

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

The Sphericity Assumption

The sphericity assumption states that the variance of the difference scores in a within-subjects design (the s_d^2 in a paired t-test) are equal across all the groups. Imagine a within-subjects design with three levels of the independent variable (e.g., pretest, posttest, and follow-up), 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 assumes that the variances of each of these sets of difference scores are not statistically different from one another. The sphericity assumption is similar to the homogeneity of variance assumption with between subjects ANOVA. 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. One could say that the F-test is *positively biased* under these conditions. There are several different approaches to correcting for this bias.

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 $df_{A \times S} = s - 1$ are used instead of the usual $df_A = a - 1$ and $df_{A \times S} = (a - 1)(s - 1)$. 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 cells.

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 $df_{A \times S}$ and df_A are adjusted by this factor, 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.

Mauchly's test. Mauchly's test is a test of the sphericity assumption using the chi-square test. Unfortunately, this test is not a very useful one. For small sample sizes, it tends to have too many Type II errors (i.e., it misses sphericity violations), and for large sample sizes, it tends to be significant even though the violation is small in magnitude (Type I error; e.g., Kesselman, Rogan, Mendoza, & Breen, 1980). Thus, the Mauchly's test may not be much help and may be misleading when deciding if there is a violation of the sphericity assumption.

Compound Symmetry

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

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

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 are calculated among all those difference scores. Under the compound symmetry assumption these correlations (or covariances, actually) must not be different. A violation of this stricter compound symmetry assumption does not necessarily mean the sphericity assumption will be violated. SPSS does not currently provide a test of this assumption.

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 thus 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. This is the so-called “additivity” assumption that there is no interaction between A and S that is not error (interactions are multiplicative or “nonadditive”). Because nonadditivity (a violation of this assumption) implies heterogeneous variances for the difference scores, the sphericity assumption will be violated if nonadditivity occurs. A test exists for this assumption, called the Tukey test for nonadditivity, but it is not currently available in SPSS.

Comments and Recommendations

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). The sphericity assumption, for instance, does not stipulate that the variances of the scores for each level are the same, but that difference scores are calculated for pairs of levels (e.g., Time 2 scores minus Time 1 scores vs. Time 3 scores minus Time 2 scores) because these assumptions involve the differences between levels being equal. 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.

Page 359 of the Myers and Well text makes a couple of important points regarding these corrections (based on recommendations of Greenhouse & Geisser): 1) If F is nonsignificant, 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, I would recommend the Huynh and Feldt correction, because its correction is the least severe. However, it might be best to report the results from all of them in this case. 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.

There are two other possible solutions: 1) transform the dependent variable scores using a square root transformation or raise the score to a fractional power (e.g., .33), or 2) use a multivariate analysis of variance (MANOVA) test. For the transformation approach, you may have to try different transformations until the sphericity problem is resolved. The

transformation approach may be quite helpful in resolving the problem, but the researcher will have more difficulty interpreting the results. Under some conditions, the MANOVA approach can be more powerful than the within-subjects ANOVA, and the MANOVA test does not assume sphericity. The MANOVA test is printed on the GLM repeated measures output in SPSS. Algina and Kesselman (1997) suggest that for few levels and large sample sizes, the MANOVA approach may be more powerful. Their guidelines are to use MANOVA if a) 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 b) 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 ($a + 30$).

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. 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-Geisser, 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).

Mauchly's Test of Sphericity^b

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon ^a		
					Greenhouse e-Geisser	Huynh-Feldt	Lower-bound
TOPIC	.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. 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.
- b. Design: Intercept
 Within Subjects Design: TOPIC

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^a
TOPIC	Sphericity Assumed	1194.000	2	597.000	12.305	.000	.528	24.611	.990
	Greenhouse-Geisser	1194.000	1.262	946.103	12.305	.002	.528	15.530	.940
	Huynh-Feldt	1194.000	1.350	884.582	12.305	.002	.528	16.610	.951
	Lower-bound	1194.000	1.000	1194.000	12.305	.005	.528	12.305	.890
Error(TOPIC)	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					

a. Computed using alpha = .05

References

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

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.