Reporting Statistics in Psychology

This document contains general guidelines for the reporting of statistics in psychology research. The details of statistical reporting vary slightly among different areas of science and also among different journals.

General Guidelines

Rounding Numbers

For numbers greater than 100, report to the nearest whole number (e.g., $M = 6254$). For numbers between 10 and 100, report to one decimal place (e.g., $M = 23.4$). For numbers between 0.10 and 10, report to two decimal places (e.g., $M = 4.34$, $SD = 0.93$). For numbers less than 0.10, report to three decimal places, or however many digits you need to have a non-zero number (e.g., $M = 0.014$, $SEM = 0.0004$).

For numbers...	Round to...	SPSS	Report
Greater than 100	Whole number	1034.963	1035
10 - 100	1 decimal place	11.4378	11.4
0.10 - 10	2 decimal places	4.3682	4.37
0.001 - 0.10	3 decimal places	0.0352	0.035
Less than 0.001	As many digits as needed for non-zero	0.00038	0.0004

Do not report any decimal places if you are reporting something that can only be a whole number. For example, the number of participants in a study should be reported as $N = 5$, not $N = 5.0$.

Report exact p-values (not $p < .05$), even for non-significant results. Round as above, unless SPSS gives a p-value of .000; then report $p < .001$. Two-tailed p-values are assumed. If you are reporting a one-tailed p-value, you must say so.

Omit the leading zero from p-values, correlation coefficients (r), partial eta-squared (η_p^2), and other numbers that cannot ever be greater than 1.0 (e.g., $p = .043$, not $p = 0.043$).

Statistical Abbreviations

Abbreviations using Latin letters, such as mean (M) and standard deviation (SD), should be italicised, while abbreviations using Greek letters, such as partial eta-squared (η_p^2), should not be italicised and can be written out in full if you cannot use Greek letters. There should be a space before and after equal signs. The abbreviations should only be used inside of parentheses; spell out the names otherwise.

Inferential statistics should generally be reported in the style of:
"*statistic*(degrees of freedom) = value, p = value, *effect size statistic* = value"

Statistic	Example
Mean and standard deviation	$M = 3.45$, $SD = 1.21$
Mann-Whitney	$U = 67.5$, $p = .034$, $r = .38$
Wilcoxon signed-ranks	$Z = 4.21$, $p < .001$
Sign test	$Z = 3.47$, $p = .001$
t-test	$t(19) = 2.45$, $p = .031$, $d = 0.54$
ANOVA	$F(2, 1279) = 6.15$, $p = .002$, $\eta_p^2 = 0.010$
Pearson's correlation	$r(1282) = .13$, $p < .001$

Descriptive Statistics

Means and standard deviations should be given either in the text or in a table, but not both.

Descriptive Statistics

	N	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
age	2351	25.480	.1638	7.9445	1.869	.050	3.930	.101
Valid N (listwise)	2351							

> The average age of participants was 25.5 years (*SD* = 7.94).

> The age of participants ranged from 18 to 70 years (*M* = 25.5, *SD* = 7.94). Age was non-normally distributed, with skewness of 1.87 (*SE* = 0.05) and kurtosis of 3.93 (*SE* = 0.10)

> Participants were 98 men and 132 women aged 17 to 25 years (men: *M* = 19.2, *SD* = 2.32; women: *M* = 19.6, *SD* = 2.54).

Non-parametric tests

Do not report means and standard deviations for non-parametric tests. Report the median and range in the text or in a table. The statistics *U* and *Z* should be capitalised and italicised. A measure of effect size, *r*, can be calculated by dividing *Z* by the square root of *N* ($r = Z / \sqrt{N}$).

Mann-Whitney Test (2 Independent Samples...)

Ranks

	pill	N	Mean Rank	Sum of Ranks
sra	0	17	19.03	323.50
	1	14	12.32	172.50
	Total	31		

Test Statistics[b]

	sra
Mann-Whitney U	67.500
Wilcoxon W	172.500
Z	-2.119
Asymp. Sig. (2-tailed)	.034
Exact Sig. [2*(1-tailed...	.040[a]

a. Not corrected for ties.

b. Grouping Variable: pill

> A Mann-Whitney test indicated that self-rated attractiveness was greater for women who were not using oral contraceptives (*Mdn* = 5) than for women who were using oral contraceptives (*Mdn* = 4), *U* = 67.5, *p* = .034, *r* = .38.

Wilcoxon Signed-ranks Test (2 Related Samples...)

Ranks

		N	Mean Rank	Sum of Ranks
male - female	Negative Ranks	25[a]	17.48	437.00
	Positive Ranks	5[b]	5.60	28.00
	Ties	1[c]		
	Total	31		

a. male < female

b. male > female

c. male = female

Test Statistics[b]

	male - female
Z	-4.207[a]
Asymp. Sig. (2-tailed)	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

> A Wilcoxon Signed-ranks test indicated that femininity was preferred more in female faces (*Mdn* = 0.85) than in male faces (*Mdn* = 0.65), *Z* = 4.21, *p* < .001, *r* = .76.

Sign Test (2 Related Samples...)

Frequencies

		N
male - female	Negative Differences[a]	25
	Positive Differences[b]	5
	Ties[c]	1
	Total	31

a. male < female
b. male > female
c. male = female

Test Statistics[a]

	male - female
Z	-3.469
Asymp. Sig. (2-tailed)	.001

a. Sign Test

> A sign test indicated that femininity was preferred more in female faces than in male faces, $Z = 3.47$, $p = .001$.

T-tests

Report degrees of freedom in parentheses. The statistics *t*, *p* and Cohen's *d* should be reported and italicised.

One-sample t-test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
female	31	4.503	.6957	.1250
male	31	3.4581	.73179	.13143

One-Sample Test

	Test Value = 3.5					
					95% CI	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
female	8.029	30	.000	1.0032	.748	1.258
male	-.319	30	.752	-.04194	-.3104	.2265

> One-sample t-test indicated that femininity preferences were greater than the chance level of 3.5 for female faces ($M = 4.50$, $SD = 0.70$), $t(30) = 8.01$, $p < .001$, $d = 1.44$, but not for male faces ($M = 3.46$, $SD = 0.73$), $t(30) = -0.32$, $p = .75$, $d = 0.057$.

> The number of masculine faces chosen out of 20 possible was compared to the chance value of 10 using a one-sample t-test. Masculine faces were chosen more often than chance, $t(76) = 4.35$, $p = .004$, $d = 0.35$.

Paired-samples t-test

Report paired-samples t-tests in the same way as one-sample t-tests.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	pathogen	26.39	722	7.414	.276
	sexual	18.03	722	9.490	.353

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	pathogen & sexual	722	.373	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	pathogen - sexual	8.353	9.617	.358	7.650	9.056	23.338	721	.000

> A paired-samples t-test indicated that scores were significantly higher for the pathogen subscale ($M = 26.4$, $SD = 7.41$) than for the sexual subscale ($M = 18.0$, $SD = 9.49$), $t(721) = 23.3$, $p < .001$, $d = 0.87$.

> Scores on the pathogen subscale (M = 26.4, SD = 7.41) were higher than scores on the sexual subscale (M = 18.0, SD = 9.49), $t(721) = 23.3$, $p < .001$, $d = 0.87$. A one-tailed p-value is reported due to the strong prediction of this effect.

Independent-samples t-test

Group Statistics

	sex	N	Mean	Std. Deviation	Std. Error Mean
pathogen	male	201	24.42	7.689	.542
	female	535	27.04	7.209	.312

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
pathogen	Equal variances assumed	2.568	.109	-4.301	734	.000	-2.613	.607	-3.805	-1.420
	Equal variances not assumed			-4.177	340.008	.000	-2.613	.626	-3.843	-1.382

> An independent-samples t-test indicated that scores were significantly higher for women (M = 27.0, SD = 7.21) than for men (M = 24.2, SD = 7.69), $t(734) = 4.30$, $p < .001$, $d = 0.35$.

If Levene's test for equality of variances is significant, report the statistics for the row equal variances not assumed with the altered degrees of freedom rounded to the nearest whole number.

> Scores on the pathogen subscale were higher for women (M = 27.0, SD = 7.21) than for men (M = 24.2, SD = 7.69), $t(340) = 4.30$, $p < .001$, $d = 0.35$. Levene's test indicated unequal variances ($F = 3.56$, $p = .043$), so degrees of freedom were adjusted from 734 to 340.

ANOVAs

ANOVAs have two degrees of freedom to report. Report the between-groups df first and the within-groups df second, separated by a comma and a space (e.g., $F(1, 237) = 3.45$). The measure of effect size, partial eta-squared (η_p^2), may be written out or abbreviated, omits the leading zero and is not italicised.

One-way ANOVAs and Post-hocs

Tests of Between-Subjects Effects

Dependent Variable:female

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4.766[a]	2	2.383	6.152	.002	.010
Intercept	25473.878	1	25473.878	65762.819	.000	.981
sra3	4.766	2	2.383	6.152	.002	.010
Error	495.433	1279	.387			
Total	26234.842	1282				
Corrected Total	500.199	1281				

a. R Squared = .010 (Adjusted R Squared = .008)

Multiple Comparisons

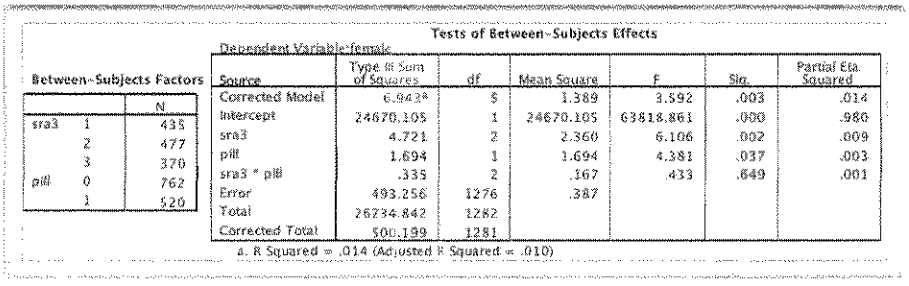
female
Tukey HSD

(I) sra3	(J) sra3	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	2	-.116*	.0413	.014	-.212	-.019
	3	-.141*	.0448	.004	-.245	-.038
2	1	.116*	.0413	.014	.019	.212
	3	-.026	.0431	.821	-.127	.075
3	1	.141*	.0448	.004	.038	.245
	2	.026	.0431	.821	-.075	.127

Based on observed means.
The error term is Mean Square(Error) = .387.
*. The mean difference is significant at the .05 level.

> Analysis of variance showed a main effect of self-rated attractiveness (SRA) on preferences for femininity in female faces, $F(2, 1279) = 6.15$, $p = .002$, $\eta_p^2 = .010$. Post-hoc analyses using Tukey's HSD indicated that femininity preferences were lower for participants with low SRA than for participants with average SRA ($p = .014$) and high SRA ($p = .004$), but femininity preferences did not differ significantly between participants with average and high SRA ($p = .82$).

2-way Factorial ANOVAs



Between-Subjects Factors

		N
sra3	1	435
	2	477
	3	370
pill	0	762
	1	520

Tests of Between-Subjects Effects

Dependent Variable:female

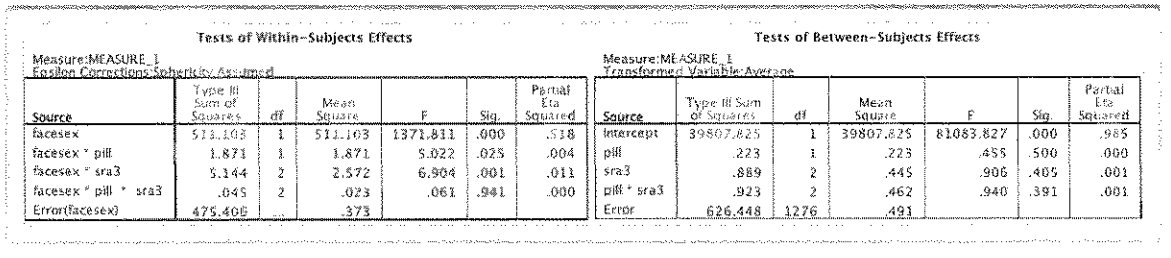
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	6.943[a]	5	1.389	3.592	.003	.014
Intercept	24670.105	1	24670.105	63818.861	.000	.980
sra3	4.721	2	2.360	6.106	.002	.009
pill	1.694	1	1.694	4.381	.037	.003
sra3 * pill	.335	2	.167	.433	.649	.001
Error	493.256	1276	.387			
Total	26234.842	1282				
Corrected Total	500.199	1281				

a. R Squared = .014 (Adjusted R Squared = .010)

> A 3x2 ANOVA with self-rated attractiveness (low, average, high) and oral contraceptive use (true, false) as between-subjects factors revealed a main effects of SRA, $F(2, 1276) = 6.11$, $p = .002$, $\eta_p^2 = .009$, and oral contraceptive use, $F(1, 1276) = 4.38$, $p = .037$, $\eta_p^2 = 0.003$. These main effects were not qualified by an interaction between SRA and oral contraceptive use, $F(2, 1276) = 0.43$, $p = .65$, $\eta_p^2 = .001$.

3-way ANOVAs and Higher

Although some textbooks suggest that you report all main effects and interactions, even if not significant, this reduces the understandability of the results of a complex design (i.e. 3-way or higher). Report all significant effects and all predicted effects, even if not significant. If there are more than two non-significant effects that are irrelevant to your main hypotheses (e.g. you predicted an interaction among three factors, but did not predict any main effects or 2-way interactions), you can summarise them as in the example below.



Tests of Within-Subjects Effects

Measure:MEASURE_1

Epsilon Corrections: Sphericity Assumed

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
facesex	513.103	1	513.103	1371.811	.000	.518
facesex * pill	1.871	1	1.871	5.022	.025	.004
facesex * sra3	5.144	2	2.572	6.904	.001	.011
facesex * pill * sra3	.045	2	.023	.061	.941	.000
Error(facesex)	475.406	1276	.373			

Tests of Between-Subjects Effects

Measure:MEASURE_1

Transformed Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	39807.825	1	39807.825	81083.827	.000	.985
pill	.223	1	.223	.455	.500	.000
sra3	.889	2	.445	.906	.405	.001
pill * sra3	.923	2	.462	.940	.391	.001
Error	626.448	1276	.491			

> A mixed-design ANOVA with sex of face (male, female) as a within-subjects factor and self-rated attractiveness (low, average, high) and oral contraceptive use (true, false) as between-subjects factors revealed a main effect of sex of face, $F(1, 1276) = 1372$, $p < .001$, $\eta_p^2 = .52$. This was qualified by interactions between sex of face and SRA, $F(2, 1276) = 6.90$, $p = .001$, $\eta_p^2 = .011$, and between sex of face and oral contraceptive use, $F(1, 1276) = 5.02$, $p = .025$, $\eta_p^2 = .004$. The predicted interaction among sex of face, SRA and oral contraceptive use was not significant, $F(2, 1276) = 0.06$, $p = .94$, $\eta_p^2 < .001$. All other main effects and interactions were non-significant and irrelevant to our hypotheses, all $F \leq 0.94$, $p \geq .39$, $\eta_p^2 \leq .001$.

Violations of Sphericity and Greenhouse-Geisser Corrections

ANOVAs are not robust to violations of sphericity, but can be easily corrected. For each within-subjects factor with more than two levels, check if Mauchly's test is significant. If so, report chi-squared (χ^2), degrees of freedom, p and epsilon (ε) as below and report the Greenhouse-Geisser corrected values for any effects involving this factor (rounded to the appropriate decimal place). SPSS will report a chi-squared of .000 and no p-value for within-subjects factors with only two levels; corrections are not needed.

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a] Greenhouse-Geisser	Huynh-Feldt	Lower-bound
subscale	.990	16.044	2	.000	.983	.986	.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 + sex
Within Subjects Design: subscale

Tests of Within-Subjects Effects

Measure:MEASURE_1
Epsilon Corrections: Greenhouse-Geisser

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
subscale	41841.140	1.905	21961.774	377.538	.000	.347
subscale * sex	3368.090	1.905	1767.859	30.391	.000	.041
Error(subscale)	78575.822	1350.773	58.171			

Tests of Between-Subjects Effects

Measure:MEASURE_1
Transformed Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	305115.373	1	305115.373	8914.519	.000	.926
sex	2695.575	1	2695.575	78.756	.000	.100
Error	24266.793	709	34.227			

> Data were analysed using a mixed-design ANOVA with a within-subjects factor of subscale (pathogen, sexual, moral) and a between-subject factor of sex (male, female). Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(2) = 16.8$, $p < .001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.98$). Main effects of subscale, $F(1.91, 1350.8) = 378$, $p < .001$, $\eta_p^2 = .35$, and sex, $F(1, 709) = 78.8$, $p < .001$, $\eta_p^2 = .10$, were qualified by an interaction between subscale and sex, $F(1.91, 1351) = 30.4$, $p < .001$, $\eta_p^2 = .041$.

ANCOVA

Tests of Between-Subjects Effects

Dependent Variable:pathogen

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1210.158[a]	3	403.386	7.502	.000	.030
Intercept	52794.932	1	52794.932	981.794	.000	.573
sex	107.679	1	107.679	2.002	.157	.003
age	174.602	1	174.602	3.247	.072	.004
sex * age	.879	1	.879	.016	.898	.000
Error	39362.526	732	53.774			
Total	550509.000	736				
Corrected Total	40572.683	735				

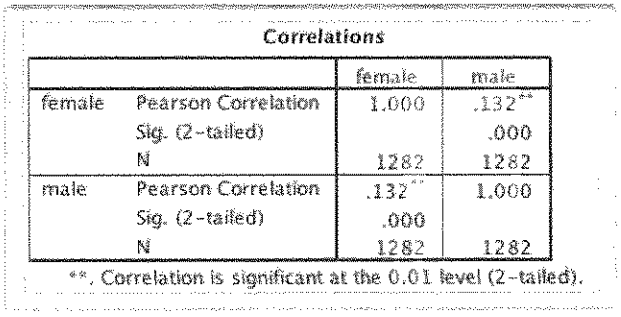
a. R Squared = .030 (Adjusted R Squared = .026)

> An ANCOVA [between-subjects factor: sex (male, female); covariate: age] revealed no main effects of sex, $F(1, 732) = 2.00$, $p = .16$, $\eta_p^2 = .003$, or age, $F(1, 732) = 3.25$, $p = .072$, $\eta_p^2 = .004$, and no interaction between sex and age, $F(1, 732) = 0.016$, $p = .90$, $\eta_p^2 < .001$.

> The predicted main effect of sex was not significant, $F(1, 732) = 2.00$, $p = .16$, $\eta_p^2 = .003$, nor was the predicted main effect of age, $F(1, 732) = 3.25$, $p = .072$, $\eta_p^2 = .004$. The interaction between sex and age were also not significant, $F(1, 732) = 0.016$, $p = .90$, $\eta_p^2 < .001$.

Correlations

Italicise *r* and *p*. Omit the leading zero from *r*.



Correlations

		female	male
female	Pearson Correlation	1.000	.132**
	Sig. (2-tailed)		.000
	N	1282	1282
male	Pearson Correlation	.132**	1.000
	Sig. (2-tailed)	.000	
	N	1282	1282

**. Correlation is significant at the 0.01 level (2-tailed).

> Preferences for femininity in male and female faces were positively correlated, Pearson's $r(1282) = .13$, $p < .001$.

References

American Psychological Association. (2005). *Concise Rules of APA Style*. Washington, DC: APA Publications.

Field, A. P., & Hole, G. J. (2003). *How to design and report experiments*. London: Sage Publications.