SS* by the appropriate degrees of freedom. All of the component parts are illustrated in a summary table, as seen in Table 3-1.

Source	Bracket Term	*SS*	*df*	*MS*	*F*
A	$[A] = \frac{\sum A^2}{n}$	$[A] - [T]$	$a - 1$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{S/A}}$
S/A	$[Y] = \Sigma Y^2$	$[Y] - [A]$	$(a)(n - 1)$	$\frac{SS_{S/A}}{df_{S/A}}$	
Total	$[T] = \frac{T^2}{(a)(n)}$	$[Y] - [T]$	$(a)(n) - 1$		

• The degrees of freedom (*df*) represent the "number of scores with independent information that enter into the calculation of the sum of squares." Essentially, you subtract a *df* for every parameter you must estimate from the number of observations. The formula in the above table for $df_{S/A}$ is not as informative as the realization that the df really represents the summed df for each separate group. Thus, you are computing $df = n - 1$ for each of the a groups, then summing over those *df*. Similarly, you are summing over the *SS* for each group to obtain the $SS_{S/A}$.

• There are a number of psychologists who recommend that we abandon the ANOVA and focus instead on confidence intervals. I doubt that we'll see such a shift in analytic approach anytime soon, but there is some good logic behind the proposal. For example, if you computed confidence intervals for two sample means and found that they did not overlap at all, then it seems extraordinarily likely that the two sample means were drawn from populations whose means differed. Of course, data are rarely that accommodating. Nonetheless, you should become comfortable with the notion of a confidence interval.

• A typical confidence interval is generated with $\alpha = .05$, or a 95% confidence interval. You would interpret a 95% confidence interval as the range that would include the population mean (μ) 95% of the time (and 5% of the time it will not capture the population mean). The lower and upper confidence values are computed as seen below:

$$\bar{Y}_j - ts_{M_j} \leq \mu_j \leq \bar{Y}_j + ts_{M_j}$$

- s_M is a symbol for the standard error estimate from a sample. So, what you would do is to compute the sample mean ($\overline{Y}_j$) and the sample standard deviation (s). You would compute the standard error for a sample mean as:

$$s_{M_j} = \frac{s_j}{\sqrt{n_j}}$$

- Next, you would select the appropriate t value, which you would multiply by the standard error and then subtract from the sample mean (to get the lower limit) and add to the sample mean (to get the upper limit). For 95% confidence intervals, you would look up the t value for $\alpha = .05$. For 90% confidence intervals, you would look up the t value for $\alpha = .10$.
- For practice, let's try the following examples.

	Sample 1: $\overline{Y} = 10, s = 6, n = 36$		Sample 2: $\overline{Y} = 15, s = 9, n = 36$	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit
95% Confidence				
90% Confidence				

Based on the confidence intervals you construct, would you feel comfortable concluding that the two samples were drawn from populations with different means?

- We have already seen that we can improve estimates of parameters by using multiple samples (and averaging). We could take a similar approach with estimation of confidence intervals. That is, instead of using the specific standard deviation for the sample being used, we could use an estimate of the population standard deviation/variance based on a set of samples that were presumably drawn from the same or similar populations. Of course, such an approach would make sense only if you had some reason to believe that the variability in the samples was similar (homogeneity of variance assumption). For this approach, simply substitute the mean of the sample variances (as $MS_{S/A}$) into the formula below:

$$\text{pooled } s_M = \sqrt{\frac{MS_{S/A}}{n_j}}$$

• Use this approach for the examples below. Note that you need to use the sample variance (s^2) and not the sample standard deviation (s). Furthermore, when looking up the t value, you would use $a(n\text{-}1)$ for df. Of course, if the sample variances are quite similar and n is large, there will be little difference between the two approaches.

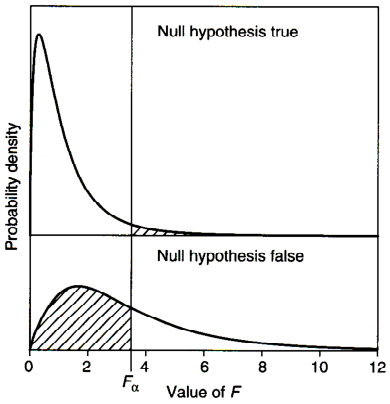
	Sample 1: $\overline{Y} = 10, s^2 = 36, n = 36$		Sample 2: $\overline{Y} = 15, s^2 = 81, n = 36$	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit
95% Confidence				
90% Confidence				

3.2 Evaluating the *F* Ratio

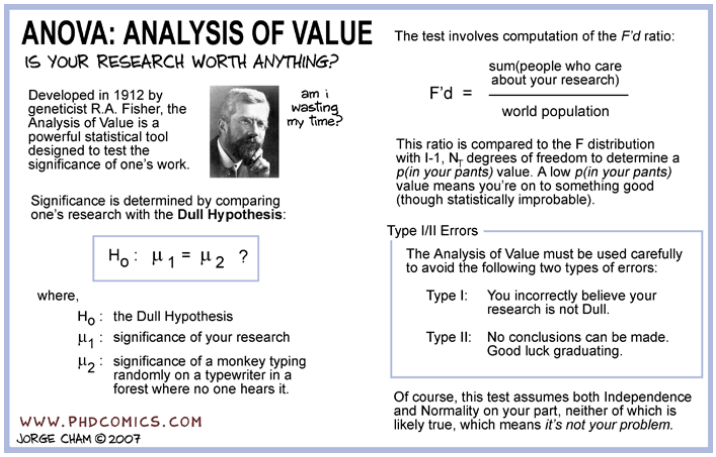
• The F ratio represents two different components of variation. The first component, MS_A, is an assessment of the variability in the data due to treatment effects, which is what we're really trying to assess with ANOVA. However, random effects (individual differences and random variability) also influence the MS_A. If you'll review the notes from Chapter 2, you'll note that the population with greater variability produced a larger MS_A than the population with less variability. Thus, the second component, $MS_{S/A}$, becomes crucial. The $MS_{S/A}$ assesses the variability that one would find in the population due to individual differences and random variability, the population variance (σ^2). The ratio of the two components is revealing:

$$F = \frac{\text{Treatment Effects + Individual Differences + Random Variability}}{\text{Individual Differences + Random Variability}}$$

• When H_0 is true (no treatment effects are present), repeated experiments would yield F ratios around 1.0. Occasionally, however, even though H_0 is false, the F ratios would be larger than one, which causes the F distribution to be positively skewed. When H_0 is false, the distribution of F ratios wouldn't be positively skewed.
• You've already encountered a sampling distribution of a statistic (the sampling distribution of the mean). In a similar fashion, we can generate the sampling distribution of F.
• There is not a single F distribution, but instead a family of curves. The shape of any curve is determined by the two df associated with the distribution.
• We won't actually make much use of the F tables because of our reliance on computer analyses, which print out the probability levels directly. The tables in your text illustrate F_{Crit} for a number of different α levels.
• The shape of the F distribution changes when H_0 is false (see K&W Fig 3.3 below). This distribution is actually called F' or noncentral F. Because we don't know the level of treatment effects present in a study, we have no real idea of the F' distribution.



- In an experiment, our $F_{Observed}$ must have come from either the F or the F' distribution. To reject the possibility that the $F_{Observed}$ was drawn from the F distribution, we determine a value that cuts off the upper 5% of the F distribution ($\alpha = .05$). When $F_{Observed}$ exceeds the critical value, we reject H_0. Otherwise we retain H_0.
- Dictated largely by tradition (and journal editors), most people use $\alpha = .05$ as their significance level.
- We'll talk about concerns that many people have expressed regarding Null Hypothesis Significance Testing (NHST). K&W argue that NHST remains useful. NHST "serves the important function of allowing a researcher to decide whether the outcome of an experiment could have reasonably happened by chance. Thus, it provides a filter to prevent us from wasting our time talking about things that may well be unstable or irreproducible."
- As an antidote to improper use of NHST, K&W offer five suggestions:
  1. Plot your data and look at them, trying to determine what they might mean.
  2. Not all null hypotheses are theoretically interesting, so rejecting them may not be informative.
  3. Do not confuse a significant result with effect size or the probability of replication.
  4. Ask yourself what a significant effect might mean. (Interpretation is usually difficult!)
  5. Interpret your results as an incremental contribution to a larger literature.



3.3 Errors in Hypothesis Testing

- "The procedures we follow in hypothesis testing do not guarantee that the decision rule will produce a correct inference." Ah, yes, the dreaded uncertainty of decision-making. Of course, we're talking about the notion of Type I errors (α errors) and Type II errors (β errors).

	Reality	
Decision ↓	**H_0 True, H_1 False**	**H_0 False, H_1 True**
Reject H_0, Accept H_1	Type I error α error (probability set at α)	Correct decision Power (probability 1- β = ?)
Retain H_0, Do not accept H_1	Correct decision (probability set at 1 - α)	Type II error β error (probability β = ?)

- When H_0 is true, we get the usual *F* distribution, for which we can specify the probabilities associated with particular outcomes of *F* ratios. For the most part, we would set α to .05, which means that we're willing to tolerate a 5% chance of Type I error.
- When H_0 is false, we get the *F'* distribution, for which we can only make guesses about the probabilities involved because we don't know the extent of the treatment effects. As we'll see later, we try to achieve a level of power that will insure that our probability of Type II error (β) is some acceptable level—typically around 20%.
- Power is the "probability of rejecting the null hypothesis when a specific alternative hypothesis is true." Of course, power is the complement of Type II error.
- K&W have more to say about increasing power later, but at this point they mention two possibilities: "to add to the number of observations in each treatment condition, and…to reduce error variance with a more precisely controlled experiment."
- K&W provide a concrete illustration of Type I and Type II errors. In dividing the set of 30 scores into two groups, the two groups should have similar means. That is, there is no treatment effect, so the two sample means should be similar (and near 9). That they are not identical (mean of 9 for both samples) for most of the participants is no real surprise, but illustrates the kind of variability that we typically find in research. Nonetheless, for 3 of the 45 participants, their two means are sufficiently discrepant that an *F* computed on the two means would be significant. The subsequent decisions to reject H_0 ($\mu_1 = \mu_2$) for those three sets of data would be considered Type I errors. And the rate is about what you'd expect (.05 * 45 = 2.25). When a "treatment" was added (+3 to scores in one group), you'd then expect your *F* to be big enough to detect the treatment. However, with that small a treatment effect (relative to error variability), only 16 of the *F*'s were significant (the other 29 *F*'s would be Type II errors). Thus, the probability of a Type II error in this case would be 29/45 = .64 (and so power would be .36).

3.4 A Complete Numerical Example

• Ahhh, here is the ubiquitous K&W p. 51 example. The experiment is a vigilance task, where people look for changes on a screen. The IV is the amount of sleep deprivation (4, 12, 20, or 28 hours) and the DV is the number of failures to spot the target in a 30-min period. The data, summary statistics, etc. are shown below.

	HOURS WITHOUT SLEEP (FACTOR A)				
	4 hrs. a_1	12 hrs. a_2	20 hrs. a_3	28 hrs. a_4	
	37	36	43	76	
	22	45	75	66	
	22	47	66	43	
	25	23	46	62	Sum
Sum (A)	106	151	230	247	734
Mean ($\bar{Y}_A$)	26.5	37.75	57.5	61.75	
ΣY^2	2962	6059	13946	15825	38792
Variance (s^2)	51	119.58	240.33	190.92	601.83
Standard Deviation (s)	7.14	10.94	15.5	13.82	

• Okay, here are the three bracket terms:

$$[T] = \frac{T^2}{an} = \frac{734^2}{16} = 33,672.25$$

$$[A] = \frac{\sum A^2}{n} = \frac{106^2 + 151^2 + 230^2 + 247^2}{4} = 36,986.5$$

$$[Y] = 38,792$$

• The source table would look like this:

Source	Bracket terms	*SS*	*df*	*MS*	*F*
A	[*A*] - [*T*]	3314.25	3	1104.75	7.34
S/A	[*Y*] - [*A*]	1805.50	12	150.46	
Total	[*Y*] - [*T*]	5119.75	15		

• To test whether H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ or H_1: Not H_0 is more reasonable, we'd select $\alpha = .05$, which leads us to an $F_{Crit}(3,12) = 3.49$. Because $F_{\text{Observed}} \geq F_{Crit}$, we would reject H_0. We could then conclude that sleep deprivation has an impact on the number of missed targets in a vigilance task.

• Note 1: Average the variances for the four groups (601.83 / 4) and you get $MS_{S/A}$, or 150.46. This is a very important conceptual fact. You are attempting to estimate σ^2, which you do

by averaging the separate s^2 from each group. The computational approach (using the Bracket Terms) potentially obscures this important fact.

- Note 2: The variance of the 4 means (276.19) multiplied by the sample size (n) equals MS_A, or 1104.75. Once again, this is an important conceptual fact. It points out that the MS_A is really assessing the variability of the group means.
- Suppose that you had the same 4 group means, but the data were less variable, as seen below:

a_1	a_2	a_3	a_4
28	37	58	61
27	37	58	61
26	38	57	62
25	39	57	63

- Given the same 4 group means, but decreased variability within each group, can you predict how the F ratio (and the source table) will change without doing the computations?

Complete as much of the source table as you can:

Source	***SS***	***df***	***MS***	***F***
A				
S/A				
Total				

You should be able to complete the source table if I tell you that the $SS_{S/A} = 11.5$.

- Below is an exercise similar to that found in K&W's Table 2.3, but for the data in K&W51.

Group	Score	$Y - \bar{Y}_A$	$\bar{Y}_A - \bar{Y}_T$	$Y - \bar{Y}_T$
a_1	37	37 - 26.5 = 10.5	26.5 - 45.875 = -19.375	37 - 45.875 = -8.875
a_1	22	22 - 26.5 = -4.5	26.5 - 45.875 = -19.375	22 - 45.875 = -23.875
a_1	22	22 - 26.5 = -4.5	26.5 - 45.875 = -19.375	22 - 45.875 = -23.875
a_1	25	25 - 26.5 = -1.5	26.5 - 45.875 = -19.375	25 - 45.875 = -20.875
a_2	36	36 - 37.75 = -1.75	37.75 - 45.875 = -8.125	36 - 45.875 = -9.875
a_2	45	45 - 37.75 = 7.25	37.75 - 45.875 = -8.125	45 - 45.875 = -0.875
a_2	47	47 - 37.75 = 9.25	37.75 - 45.875 = -8.125	47 - 45.875 = 1.125
a_2	23	23 - 37.75 = -14.25	37.75 - 45.875 = -8.125	23 - 45.875 = -22.875
a_3	43	43 - 57.5 = -14.5	57.5 - 45.875 = 11.625	43 - 45.875 = -2.875
a_3	75	75 - 57.5 = 17.5	57.5 - 45.875 = 11.625	75 - 45.875 = 29.125
a_3	66	66 - 57.5 = 8.5	57.5 - 45.875 = 11.625	66 - 45.875 = 20.125
a_3	46	46 - 57.5 = -11.5	57.5 - 45.875 = 11.625	46 - 45.875 = .125
a_4	76	76 - 61.75 = 14.25	61.75 - 45.875 = 15.875	76 - 45.875 = 30.125
a_4	66	66 - 61.75 = 4.25	61.75 - 45.875 = 15.875	66 - 45.875 = 20.125
a_4	43	43 - 61.75 = -18.75	61.75 - 45.875 = 15.875	43 - 45.875 = -2.875
a_4	62	62 - 61.75 = .25	61.75 - 45.875 = 15.875	62 - 45.875 = 16.125
		If you square each of these differences and sum them, you'll get 1805.5 ($SS_{S/A}$)	If you square each of these differences and sum them, you'll get 3314.25 (SS_A)	If you square each of these differences and sum them, you'll get 5119.75 (SS_{Total})

This exercise is an attempt to clarify the nature of the ANOVA, showing how each of the scores contributes to each *SS*. Note that SS_A is influenced only indirectly by individual scores, because the only scores that contribute are the group means.

- You can readily produce a graph that shows the means with 95% confidence intervals around the means. The procedure for getting a computer program to display the data this way is usually straightforward. (In SPSS, you need to use the Graph menu and then Error Bars…) Note, however, that the error bars that K&W use in Fig 3.4 are standard errors and not 95% confidence limits.

- To compute 95% confidence intervals, you first need to look up the critical value of *t* for the appropriate *df*. For K&W51, with $n = 4$, assuming homogeneity of variance would lead you to pool the variances, so the $df = ((4\text{-}1) * 4) = 12$. Thus, $t_{Crit}(12) = 2.18$.
- Next, compute the standard error estimate:

$$\hat{\sigma}_M = \sqrt{\frac{MS_{S/A}}{n}} = \sqrt{\frac{150.46}{4}} = 6.13$$

- The confidence interval is $\pm[\, t_{Crit} * \hat{\sigma}_M]$, which in this case would be $\pm[2.18 * 6.13] = 13.36$.
- Thus, for the 4 hr deprivation group, with a mean of 26.5, the upper limit would be 39.9 and the lower limit would be 13.14. You would use the same interval for all four groups.
- It's also possible to determine confidence intervals using the variability within the sample for which the confidence interval is being computed. For the 4 hr deprivation group, with a standard deviation of 7.14, the standard error estimate would be 3.57. The $t_{Crit}(3) = 3.18$, so the confidence interval is $\pm[3.18 * 3.57] = 11.35$. Using this approach, for the 4 hr deprivation group, the upper limit would be 37.85 and the lower limit would be 15.15.
- Which approach (separate variance for each sample or pooled sample variance) makes more sense to you? We'll return to this discussion in a later chapter.

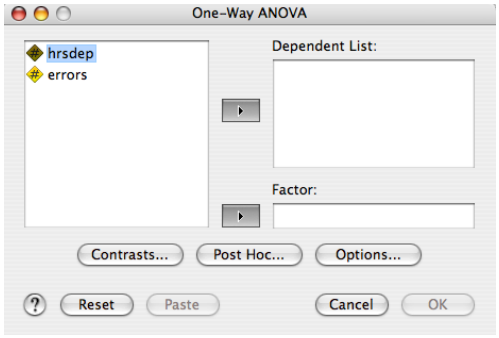
3.5 Unequal Sample Sizes

- In general, you should try to achieve equal (or roughly equal, with large *n*) sample sizes in your groups. In part, "equal sample sizes reduce any problems associated with violations of the assumptions underlying the analysis of variance."
- That said, of course, you may well find yourself in a situation where the sample sizes are not equal. If the sample sizes are large and if the discrepancy in size is not too great, you're probably in good shape. The computations are actually fairly straightforward, as K&W show. However, given that you're likely to be conducting most analyses with a computer program, you should probably determine exactly how the program is dealing with the analysis of unequal sample sizes.

SPSS Analysis

• There are two ways to approach the analysis of a single factor independent groups design in SPSS. One approach is to input the levels of your factor as numerals. Thus, for K&W51, you could use 1 = 4 hours of deprivation, 2 = 12 hours, 3 = 20 hours, and 4 = 28 hours. In *Data View*, your data would look like the portion of the window seen below left:

	hrsdep	errors
1	1	37
2	1	22
3	1	22
4	1	25
5	2	36
6	2	45
7	2	47
8	2	23
9	3	43
10	3	75
11	3	66
12	3	46
13	4	76
14	4	66
15	4	43
16	4	62



• This approach is not optimal, because you need to somehow record the meaning of the numbers representing the levels of your factor. However, I think that it's the only way to make use of the One-Way analysis in SPSS. Once your data are entered, to select that analysis, choose Analyze->Compare Means->One-Way ANOVA. You'll then see a window like the one above right.

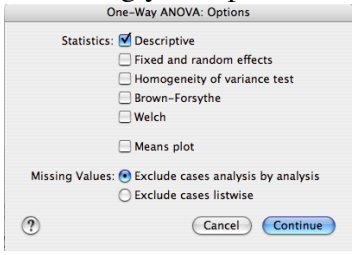
• To create the analysis, you'd need to tell SPSS what your Factor variable is (*hrsdep*) and what variable contains your scores (*errors*). Do so be selecting the variable on the left (by clicking on it) and then click on the appropriate arrow (to move the variable to Factor or to Dependent List).

• Without further instruction, the ANOVA table generated will look like this:

ANOVA
ERRORS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3314.250	3	1104.750	7.343	.005
Within Groups	1805.500	12	150.458		
Total	5119.750	15			

• However, if you want to know the means, etc., you can instruct SPSS to compute such information for you. Simply click on the Options button in the window above right and you'll see the following screen showing your options:

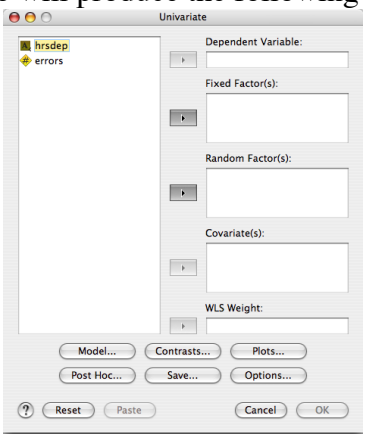


By choosing Descriptive, you'll get an output that looks like this one (in addition to the ANOVA table):

Descriptives
ERRORS

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	4	26.50	7.141	3.571	15.14	37.86	22	37
2	4	37.75	10.935	5.468	20.35	55.15	23	47
3	4	57.50	15.503	7.751	32.83	82.17	43	75
4	4	61.75	13.817	6.909	39.76	83.74	43	76
Total	16	45.88	18.475	4.619	36.03	55.72	22	76

- If you want to be able to use string variables to label the levels of your IV, you'll have to use a different approach. For the same data, I've changed the *hrsdep* variable to "string" (in the Variable View window). Doing so allows me to label the levels appropriately, which means that they will be labeled clearly in any output. I've chosen to use 4hrs, 12hrs, 20hrs, and 28hrs. For the analysis, you would choose Analyze->General Linear Model->Univariate. Doing so will produce the following window:



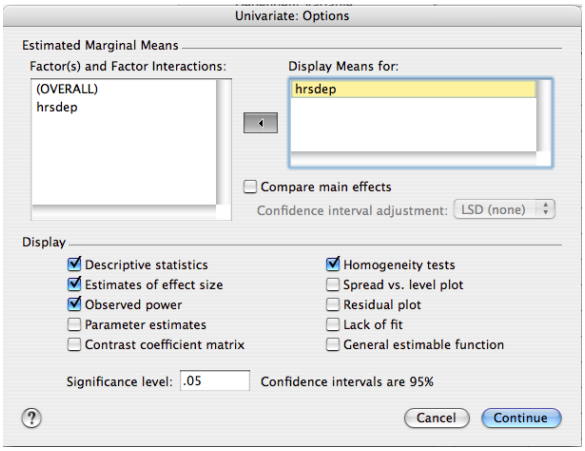
- Once again, move *errors* to the Dependent Variable window and *hrsdep* to the Fixed Factor(s) window. This window is a bit complex, with a number of different buttons for selecting different options, etc. If you do no more than move your variables appropriately and click on the OK button, you'll get a slightly more complicated ANOVA source table, as seen below:

Tests of Between-Subjects Effects
Dependent Variable: ERRORS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3314.250	3	1104.750	7.343	.005
Intercept	33672.250	1	33672.250	223.798	.000
HRSDEP	3314.250	3	1104.750	7.343	.005
Error	1805.500	12	150.458		
Total	38792.000	16			
Corrected Total	5119.750	15			

a R Squared = .647 (Adjusted R Squared = .559)

• If you simply ignore the Corrected Model, Intercept, and Corrected Total lines, you'll be looking at an ANOVA table that should seem quite familiar to you. Moreover, you'll be able to tailor your output to your needs by selecting other types of output. In this analysis, you have quite a few options available to you, as seen below:



By choosing the ones seen above, the output will now be a bit more complex. You'll see information about your group means:

Descriptive Statistics
Dependent Variable: ERRORS

HRSDEP	Mean	Std. Deviation	N
12hrs	37.75	10.935	4
20 hrs	57.50	15.503	4
28 hrs	61.75	13.817	4
4hrs	26.50	7.141	4
Total	45.88	18.475	16

In addition, your source table will now be more complex:

Tests of Between-Subjects Effects
Dependent Variable: ERRORS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Corrected Model	3314.250	3	1104.750	7.343	.005	.647	22.028	.930
Intercept	33672.250	1	33672.250	223.798	.000	.949	223.798	1.000
HRSDEP	3314.250	3	1104.750	7.343	.005	.647	22.028	.930
Error	1805.500	12	150.458					
Total	38792.000	16						
Corrected Total	5119.750	15						

a Computed using alpha = .05
b R Squared = .647 (Adjusted R Squared = .559)

Nonetheless, the source table should still be familiar to you, with all the essentials you're used to seeing.