Some Comments and Definitions Related to the Assumptions of Within-subjects ANOVA

The Sphericity Assumption

The sphericity assumption states that the variance (or standard deviation) of the difference scores taken between levels of the repeated measures factor in the population (i.e., σ^2_{Y-Y} for all pairs of difference scores) are all equal. To illustrate the idea in a sample, imagine a within-subjects design with three levels of the independent variable, such as pretest (Y_1), posttest (Y_2), and follow-up (Y_3), and that difference scores are calculated for each subject comparing pretest with posttest, posttest with follow-up, and pretest with follow-up.[1] The sphericity assumption (sometimes called the "circularity" assumption) would imply that the variances of each of these sets of difference scores are not statistically different from one another. The table below illustrates the idea. The sphericity assumption is violated if s^2_{2-1}, s^2_{3-2}, and s^2_{2-1} are not all equal.

Y_1	Y_2	Y_3	Y_2-Y_1	Y_3-Y_2	Y_3-Y_1
53	47	45	-6	-2	-8
49	42	41	-7	-1	-8
47	39	38	-8	-1	-9
42	37	36	-5	-1	-6
51	42	35	-9	-7	-16
34	33	33	-1	0	-1
44	13	46	-31	33	2
48	16	40	-32	24	-8
35	16	29	-19	13	-6
18	10	21	-8	11	3
32	11	30	-21	19	-2
27	6	20	-21	14	-7
			$s^2_{2-1} = 108.73$	$s^2_{3-2} = 154.64$	$s^2_{3-1} = 27.73$

The sphericity assumption is similar to the homogeneity of variance assumption with between-subjects ANOVA in some ways. When this assumption is violated, there will be an increase in Type I errors, because the critical values in the F-table are too small. Here the variances of the difference scores look quite unequal, particularly the last set of difference scores compared with the first two sets. There are two major approaches to dealing with this problem—univariate tests with corrections for sphericity and multivariate tests that do not assume sphericity.

A note on determining sphericity violations. Determining whether there is a sphericity violation for certain is not very feasible. There are tests that attempt to do this, such as Mauchly's chi-square test. But much as with Levene's test for equal variances, it tends to miss sphericity violations for smaller samples and tends to be significant even though the violation is small in magnitude for larger sample sizes or nonnormal distributions (Type I error; e.g., Kesselman, Rogan, Mendoza, & Breen, 1980). Moreover, the univariate and multivariate remedies work well when applied under the right circumstances (see below), so there is not a critical need for determining whether a violation exists.

Univariate Sphericity Corrections

Lower bound correction. This is a correction for a violation of the sphericity assumption. The correction works by using a higher F-critical value to reduce the likelihood of a Type I error. In the non-SPSS world, the lower bound correction is really referred to as the "Geisser-Greenhouse" correction in which $df_A = 1$ and $df_{AxS} = s - 1$ are used instead of the usual $df_A = a - 1$ and $df_{AxS} = (a - 1)(s - 1)$. Under this correction to the dfs, the F-critical values will be larger and it will be harder to detect significance, thus reducing Type I error. Unfortunately, this correction approach tends to overcorrect, so there are too many Type II errors with this procedure (i.e., low power). This correction assumes the maximum degree of heterogeneity among the differences.

[1] The sphericity assumption does not apply to within-subjects ANOVAs that have only two levels.

Huynh & Feldt correction. This correction is based on a similar correction by Box (1954). In both cases, an adjustment factor based on the amount of variance heterogeneity (i.e., how much the variances are unequal) is computed (the adjustment factor is called epsilon). Then, both df_{AxS} and df_A are adjusted by this amount, so that the F-critical will be somewhat larger. The correction is not as severe as the "lower bound" correction.

Geisser-Greenhouse correction. The Geisser-Greenhouse correction referred to in SPSS is another variant on the procedure described above under Huynh & Feldt. A slightly different correction factor (epsilon) is computed, which corrects the degrees of freedom slightly more than the Huynh & Feldt correction. So, significance tests with this correction will be a little more conservative (higher p-value) than those using the Huynh-Feldt correction.

Multivariate Tests (Sphericity Not Assumed)

The multivariate testing approach treats the analysis as a special case of a more general analysis called multivariate analysis of variances (usually abbreviated as “MANOVA” by authors and in software packages). As mentioned earlier, we can think about the repeated measures design as having either multiple levels of a within-subjects factor or as multiple dependent variables. The MANOVA approach does not require the sphericity assumption. MANOVA does assume, however, that the data have a “multivariate” normal distribution—that the analysis variable is jointly normally distributed when all levels are considered together. With larger samples sizes, the MANOVA approach is more powerful than the repeated measures univariate ANOVA. The MANOVA test results are included by default in the output for the GLM repeated measures analysis in SPSS and we can request separately in R. Algina and Kesselman (1997) suggest guidelines for when to use MANOVA instead of the univariate ANOVA with sphericity corrections. Their guidelines are to **use MANOVA if 1) the number of levels is less than or equal to 4 ($a \leq 4$) and n greater than the number of levels plus 15 ($a + 15$); or 2) the number of levels is between 5 and 8 ($5 \leq a \leq 8$) and n is greater than the number of levels plus 30 ($a + 30$).**

Other Assumptions You May Hear About

*Compound Symmetry Assumption.*Another assumption of within-subjects ANOVA that you may hear about is the “compound symmetry” assumption. The compound symmetry assumption is a stricter assumption than the sphericity assumption. Not only do the variances of the difference scores need to be equal for pairs of conditions, but their correlations (technically, the assumption concerns covariances—the unstandardized version of correlation) must also be equal. Imagine taking differences between scores for each possible pair of cells. Then correlations (covariances) are calculated among all those difference scores. Under the compound symmetry assumption these correlations (or covariances, actually) must not be different in the population (e.g., the covariances among the $Y_{2\text{-}1}$, $Y_{3\text{-}2}$, and $Y_{3\text{-}1}$ columns in the table are not significantly different from one another). The compound symmetry assumption is often considered overly restrictive (Edwards, 1985), and a violation of this stricter compound symmetry assumption does not necessarily indicate that the sphericity assumption will be violated. SPSS does not currently provide a test of this assumption within the ANOVA commands.

Nonadditivity. The error term for within-subjects is the interaction term, $S \times A$. In other words, we assume that any variation in differences between levels of the independent variable is due to error variation. It is possible, however, that the effect of the independent variable A is different for different subjects, and there is truly an interaction between S and A. Thus, some of what we consider to be error when we calculate $S \times A$ is really an interaction of subject and treatment and not error variation. For example, if Factor A represents program groups, then an $S \times A$ interaction suggests the program is not equally effective for each subject. This is the so-called “additivity” assumption that there is no interaction between A and S that is not unexplained error (interactions are multiplicative or “nonadditive”). Because nonadditivity (a violation of this assumption) implies hetergeneous variances for the difference scores, the sphericity assumption will be violated if nonadditivity occurs. The Tukey test for nonadditivity (Tukey, 1949) is usually used to test for violations, but alternatives and variants on the test have been proposed to address its low power in some circumstances (see Šimeček, P., & Šimečková, 2013). The Tukey test

for nonadditivity can be obtained in SPSS under the scale reliability command or in R with the `additivityTests` package, which also includes some alternative tests and a modified Tukey test that may have better power in some cases. The SPSS output includes a suggested transformation of the data (raising each score to a certain power) that can be employed if the test is significant.

Comments and Recommendations

Importantly, the sphericity and compound symmetry assumptions do not apply when there are only two levels (or cells) of the within-subjects factor (e.g., pre vs. post test only).[2] The sphericity assumption, for instance, does not stipulate that the variances of the scores for each level of the independent variable are equal, but that the variances of the difference scores, calculated for pairs of levels (e.g., Time 2 scores minus Time 1 scores vs. Time 3 scores minus Time 2 scores), in the population are all equal.

When univariate tests are recommended (small sample sizes), it is a bit difficult to know which test is preferred. But, I'll make a couple of important points regarding the univariate test corrections (based on recommendations of Greenhouse & Geisser): 1) If F is nonsignifcant, do not worry about the corrections, because the corrections will only increase the p-values; 2) If the F is significant using all three approaches, do not worry about the corrections. That is, if the most conservative approach is still significant, there is no increased risk of Type I error. If the various correction tests lead to different conclusions, there is no one perfect solution. I generally use the Huynh and Feldt correction, because it addresses the Type I error problem and is more powerful than the "lower bound" correction. It does not perform perfectly, however, and it might be safest to report the results from all correction approaches when they do not show the same result. With large sample sizes, a small departure from sphericity might be significant, but the correction in these situations should be relatively minor and there is not likely to be difference in the conclusions drawn from the significance tests using the various corrections.

Another possible solution you may hear about is to transform the dependent variable scores using a square root transformation or raise the score to a fractional power (e.g., .33). For the transformation approach, you may have to try different transformations until the sphericity problem is resolved. There are also automated normalizing transformation procedures. The transformation approach may be quite helpful in resolving the problem, but the researcher will havemore difficulty interpreting the results.

Example

Below is an example of a within-subjects ANOVA with three levels of the independent variable which shows the sphericity assumption test and corrections. This hypothetical study compares performance on a vocabulary test after different lecture topics (e.g., physical science, social science, history). Each student hears each lecture topic and takes a vocabulary test afterward. Notice that the adjustments are not made to the calculated F-values. Instead, the corrections are made to adjusted critical values (not shown), so the only difference you may see is in the “Sig” values (p-values). This is accomplished by adjusting the degrees of freedom. The biggest correction to df is the Lower-bound, followed by the Greenhouse-Geiser, and the Huynh-Feldt. Thus, the Lower-bound will have the largest p-value (the most conservative significance test), the Greenhouse-Geisser will have an intermediate p-value, and the Huynh-Feldt will have the smallest p-value (the most liberal significance tests of the corrections).

SPSS

Syntax

```
glm vocab1 vocab2 vocab3
  /wsfactor=vocab 3
  /wsdesign=vocab
  /print=parameter.
```

[2] In SPSS, the Mauchly’s chi-square is reported as 1 and the sig as “.” when there are only two levels of the within-subjects factor.

Mauchly's Test of Sphericity [a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon [b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
vocab	.415	8.789	2	.012	.631	.675	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: vocab

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
vocab	Sphericity Assumed	1194.000	2	597.000	12.305	.000	.528
	Greenhouse-Geisser	1194.000	1.262	946.103	12.305	.002	.528
	Huynh-Feldt	1194.000	1.350	884.582	12.305	.002	.528
	Lower-bound	1194.000	1.000	1194.000	12.305	.005	.528
Error(vocab)	Sphericity Assumed	1067.333	22	48.515			
	Greenhouse-Geisser	1067.333	13.882	76.885			
	Huynh-Feldt	1067.333	14.848	71.885			
	Lower-bound	1067.333	11.000	97.030			

There are a = 3 levels and 12 cases, so there were not more than a + 15 cases in this example, and I would not recommend the use of the MANOVA test. If MANOVA (Multivariate Tests) was appropriate, there are four different tests presented by SPSS, and with a large sample size they will all tend to show the same results. Roy's largest root is the most liberal of the tests and Pillai's trace is said to be the most robust to variance differences with small samples (Olson, 1979). Wilk's lambda seems to be the most commonly reported for some reason.

Multivariate Tests [a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
vocab	Pillai's Trace	.790	18.833 [b]	2.000	10.000	.000	.790
	Wilks' Lambda	.210	18.833 [b]	2.000	10.000	.000	.790
	Hotelling's Trace	3.767	18.833 [b]	2.000	10.000	.000	.790
	Roy's Largest Root	3.767	18.833 [b]	2.000	10.000	.000	.790

a. Design: Intercept
Within Subjects Design: vocab

b. Exact statistic

Below I ran a reliability analysis, **Analyze→ Scale → Reliability analysis**, and chose the *statistics* button and check the *Tukey's test of additivity*. The reliability analysis is usually used to assess internal reliability of a scale and obtain Cronbach's alpha, but we can use it here for the nonadditivity test.

ANOVA with Tukey's Test for Nonadditivity

			Sum of Squares	df	Mean Square	F	Sig
Between People			3531.667	11	321.061		
Within People	Between Items		1194.000	2	597.000	12.305	.000
	Residual	Nonadditivity	93.969[a]	1	93.969	2.027	.169
		Balance	973.364	21	46.351		
		Total	1067.333	22	48.515		
	Total		2261.333	24	94.222		
Total			5793.000	35	165.514		

Grand Mean = 33.50000

a. Tukey's estimate of power to which observations must be raised to achieve additivity = 1.949.

The row of the table labeled "Nonadditivity" gives the test of significance of this assumption violation, F(1,24) = 2.027, ns, and suggests no evidence of the $A \times S$ interaction. The footnote below the table

contains the recommended transformation of the data that could be used if the nonadditivity test was significant.

R

I use the `Manova` from the car package which also prints the univariate ANOVA with sphericity corrections. As with the `anova_test` function we used previously, the data need to be reshaped into long format.

```
> #create id numbers
> d$id <- 1:nrow(d)
>
> d$one <- 1
> d$one <- factor(d$one)
>
>
>
> #reshape data going from wide to long format
> library(reshape2)
> longdata <- melt(d,
+                  measure.vars = c("vocab1", "vocab2", "vocab3"),  #old variables
+                  variable.name = "level",       #name new variable for the value labels
+                  value.name = "vocab")         #name a new variable for the values
>
> library(tibble)
> library(car)

> #need to create a factor and frame for appropriate number of IV levels, called "condition" here
> condition <- c(1,2,3)
> condition <- as.factor(condition)
> condframe <- data.frame(condition)
> #use lm to run model and Manova function (Anova also works) to get univariate with corrections and
manova results
> model2 <- lm(cbind(vocab1, vocab2, vocab3) ~ 1, data=d)
> analysis <- Manova(model2, idata=condframe, idesign=~condition, type="III")
> summary(analysis)

Type III Repeated Measures MANOVA Tests:

------------------------------------------

Term: (Intercept)

 Response transformation matrix:
       (Intercept)
vocab1           1
vocab2           1
vocab3           1

Sum of squares and products for the hypothesis:
             (Intercept)
(Intercept)      121203

Multivariate Tests: (Intercept)
                 Df test stat approx F num Df den Df        Pr(>F)
Pillai            1  0.919612 125.8361      1     11 0.00000023188
Wilks             1  0.080388 125.8361      1     11 0.00000023188
Hotelling-Lawley  1 11.439641 125.8361      1     11 0.00000023188
Roy               1 11.439641 125.8361      1     11 0.00000023188

------------------------------------------

Term: condition

 Response transformation matrix:
       condition1 condition2
vocab1          1          0
vocab2          0          1
vocab3         -1         -1

Sum of squares and products for the hypothesis:
           condition1 condition2
condition1        363       -561
condition2       -561        867

Multivariate Tests: condition
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1  0.790206 18.83291      2     10 0.00040641
Wilks             1  0.209794 18.83291      2     10 0.00040641
Hotelling-Lawley  1  3.766582 18.83291      2     10 0.00040641
Roy               1  3.766582 18.83291      2     10 0.00040641
```

```
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

            Sum Sq num Df Error SS den Df F value       Pr(>F)
(Intercept)  40401      1   3531.7     11 125.836 0.0000002319
condition     1194      2   1067.3     22  12.305     0.000259

Mauchly Tests for Sphericity

          Test statistic  p-value
condition        0.41524 0.012345


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

           GG eps Pr(>F[GG])
condition 0.63101    0.00225

             HF eps Pr(>F[HF])
condition 0.6748954 0.00173583
```

Another way to get univariate repeated measures ANOVA with sphericity corrections is with the exANOVA procedure from the ez package

```
> #ezANOVA is alternative--gives sphericity tests but no MSE
> library('ez')

> mymodel = ezANOVA(data = longdata,
+       dv = vocab, #dependent variable
+       wid = .(id), #id variable
+       within = .(level), #levels of the independent variable
+       detailed = TRUE)  #print some extra details
```

Write-up Example

A within-subjects ANOVA was used to compare the vocabulary scores in the three lecture conditions. The average vocabulary score was the highest in the first topic (M_1 = 40.00, SD = 10.80), followed by the third topic, M_3 = 34.50 and SD = 8.39, and then second topic (M_2 = 26.00, SD = 26.00) *[means not shown in the above output excerpts to save space]*. The results indicated that there was a significant difference among the three conditions, F(2,22) = 12.31, p < .01, for all sphericity assumption correction tests.

The sample size was only 12 in my example, so I would not use the MANOVA repeated-measures test in this case. Had there been a larger sample size, which is likely almost always the case in practice with applied research, I would have stated something like: "The multivariate F-value was examined, because the multivariate analysis of variance approach to repeated measures is not subject to the sphericity assumption and performs better than sphericity-corrected F-tests if the sample size is sufficient (Algina & Kesselman, 1997)."

References

Algina, J., & Kesselman, H.J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods, 2*, 208-218.

Edwards, A. L. (1985). *Experimental design in psychological research (5th ed.*). New York: Harper & Row.

Kesselman, H.J., Rogan, J.C., Mendoza, J.L., & Breen, L.J. (1980). Testing the validity conditions of repeated measures F tests. *Psychological Bulletin, 87*, 479-481.

Šimeček, P., & Šimečková, M. (2013). Modification of Tukey's additivity test. *Journal of Statistical Planning and Inference*, *143*(1), 197-201.

Tukey, John (1949). One degree of freedom for non-additivity. *Biometrics, 5* (3): 232–242