

One-way Repeated-Measures ANOVA

Relevant research questions and data requirements

Research question: An ANOVA (Analysis Of Variance) is a mean-difference test. In a one-way repeated-measures (RM) ANOVA we *simultaneously compare* the means of *two or more conditions* that are the result of manipulating *one independent variable* (hence “one-way” ANOVA). Importantly, every participant is exposed to each condition (i.e., a within-subjects design). As such, it is an extension of the dependent-samples t-test. While the dependent-samples t-test can “only” compare 2 conditions, a RM-ANOVA can compare two or more conditions at the same time. It calculates the ratio of variability between groups and within groups. If the variability between groups exceeds the variability within groups, this may be evidence of a treatment effect.

As an example, let’s say we want to know if some toys are more appealing to preschoolers than others. We present 15 preschoolers with three different toys (a coloring book, a marble run, and a puzzle) and measure (in minutes) how long each child plays with each toy. In other words, we get 3 data points from each child because every child gets *every treatment* (i.e., this is a within-subjects design).

How many conditions? Two or more. Remember: ANOVA can do anything a t-test can do, but a t-test is limited to comparing 2 groups!

Data requirements? Interval/ratio outcome data (here: minutes of playing with toy), which are (roughly) normally distributed. It is also assumed that the variance of the difference scores are roughly the same (e.g., $VAR(\text{puzzle} - \text{marble run}) = VAR(\text{puzzle} - \text{coloring book}) = VAR(\text{coloring book} - \text{marble run})$). This is called sphericity (try saying it out loud three times, fast!). If this assumption is violated, we risk an increased type I error, so we need to correct for that (more on this below). Also check for outliers.

Checking our assumptions

First, let’s check our assumptions and have a look at descriptives and the distribution of scores in each condition. (For more detailed information, refer to the Descriptives worksheet.)

Notice that the setup of our data set is slightly different: Each child still gets its own row, but instead of having just one outcome column, we now have three... one for each condition.

| ID | coloringbook | marblerun | puzzle |
|----|--------------|-----------|--------|
| 1 | 7 | 5 | 2 |
| 2 | 3 | 6 | 4 |
| 3 | 2 | 4 | 2 |
| 4 | 4 | 6 | 3 |
| 5 | 5 | 3 | 4 |
| 6 | 6 | 4 | 3 |
| 7 | 4 | 5 | 3 |
| 8 | 1 | 4 | 3 |
| 9 | 3 | 2 | 1 |
| 10 | 5 | 3 | 4 |
| 11 | 3 | 8 | 4 |
| 12 | 5 | 4 | 1 |
| 13 | 4 | 7 | 5 |
| 14 | 5 | 5 | 4 |
| 15 | 4 | 3 | 4 |

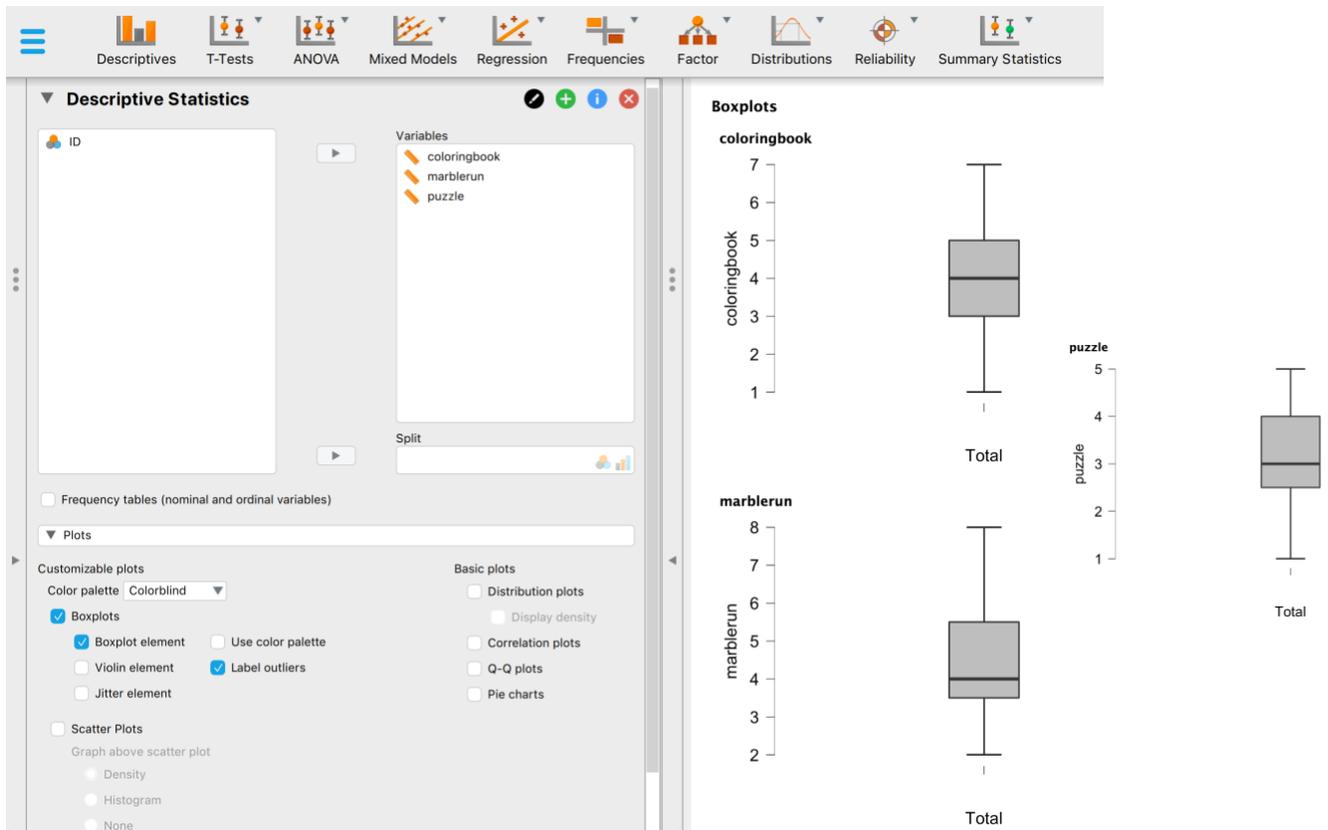
The screenshot shows the JASP software interface. On the left, the 'Descriptive Statistics' panel is active, with 'ID' selected as the variable. The 'Variables' list on the right contains 'coloringbook', 'marblerun', and 'puzzle'. The 'Statistics' section is expanded, showing options for Percentile Values, Dispersion, Central Tendency, and Distribution. The 'Central Tendency' section has 'Mean' selected. The 'Dispersion' section has 'Std. deviation', 'Minimum', and 'Maximum' selected. The 'Distribution' section has 'Shapiro-Wilk test' selected. On the right, the 'Results' panel displays a table of Descriptive Statistics for the three variables. The 'P-value of Shapiro-Wilk' row is circled in red, with a red arrow pointing from the text below to it.

| | coloringbook | marblerun | puzzle |
|-------------------------|--------------|-----------|--------|
| Valid | 15 | 15 | 15 |
| Missing | 0 | 0 | 0 |
| Mean | 4.067 | 4.600 | 3.133 |
| Std. Deviation | 1.534 | 1.639 | 1.187 |
| Shapiro-Wilk | 0.968 | 0.955 | 0.891 |
| P-value of Shapiro-Wilk | 0.828 | 0.600 | 0.070 |
| Minimum | 1.000 | 2.000 | 1.000 |
| Maximum | 7.000 | 8.000 | 5.000 |

I have moved our three outcome variables to the “variables” window and asked for M, SD, Min, Max, and Shapiro-Wilk for now.

The table shows us the sample size of each condition, how many missing cases we have, M, SD, Variance, and the Shapiro-Wilk test for normality. Just by eye-balling the means, we can see that “marblerun” has the highest play time, and “puzzle” the lowest, with “coloringbook” inbetween those two. To see if these 3 means differ statistically, we will need to run our RM-ANOVA.

Normal distribution of data. For now, let’s check on the assumption of normality. Recall that the null hypothesis of the Shapiro-Wilk test is that “normality is met” – hence, any p-value below .05 should alert us to deviations from normality. This is not the case for any condition, so we are good to go. (Recall, ANOVA is generally fairly robust to violations of normality, but if this assumption is violated, you can run the nonparametric equivalent, the Friedman’s test, which is an option at the very bottom of the RM-ANOVA screen in JASP, as we will see below.)

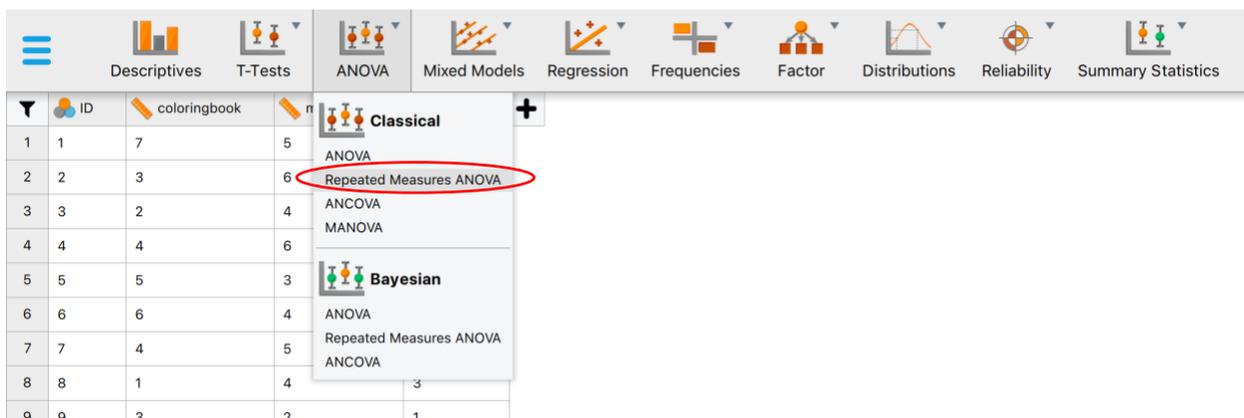


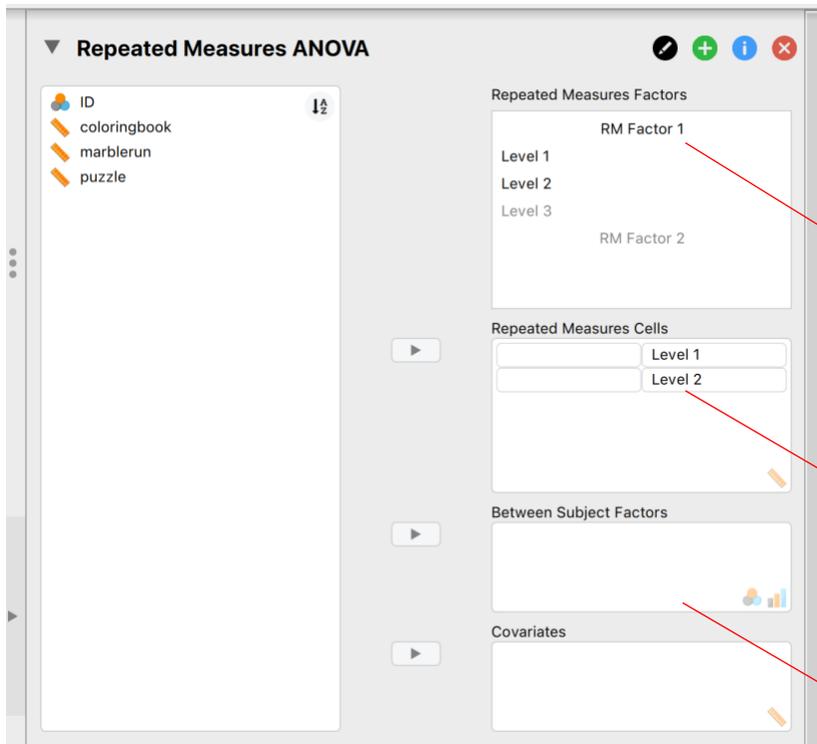
Outliers. To check for outliers, I asked for boxplots and told JASP to label outliers. Looking at our boxplots, we see no outliers (they would be marked with a black dot), but get a more visual representation of the 3 different groups’ distribution of scores.

Sphericity. This cannot be checked under “Descriptives.” Proceed to the ANOVA (below) to check for this assumption.

Running the test in JASP

Next, to proceed with our analysis, click on “Repeated-measures ANOVA” in the main menu and select “ANOVA” under “Classical” (we will ignore all Bayesian analyses for now).





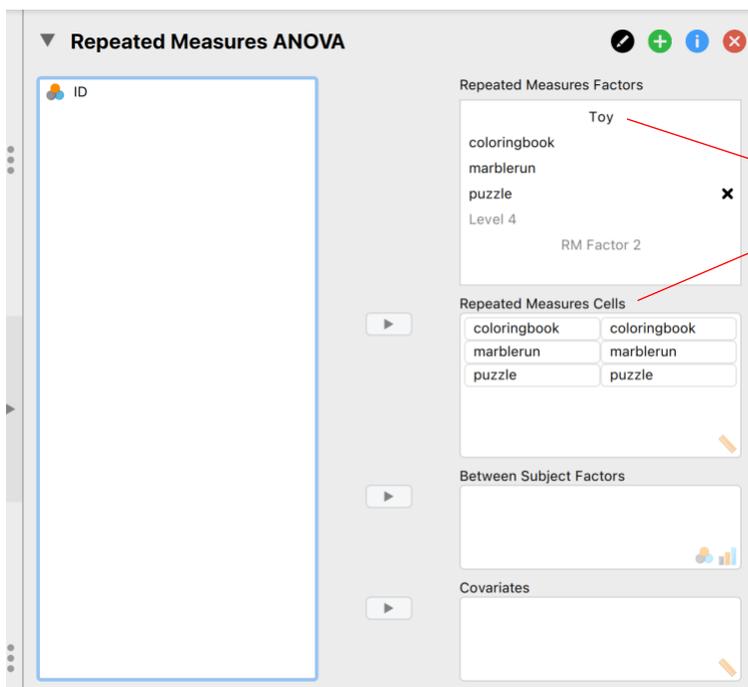
JASP wants a bit more information before we can move our variables of interest into the appropriate windows.

We need to tell JASP what the name of our repeated-measures factor is and what its levels are. In our example, the factor is “toys” and the three levels are “coloringbook”, “marblerun”, and “puzzle.”

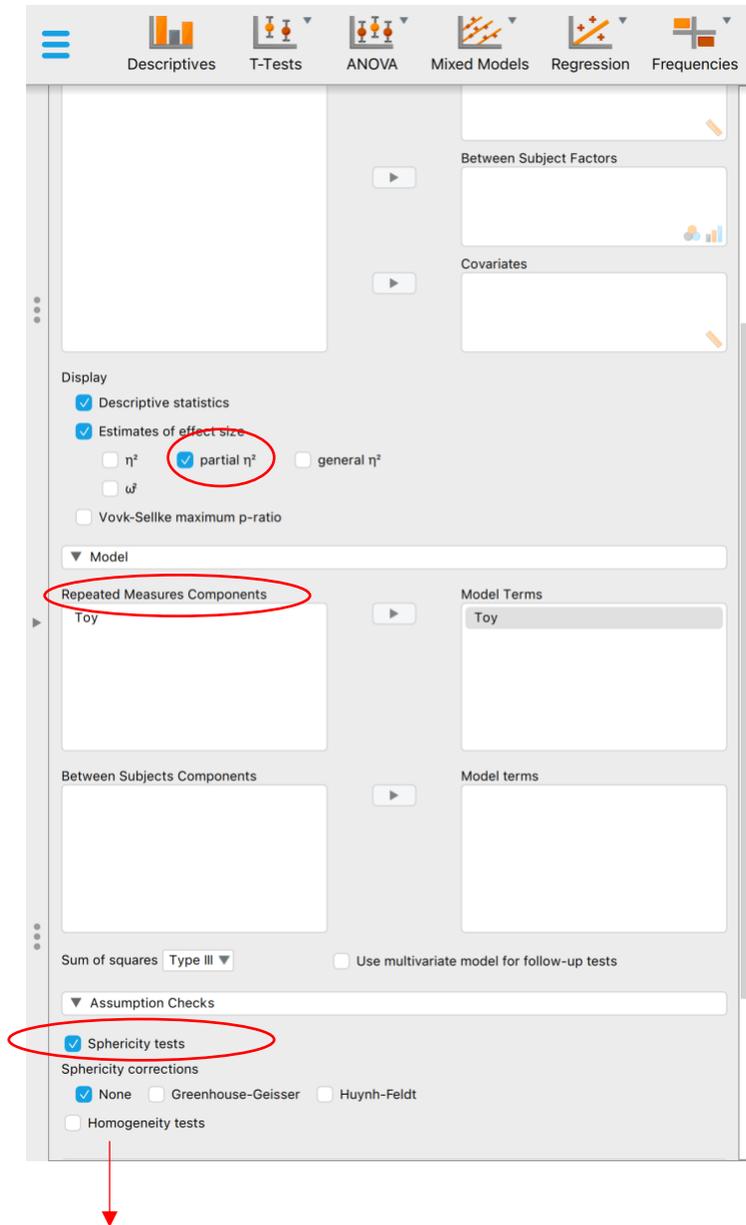
We type that information into the top window, and then move our conditions into the appropriate cells in the “Repeated-measures cells” window. Again, JASP reminds us that the variables here need to be scale by showing us the little ruler to the right.

For a one-way RM-ANOVA, we can ignore the “between-subjects” and “covariates” fields.

When it’s all said and done, your screen might look something like this:



Here, our factor and conditions are labeled and the conditions were moved to “repeated-measures cells.” We are now ready to specify the statistics we want to run.



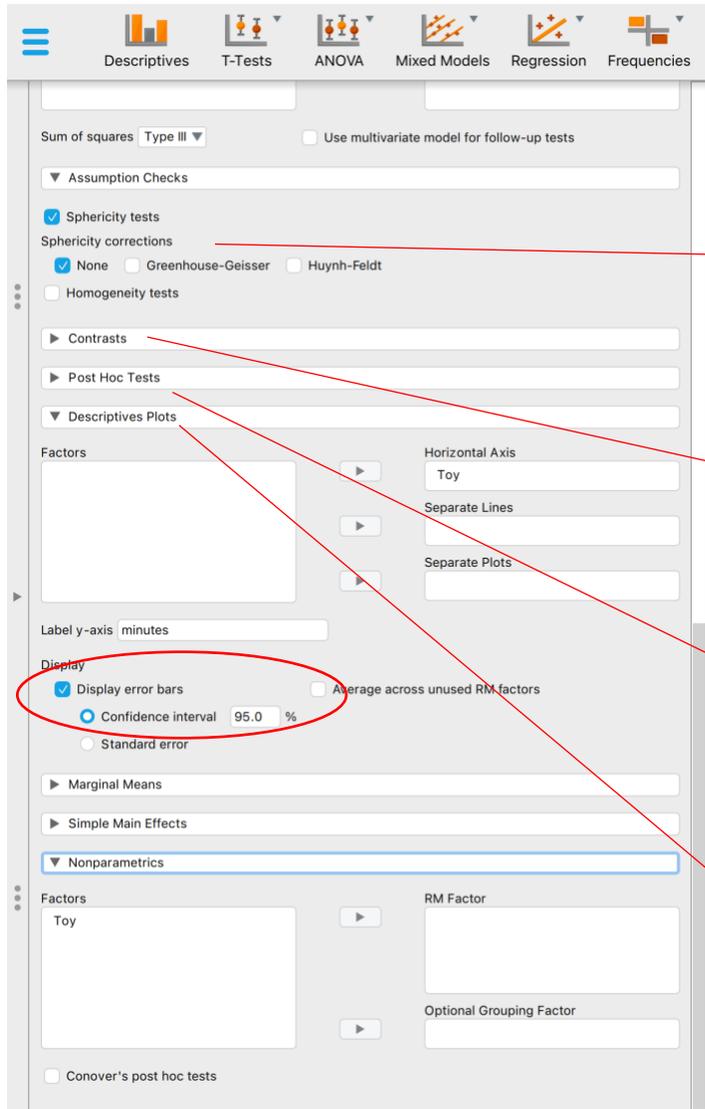
The top part of this is shown on the last image.

Under **Display**, select “Descriptive Statistics” and “Estimates of effect size” and “partial eta squared” (this is the preferred effect size for RM ANOVA)

Under **Model**, select your repeated-measures IV (toy) and move it over to “Model Terms” ... notice, there is a field for repeated-measures components and one for between-subjects components right below it. Here we only have a repeated-measured factor (toy) so we can ignore the between-subjects components fields.

Under **Assumption Checks**, check “Sphericity tests” – this tests the null hypothesis that the variances of the differences between the groups are roughly equal (i.e., that our assumption is met). This is the last assumption we need to check.

We can ignore the “homogeneity tests” because that is for between-subjects factors and we don’t have any of those in this example.



Assumption Checks, continued: IF your sphericity assumption is violated, you can go under “Sphericity Corrections” and select “Greenhouse-Geisser” and “Huynh-Feldt”, both of which correct the df to make up for the violation and to keep the type I error low. Both will yield an epsilon value. Report the Greenhouse-Geisser correction if the epsilons are below .75; otherwise report the Huynh-Feldt correction.

Contrasts: These are specific ways to compare the different groups. We don’t teach this at the undergraduate level, so you can skip this.

Post-Hoc Tests: IF your overall ANOVA is significant, you will want to run post-hoc tests to see *which conditions* differ (the overall test only asks *if* any of the conditions differ). We will return to this.

Descriptive Plots: Select this if you would like a visual representation of your means and display error bars as 95% Confidence Intervals

Marginal Means: This becomes relevant in a factorial ANOVA (i.e., when you have two or more IVs). Because we are running a one-way ANOVA here, we can skip this. Our marginal means are identical to the means of the three groups that we get by selecting “Descriptive Statistics” above.

Simple Main Effects: Simple main effects help us make sense of interactions in factorial ANOVAs (i.e., ANOVAs with 2 or more IVs). This, too is irrelevant for a one-way ANOVA.

Nonparametrics: This is where you could run the Friedman test (the nonparametric equivalent to a RM-ANOVA) if you have an ordinal DV, or if your data vastly violate the assumption of normality (again, though, ANOVA, especially with larger samples, is believed to be robust against this violation). Still, you may run this nonparametric test and report it alongside the ANOVA to let your reader know if the parametric and nonparametric analyses are different or identical in their conclusion.

Reading and understanding the output

Assumption Checks ▾

Test of Sphericity

| | Mauchly's W | Approx. X ² | dfSphericity | p-value | Greenhouse-Geisser ε | Huynh-Feldt ε | Lower Bound ε |
|-----|-------------|------------------------|--------------|---------|----------------------|---------------|---------------|
| Toy | 0.799 | 2.919 | 2 | 0.232 | 0.833 | 0.931 | 0.500 |

Even though Assumption checks are reported at the bottom of the output, it is important to look at this first. Here we see that $p = .232$, so our null hypothesis that the variances of the differences are equal cannot be rejected, so our assumption is met. IF $p < .05$, check the epsilons reported next to the p-value. Here, they are .83 and .93 – so, both are above .75, which means we would want to ask JASP to report the Huynh-Feldt correction under Sphericity corrections as indicated on the previous page to correct for the violation of the assumption. This will be reported as a separate line in the output.

Results ▾

Repeated Measures ANOVA ▾

Within Subjects Effects

| Cases | Sum of Squares | df | Mean Square | F | p | η_p^2 |
|-----------|----------------|----|-------------|-------|-------|------------|
| Toy | 16.533 | 2 | 8.267 | 4.276 | 0.024 | 0.234 |
| Residuals | 54.133 | 28 | 1.933 | | | |

Note. Type III Sum of Squares

Between Subjects Effects

| Cases | Sum of Squares | df | Mean Square | F | p |
|-----------|----------------|----|-------------|---|---|
| Residuals | 36.133 | 14 | 2.581 | | |

Note. Type III Sum of Squares

This is our hypothesis test. Shown in order are the SS, df, MS (aka variance, which is SS/df), the value of the F-ratio, the p-value, and our effect size partial eta squared.

Note: For SS, df, and MS, the top row always shows the values between conditions, the bottom row shows values within conditions!

This table is for between-subjects IVs, which we did not have (we only had one repeated-measures IV!). hence, it is empty.

We see that the value of our F-ratio (recall: $F = MS_{\text{between}} / MS_{\text{within}}$) is $F(2,28) = 4.28$, and the **p-value is .024**, which is below .05. Thus, we reject the null hypothesis that $\mu_1 = \mu_2 = \mu_3$. The variability between groups much exceeds the variability within groups, which suggests a treatment effect. In other words: The means of our three groups significantly differ – however, we only know *that* there is a difference among the three toys, not yet *where* it is, i.e. which groups, exactly, differ from one another. For that, we will need to run post-hoc tests.

Partial eta squared is 0.23 and tells us that 23% of the variability in time spent playing can be explained by the type of toy. *Note:* This is considered a large effect. General guidelines say that for partial eta squared, .01 is a small effect, .06 a medium effect, and .14 or above a large effect.

Within Subjects Effects

| Cases | Sphericity Correction | Sum of Squares | df | Mean Square | F | p | η_p^2 |
|-----------|-----------------------|----------------|--------|-------------|-------|-------|------------|
| Toy | None | 16.533 | 2.000 | 8.267 | 4.276 | 0.024 | 0.234 |
| | Huynh-Feldt | 16.533 | 1.863 | 8.876 | 4.276 | 0.027 | 0.234 |
| Residuals | None | 54.133 | 28.000 | 1.933 | | | |
| | Huynh-Feldt | 54.133 | 26.078 | 2.076 | | | |

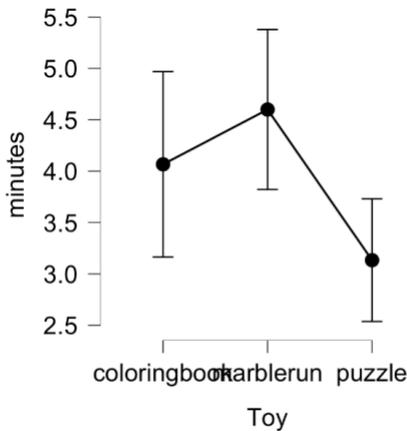
Note. Type III Sum of Squares

FYI: This is what the output would look like if we ran the H-F Sphericity correction. You would report the second row (notice the df changed).

Descriptives ▼

| Toy | Mean | SD | N |
|--------------|-------|-------|----|
| coloringbook | 4.067 | 1.534 | 15 |
| marblerun | 4.600 | 1.639 | 15 |
| puzzle | 3.133 | 1.187 | 15 |

Descriptives plots ▼



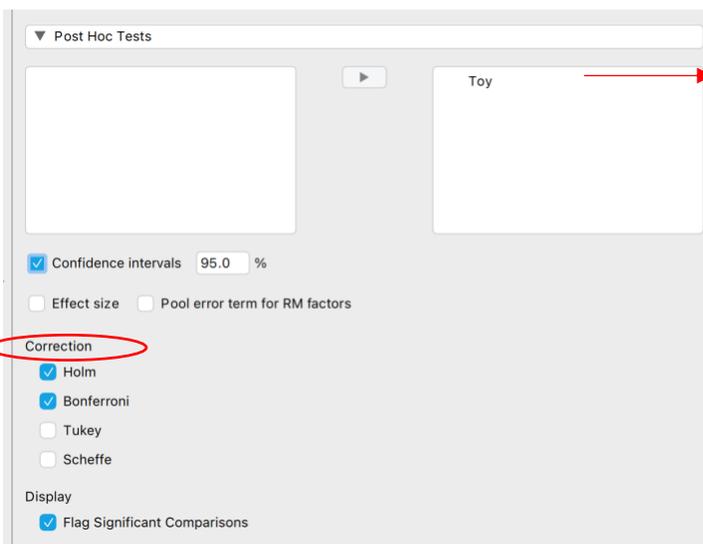
Here are our **descriptive statistics** and descriptive plots. Condition labels will be reflected here, so always label your conditions in a way that makes the output easy to read (here: “coloringbook”, etc. rather than “1”, “2”, etc.).

Our plot shows the means and their 95% confidence intervals. Right away, we can see that time spent playing was lowest in the “puzzle” condition and confidence intervals do not overlap between that and the “marblerun”, so those two groups likely differ from each other because the ranges of true population means don’t overlap. We will test this with post-hoc tests below.

Also notice that the y-axis is scaled from 2.5-5.5, not 0-10, so differences between the groups appear slightly more pronounced in this image than they would if we scaled the y-axis from 0-10 (which JASP does not let us do).

Running and understanding post-hoc tests

Because our overall ANOVA showed that there was a significant difference somewhere between the 3 means, we need to run **post-hoc tests** to see *what groups* differ from each other, exactly. To run those, we go back to our analysis options on the left hand side of the screen and scroll to post-hoc tests.



First, move your IV into the window on the right.

Under post-hoc test corrections, you have several options. For simplicity, here are some general guidelines:

For a simple RM-ANOVA, choose either Holm or Bonferroni (JASP will not run Tukey or Scheffe if you only have a repeated-measures IV, as these are not deemed appropriate tests for repeated measures). Bonferroni is considered very “conservative” (strict), so I suggest using Holm.

Ask for 95% CIs and flag significant comparisons for easier identification in the output.

Post Hoc Tests ▾

Post Hoc Comparisons – Toy ▾

| | | Mean Difference | 95% CI for Mean Difference | | SE | t | P _{bonf} | P _{holm} |
|--------------|-----------|-----------------|----------------------------|-------|-------|--------|-------------------|-------------------|
| | | | Lower | Upper | | | | |
| coloringbook | marblerun | -0.533 | -2.165 | 1.099 | 0.601 | -0.888 | 1.000 | 0.389 |
| | puzzle | 0.933 | -0.431 | 2.298 | 0.502 | 1.859 | 0.253 | 0.168 |
| marblerun | puzzle | 1.467 | 0.377 | 2.556 | 0.401 | 3.659 | 0.008** | 0.008** |

** p < .01

Note. P-value and confidence intervals adjusted for comparing a family of 3 estimates (confidence intervals corrected using the bonferroni method).

Our post-hoc test output shows us the comparison of each pair of groups. The first row compares “coloringbook” to “marblerun”, the second row compares “coloringbook” to “puzzle”, and the last row compares “marblerun” to “puzzle.”

Because we asked significant comparisons to be flagged, we see quickly that only “marblerun” and “puzzle” differ from each other. This supports our hunch from looking at the means plot and its 95% CIs earlier. To remind ourselves which group had higher social attractiveness ratings, we can refer back to the Descriptives. (Notice that Bonferroni and Holm p-values don’t differ for the last comparison, but for the other two, Holm as much lower p-values than Bonferroni – again, Bonferroni is considered to be very strict, so we can trust that we don’t have an inflated Type 1 error, but this comes at the expense of reduced power!)

Writing up results in APA style

A one-way repeated-measures ANOVA indicated that time spent playing differed by type of toy, $F(2,28) = 4.28, p = .024, \eta^2 = 0.23$. Post-hoc comparisons using Holm’s correction showed that children played longer with the marble run ($M = 4.60, SD = 1.64$) than the puzzle ($M = 3.13, SD = 1.19$), $p = .008$. There was no difference in time spent playing between the marble run and the coloring book ($M = 4.07, SD = 1.53$), $p = .39$, nor the coloring book and the puzzle, $p = .17$.

*Note that there are spaces before and after equal signs; M, SD, F, p are italicized; everything is rounded to two digits *except* for p-values, which should be reported exactly as given in the output. Only report leading zeros for values that can exceed 1 (hence, p-values should *not* be reported with leading 0s). For F-ratios, you *always* list the df_{between} first, then the df_{within} .

If there are **three groups** (three conditions), you can probably describe all comparisons (and each group’s M and SD) in the text as shown in this example.

If there are **more than three groups** (3+ conditions), you might want to move to minimal in-text description and then display your findings in a table (again, to give the reader each group’s M and SD). See the “one-way ANOVA between” for an example.