Parents, teachers, and media: agents of biased socialization

Kroes, A.D.A.

Chapter 2

Demystifying Omega Squared

Practical Guidance for Effect Size in Common ANOVA Designs

Antoinette D. A. Kroes, Jason R. Finley

Advanced online publication in *Psychological Methods*, 2023, https://doi.org/10.1037/met0000581

Abstract

Omega squared ($\widehat{\omega}^2$) is a measure of effect size for ANOVA designs. It is less biased than eta squared, but reported less often. This is in part due to lack of clear guidance on how to calculate it. In this paper, we discuss the logic behind effect size measures, the problem with eta squared, the history of omega squared, and why it has been underused. We then provide a user-friendly guide to omega squared and partial omega squared for ANOVA designs with fixed factors, including one-way, two-way, and three-way designs, using within-subjects factors and/or between-subjects factors. We show how to calculate omega squared using output from SPSS. We provide information on the calculation of confidence intervals. We examine the problems of non-additivity, and intrinsic versus extrinsic factors. We argue that statistical package developers could play an important role in making the calculation of omega squared easier. Finally, we recommend that researchers report the formulas used in calculating effect sizes, include confidence intervals if possible, and include ANOVA tables in the supplemental materials of their work.

Keywords: Omega squared; partial omega squared; effect size; ANOVA

Comparison of means is a common analysis in psychological science. For example, researchers often want to compare performance (dependent variable) across different groups of people or different conditions of a treatment (independent variable). ANOVA (analysis of variance) is the inferential statistical test for comparing means across three or more groups/conditions, and/or for comparing means across two or more independent variables. The *F* test statistic and *p*-value from an ANOVA indicate the statistical significance of the test—that is, the probability of obtaining these differences in means just due to chance. However, it is crucially important for researchers to also report a measure of *effect size*, which tells us not just whether there was likely a difference in means at all, but how *large* that difference was, or how *strongly* an independent variable effected the dependent variable (American Psychological Association, 2020; Keppel & Wickens, 2004; Thompson, 1999a; L. Wilkinson, 1999).

Effect sizes are also important for meta-analysis, in which treatment effects are compared across studies (Thompson, 1999a), and for calculating sample size required to obtain certain levels of power (Keppel & Wickens, 2004).

A commonly used measure of effect size for ANOVA is eta squared (η^2) or partial eta squared (η_p^2). Eta squared estimates the amount of variance in the dependent variable that is accounted for by one or more independent variables. However, eta squared is problematic because it is biased: it tends to overestimate the true effect size in a population. Many authors have pointed out this flaw and have advised alternatives (e.g., Albers & Lakens, 2018; Field, 2017; Lakens, 2015; Okada, 2013; Olejnik & Algina, 2000; Tabachnick & Fidell, 2007a; Yigit & Mendes, 2018). The foremost alternative measures of effect size are omega squared ($\widehat{\omega}^2$) and partial omega squared ($\widehat{\omega}_p^2$), which were first proposed over 50 years ago (Hays, 1963). Omega squared is much less biased than eta squared, and thus is a superior measure. However, it is still rarely used (Alhija & Levy,

2009; Zhou & Skidmore, 2017). The goal of this paper is to help remedy that by providing user-friendly explanations and instructions.[1]

In this paper, we will explain the logic of eta and omega squared, and the shortcomings of eta squared. We will give a brief history of omega squared, and we will discuss the reasons we believe it is still underused. Most helpfully, we will provide formulas to calculate omega squared for the most commonly used ANOVA designs in the behavioral and social sciences, up to three-way ANOVAs. We will explain how to use these formulas with output from SPSS. We will conclude with recommendations on how to calculate and report omega squared.

Clarifying Our Scope and Terminology

Let us first clarify some terminology, and define the scope of what we will and will not cover.

Variables (aka factors): are characteristics that vary across entities. In experiments, an *independent variable* is one that is manipulated by the researcher (i.e., they determine the possible values and assign participants to those values), and a *dependent variable* is one that is simply measured by the researcher. Common inferential statistical tests such as *t*-test and ANOVA are used to see if an independent variable(s) has an effect on the dependent variable. ANOVA may also be used for non-experimental data, in which case the variables may be referred to as *predictor variable* and *outcome variable*, rather than independent variable and dependent variable, respectively. For convenience, we will simply use the terms independent variable (IV) and dependent variable (DV). We will

[1] Seasoned researchers may note how painstakingly we describe and explain basic concepts in this paper. We have deliberately chosen to do this, as we found that most of the previous articles on omega squared assume a good deal of specific prior knowledge, and are difficult to decipher for non-statisticians. That may be one of the obstacles to more widespread use of omega squared. Another obstacle is that a number of the sources we have scoured are out-of-print textbooks. We wish to remove such obstacles.

also use the term *treatment* as a synonym for IV. As for labeling, the outcome or dependent variable is typically labeled as Y. The predictor or independent variable may often be labeled as X, though we will use the label A instead of X in order to be consistent with many of our sources. The words variable and factor are synonyms, and the latter is often used to refer to IVs in ANOVA; we will use the words variable and factor interchangeably in this paper.

ANOVA (analysis of variance): A statistical test (aka model) of the relationship between one or more categorical independent variables and one continuous dependent variable. A categorical variable has values that are treated as named categories without inherent order or numeric value, and can have two or more such categories; a continuous variable has numeric values and may be of interval or ratio scale of measurement.[2] The purpose of ANOVA is to compare means across three or more groups/conditions. The variances of the groups/conditions are simply a tool toward that end, and are not actually themselves the subject of analysis. An ANOVA produces an ANOVA summary table, which contains the outcome(s) of the test(s) for statistical significance, as well as the components needed for calculating effect size. The terms one-way, two-way, and three-way refer to the number of independent variables in the design.

Statistical significance*:* When conducting statistical analyses, we choose a significance level, represented by α (alpha). This value represents the probability of rejecting the null hypothesis while it is in fact true. Traditionally, this value is set at .05 or .01. When carrying out statistical analyses we calculate a *p*-value. This value is the probability of getting the obtained result (or a more extreme value) while the null hypothesis is true. When the *p*-value is equal to or lower than the

[2] It is also possible to use ANOVA with a Likert-scale DV (Norman, 2010) or a dichotomous DV (Lunney, 1970).

significance level α, we reject the null hypothesis. When the *p*-value exceeds α, it cannot be ruled out that the found effect is due to chance (Kirk, 1996). Reporting the *p*-value is a longstanding and important tradition in the social and behavioral sciences (American Psychological Association, 2020).

Effect size: is a measure that estimates the strength of the investigated effects of the IV(s). Whereas statistical significance only indicates whether an effect is present, effect sizes describe the quantitative size of the effect (Fritz et al., 2012). Effect sizes help us understand the expected impact of a treatment or condition. Thus, an effect size gives an indication whether an effect is meaningful in the real world, and is therefore called *practical significance* (Ellis, 2010; Kirk, 1996). While statistical significance is dependent on sample size, effect sizes are not; they should be comparable across studies, regardless of sample size (Levine & Hullett, 2002). A large effect that is not significant indicates that greater power may be needed, while a very small significant effect cautions against overvaluing the effect (Fritz et al., 2012).

Between-subjects, within-subjects, and split-plot designs: These distinctions have to do with how an independent variable(s) is manipulated. For an IV that is manipulated *between-subjects* (aka independent samples), each participant experiences only one value of the IV, and thus the comparison of DV means is made between different groups of participants. For an IV that is manipulated *within-subjects* (aka repeated measures), each participant experiences all values of the IV, and thus the comparison of DV means is made within that one group of participants such that each participant's score in one condition is compared to their scores in the other conditions. By convention, the values of a between-subjects IV are often called *groups*, and the values of a within-subjects IV are often called *conditions*. The term *levels* may also be used in either design. In this paper, we will use the term groups/conditions to refer generally to values of an IV. Note also that the terms *participant* and *subject* are synonyms.

A one-way ANOVA has only one IV and will be either between-subjects or within-subjects. For two-way ANOVAs and higher, a *split-plot* design is one in which there is at least one between-subjects IV and at least one within-subjects IV. The name originates from agricultural research, where experiments were conducted on different plots and subplots of land (Goos, 2010). Some authors have referred to this model as a mixed design, as it is a mix of within-subjects and between-subjects factors (Gaebelein & Soderquist, 1978; Keppel & Wickens, 2004). However, this can be quite confusing, as the term "mixed" has different meanings in other contexts, such as a "mixed effects model" (aka mixed model, linear mixed model, multilevel model, ANOVA Model III) which is one that includes both fixed factors and random factors, or "mixed methods research" (aka mixed research) which combines qualitative and quantitative methods. Thus, in this paper we use the term split-plot to avoid confusion. Note that the distinction of between-subjects versus within-subjects is unrelated to the distinction of fixed effects versus random effects, which we will discuss next.[3]

Fixed effects versus random effects (aka fixed factors vs. random factors): This distinction has to do with how the conditions of an independent variable are chosen. A fixed effect is when the researcher chooses a fixed set of conditions for an IV. A random effect is when the researcher randomly samples conditions from a range of possible values, so that the conditions used may vary across experiments.[4]

The distinction is important for several reasons. First, the conclusions from an experiment using fixed effects should be limited to just those conditions that were used, whereas the conclusions based on random effects can be broader. Second, the expected mean squares, and thus the

[3] The encyclopedia entry Mixed Model Design (Kraska, 2010) confuses these two uses of the word "mixed."

[4] More explanation of fixed versus random effects can be found in the following textbooks: Keppel and Wickens (2004, pp. 533-549), and Myers, Well, and Lorch (2010, p. 335).

appropriate error terms, differ for fixed versus random effects. This is why there are three different overall ANOVA models. ANOVA Model I is for designs including only fixed effects, ANOVA Model II is for designs including only random effects, and ANOVA Model III is for designs including both fixed and random effects. The way that omega squared is calculated differs across these three models. The most common scenario in psychology research is an experiment using only fixed effects (i.e., ANOVA Model I). Thus, we will limit the scope of this paper to only fixed effect IVs (aka fixed factors). If you want to know about calculating omega squared for designs that include random effects consult Dodd and Schultz (1973), Olejnik and Algina (2000), and Vaughan and Corballis (1969). Alternatively, the intraclass correlation coefficient (ICC, or $\hat{\rho}^2$) has been recommended for designs that include random effects (Kirk, 2012; Maxwell et al., 1981).

Summary of scope: In this paper, we will address between-subjects, within-subjects, and split-plot designs, up to three-way ANOVAs, for fixed factors only.

Logic of ANOVA Effect Size Measures

In order to understand the logic of effect size measures for ANOVA, including omega squared, we must first consider the larger context. When we measure a group of people's performance on some task, their scores will vary from each other. The job of psychological science is to understand *why* that variance happens, to "account for" the variance.[5]

Let us use Y to represent a variable we have measured, for example performance on a memory test for a list of words. In the context of an experiment, we call this the dependent variable. If we have no other information about the Y scores—that is, we know nothing about any other

[5] For a further discussion of what "accounting for variance" means, see Sechrest and Yeaton (1982).

variables—then all we can do is *describe* the variance of those scores. We cannot explain any of it. Every possible conceivable other variable in the world could be influencing the spread of those Y scores.

But suppose that we do have more information about those Y scores. Suppose we know the value of another variable, let us call it A, that goes along with each Y score. In fact, the reason we know the A values in this example is because we randomly assigned people to three different conditions of A (e.g., short, medium, and long amount of time to study the list) before we measured Y. That is, we ran a between-subjects experiment and A was the independent variable.

An inferential statistical test, in this case a one-way between-subjects ANOVA, would tell us the extent to which we might want to believe that there is truly any effect at all of variable A on variable Y. If the ANOVA tells us that $p < .05$, the effect of A on Y is statistically significant. But the statistical significance does not tell us how *big* the effect of A on Y is (i.e., the *effect size*). For that we must see *how much of the variance* in Y may be attributable to A. That is, how much of a role did study time (A) play in the variance of peoples' memory test performance (Y)?

To do so, we must *partition the total variance into two different components*. The first component is the variance in Y attributable to A, which might be called between-groups variance, or treatment effect. The second component is all the remaining leftover variance in Y, which could be due to any other variables in the world (aka extraneous variables). This component is often called error variance. The word error does not mean mistake. It means the unexplained deviation of a Y score from what we would have predicted it to be, based on the overall mean of Y and the mean of the relevant A condition. Error variance is simply variance in Y that is due to variables that are not included in our statistical model.

We quantify variance components by calculating several different *sums of squared deviations*. SS_{total} is the sum of squared deviations of all Y scores from the grand mean (i.e., the overall mean of all Y scores,

disregarding any other variable). This represents the total variability of all the Y scores:

$$SS_{total} = \sum_j \sum_i \left(Y_{ij} - M\right)^2 \quad (1)$$

In the above formula, Σ means summation, *j* indicates a particular group/condition of the independent variable (A), *i* indicates a particular participant, Y_{ij} is the dependent variable score of participant *i* in group/condition *j*, and *M* is the grand mean.

SS_{total} can be partitioned into two components: $SS_{between}$ and SS_{within}. $SS_{between}$ (also called SS_{effect} or $SS_{treatment}$) is the between-groups variance component, which tells us how much of the variance in Y is due to A. It is quantified by calculating how much the group means (e.g., mean test performance for the short, medium, and long study time groups) vary around the grand mean:

$$SS_{between} = \sum_j n_j \left(M_j - M\right)^2 \quad (2)$$

In the above formula, n_j is the number of participants in group/condition *j*, and *Mj* is the mean of group/condition *j*.

SS_{within} (also called SS_{error}) is the leftover variability, which is our best estimate of the influence of all other conceivable variables on Y. It is quantified by the sum of squared deviations of individual Y scores from their respective group means:

$$SS_{within} = \sum_j \sum_i \left(Y_{ij} - M_j\right)^2 \quad (3)$$

This formula is very similar to the SS_{total} formula, except that we are comparing each participant's score to their respective group/condition mean, instead of the grand mean. Adding $SS_{between}$ and SS_{within} gives us SS_{total}.

That is, SS_{total} consists of two components: SS_{between} and SS_{within}. Keep in mind, this is for a one-way between-subjects ANOVA.

Now that we have partitioned the variance, we can see *how much of the total variance* is due to a particular component. That is, effect size. Let us start by using the eta squared measure of effect size. Eta squared gives an intuitive use of the partitioned variance:

$$\eta^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}} \tag{4}$$

Eta squared expresses the between-groups variance as a proportion of the total variance, telling us how much of the variance in dependent variable Y can be attributed to independent variable A. For example, how big of a role did study time play in performance on our memory test? Eta squared can be thought of conceptually as either the *proportion of variance accounted for*, or as a *proportional reduction in error/uncertainty*. As a proportion, its possible values range from 0 to 1. So, for example, a value of .50 would mean that independent variable A accounts for 50% of the variance in dependent variable Y. This makes sense, and eta squared works well as a descriptive statistic of our sample data. However, there are problems with eta squared.

What is Wrong With Eta Squared?

In reporting on ANOVA, eta squared and partial eta squared are the most popular effect sizes (Peng et al., 2013; Zhou & Skidmore, 2017). It has become common practice in statistics to use Greek letters to indicate a population parameter (American Psychological Association, 2020; Keppel & Wickens, 2004). This might give the impression that eta squared (η^2) is an estimator for the effect size in the population. This however, is not the case. Eta squared is simply a descriptive statistic of the sample data

(Lakens, 2013; Maxwell et al., 2018; Tabachnick & Fidell, 2007a).[6] If we want to make inferences to the broader population of people (i.e., everyone not in our sample), eta squared is flawed. This is why eta squared is also known as R^2 or the correlation ratio in the context of regression (Keppel & Wickens, 2004).

To understand the problem with eta squared, let us remind ourselves about the difference between population and sample. In psychology research, the population is the hypothetical set of all possible participants of interest (e.g., all humans, past, present, and future), whereas the sample is a single finite set of participants drawn from that population. Descriptive statistics merely describe our sample data. Inferential statistics draw conclusions about the population from the sample.

Some sample statistics, such as the mean, are unbiased estimators of their corresponding population parameters, meaning that the expected valued of the sample mean in the long run is equal to the population mean. That is, if we were to endlessly draw new samples from the population, calculate the sample mean each time, and build a sampling distribution out of those means, the mean of that sampling distribution would be equal to the population mean.

However, some sample statistics are biased estimators, such as the variance. The expected value of the sample variance is smaller than the actual population variance. An adjustment is necessary to make an unbiased estimate of the population variance. That adjustment is Bessel's correction, which uses n-1 in the denominator of the variance formula, instead of n. This adjusted version of the sample variance is often denoted as s^2.

Eta squared is a ratio of sums of squares calculated from the sample data (SS_{between} / SS_{total}). Just as the unadjusted sample variance is a biased

[6] The persistent misleading use of the Greek letter eta is probably a result of the notation used by SPSS.

estimator of the population variance, an unadjusted measure based on sample sums of squares will also be a biased estimator of the population value. In this case, eta squared tends to *overestimate* the true effect size, especially when it is calculated from small samples. The reason for this flaw is that the numerator, $SS_{between}$, consists of variance due to the factor, *as well as* some random variance in the means of the groups of that factor. This is due to sampling error. Even when the population group means do not differ, the sample group means will always differ somewhat. This is due to the random chance involved in sampling participants from the population (Ellis, 2010; Maxwell et al., 2018; Myers & Well, 2003). In eta squared, these coincidental differences between sample groups are treated as systematic (Keppel & Wickens, 2004). Treating coincidental differences as systematic results in positive bias, meaning it will overestimate the effect, especially with small sample sizes (Albers & Lakens, 2018; Keppel & Wickens, 2004; Lakens, 2015) and even more so for partial eta squared (Levine & Hullett, 2002). Simultaneously, as the explained variance due to the factor is overestimated (numerator), the unexplained variance is underestimated (denominator, Maxwell et al., 2018). Multiple studies using Monte Carlo simulations have shown the extent of the positive bias that occurs in eta squared and partial eta squared (Keselman, 1975; Okada, 2013; Yigit & Mendes, 2018). Recently, Liu (2022) proposed a bootstrapping method to correct for bias in eta squared. This method requires some proficiency with the program R. Importantly, this method is only suitable for one-way between-subjects ANOVAs and not for other designs. The positive bias in eta squared and partial eta squared is problematic as it can lead to overvaluing effects, as well as underpowering subsequent studies. In a 2015 blogpost, statistician Daniël Lakens wrote: "If η^2 was a flight from New York to Amsterdam, you would end up in Berlin." Although that may be a bit of an overstatement (Albers, 2015), the bias in eta squared should motivate researchers to choose a better measure of effect size, namely omega squared.

Development/History of Omega Squared

The origins of the effect size omega squared are somewhat murky. We can broadly think of most effect size measures as belonging to one of two families: standardized differences (such as Cohen's *d*), and associative strength (such as Pearson's *r*; Ellis, 2010, pp. 6-15). The ANOVA effect sizes belong to the latter. Using Google Books Ngram Viewer and the PsychINFO database, we conclude that the standard version of omega squared was introduced by Hays (1963) in the first edition of his well-regarded graduate-level statistics textbook (pp. 323-332, and especially pp. 381-384). Several other sources also point to Hays (1963) as the origin (Dwyer, 1974; Keren & Lewis, 1979; Sechrest & Yeaton, 1982).[7]

However, there were several precursors to omega squared that have been mentioned in histories by other authors (Dwyer, 1974; Glass & Hakstian, 1969; Huberty, 2002; Keren & Lewis, 1979). Examining these precursors can help readers in understanding effect sizes in general and omega squared specifically. It is unclear whether Hays knew of these precursors, as he did not cite anything when introducing omega squared, and simply justified his use of the omega symbol by referring to it as a "relatively neutral symbol," presumably meaning one that had not already been used very much for other measures.

The general idea of quantifying the strength of association dates back to at least Pearson (1905)[8] who defined and labeled eta as the correlation ratio. Eta squared was introduced for regression by Pearson (1911), and for ANOVA by Fisher (1925, 1928). As stated before, the squared correlation ratio, η^2 ($SS_{\text{between}}/SS_{\text{total}}$), is synonymous with R^2 (aka $r^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}}$), which is often called the coefficient of determination. The term

[7] There are other statistics that are referred to as omega: Cramer-von-Mises omega squared, an alternative to the Kolmogorov-Smirnov test; Cohen's Omega, an effect size for Chi Square tests; and McDonald's coefficient omega, an alternative to Cronbach's alpha measure of reliability. These are unrelated to the omega squared measure of effect size for ANOVA.

[8] For a more thorough history of effect sizes, see Huberty (2002).

R^2 is used in the context of regression, whereas η^2 is used in the context of ANOVA. To the best of our knowledge, omega squared is not used in regression analysis.

Several measures descended from eta squared. Bolles and Messick (1958) proposed a utility index or coefficient of utility, *U*, which appears to be equivalent to eta squared (Gaito & Firth, 1973). *U* has not apparently had a lasting impact, but was notable for its early emphasis on the importance of an effect (i.e., its utility). Another relative of eta squared was the intraclass correlation coefficient (rho-i, ρ_i), introduced by Fisher (1925) for use with random effects; it is analogous to omega squared which is used more with fixed effects. For fixed effects, Kelley (1935) recognized the bias in eta squared, which tends to overestimate the population effect size, and developed epsilon squared (ε^2) as an improvement. Epsilon squared is equivalent to adjusted R^2 in regression contexts (Vogt & Johnson, 2015, p. 142), and is very similar to omega squared, differing only in the denominator. Finally, Hays (1963) developed omega squared, also intended mainly for fixed effects ANOVA. In the years since then, there have been refinements of and debates about omega squared, including epsilon versus omega (Glass & Hakstian, 1969), slightly unequal sample sizes (Vaughan & Corballis, 1969), formulas for designs that include random effects (Dodd & Schultz, 1973), formulas for MANOVA and ANCOVA (Olejnik & Algina, 2000), the introduction of partial omega squared (Keren & Lewis, 1979), and the introduction of generalized omega squared (Olejnik & Algina, 2003).

It is worth considering the conceptual formulas that would be used to calculate omega squared if we somehow magically knew population values. A first helpful stepping stone in understanding the formulas for the effect sizes is provided by Keppel & Wickens (2004; p. 162), who express the idea of effect sizes in words:

$$\text{effect size} = \frac{\text{variability explained}}{\text{total variability}} = \frac{\text{total variability} - \text{unexplained variability}}{\text{total variability}} \quad (5)$$

Effect size, in this context, represents the proportion of variability that is accounted for by an effect (aka variable, factor, treatment). Variance in the population can be represented by sigma squared (σ^2). This results in the population formulas for omega squared shown in Table 2.1, which are all equivalent to each other.

Table 2.1

Population formulas for omega squared.

Formula	Source
$\omega^2 = \frac{\sigma^2_Y - \sigma^2_{Y\|X}}{\sigma^2_Y}$	Hays (1963, pp. 325, 381-382)
$\omega^2 = \frac{\sigma^2_Y - \sigma^2_e}{\sigma^2_Y}$	Maxwell et al. (1981, pp. 526-527)
$\omega^2 = \frac{\sigma^2_A}{\sigma^2_A+\sigma^2_e}$ $\omega^2 = \frac{\sigma^2_\alpha}{\sigma^2_\alpha+\sigma^2_\epsilon}$ $\omega^2 = \frac{\sigma^2_A}{\sigma^2_A+\sigma^2_{S/A}}$	Myers & Well (2003, p. 208), Vaughan & Corballis (1969, p. 206), Keppel (1991, p. 64)
$\omega^2 = \frac{\sigma^2_\alpha}{\sigma^2_Y}$	Cardinal and Aitken (2005)

The overall logic is the same across all these forms of the population formula. Starting with the first row (formula 10.19.2 in Hays, 1963), σ^2_Y is the total population variance of variable Y, and $\sigma^2_{Y|X}$ is the population variance of Y within a particular group/condition of variable X. That is, $\sigma^2_{Y|X}$ is the remaining variance left in Y (DV) given that you know X (IV). Assuming equal variance of Y across all groups/conditions of X (i.e., homogeneity of variance, aka homoscedasticity, which is an assumption underlying ANOVA), $\sigma^2_{Y|X}$ is simply the error variance, σ^2_e, which is the leftover variance due to unkown variables. Thus, $\sigma^2_Y - \sigma^2_{Y|X}$ gives us the variance in Y attributable to variable X. Dividing that by σ^2_Y gives us the proportion of Y's variance that is attributable to X, or the proportional

reduction in uncertainty. This last interpretation is because we can consider variance to be a kind of uncertainty (aka error). Say we want to predict one person's score on a memory test. Our best guess is the mean score across all people; but the person's actual score might be some distance from that mean. Exactly how far? We are uncertain. But the smaller the variance around that mean, the closer we can get to making a good guess, and thus the less uncertain we are. The next formula, from Maxwell et al., 1981 (pp. 526-527) simply replaces $\sigma^2_{Y|X}$ with σ^2_e.

The three formulas in the third row switch to using A or α as the label for the IV, instead of X. These formulas use the single term σ^2_A or σ^2_α to represent the variance in Y that is attributable to X, then divide by a sum that yields the total variance of Y. Finally, the simplest population formula is provided by Cardinal and Aitken (2005): the variance in Y due to X, divided by the total variance in Y.[9]

In reality, we cannot truly know the population variances, because for the purpose of inferential statistics in psychology, we typically conceive of the population as an infinitely large hypothetical distribution of all possible individuals of interest. Thus, the best we can do is to estimate those variances from sample data. This returns us to the bias in eta squared, and how omega squared improves upon that. Omega squared corrects for bias by both shrinking the numerator and enlarging the denominator.[10]

Hays (1963) made these corrections by using a combination of ANOVA expected mean squares (E(*MS*)) and some "nasty" algebra (Keppel & Wickens, 2004, p. 163). The E(*MS*) are based on the idea of repeatedly

[9] Some authors have argued that the very idea of total population variance of Y is nonsensical. See Maxwell et al. (1981) for discussion.

[10] Another very similar measure of effect size, epsilon squared, just shrinks the numerator from eta squared, and does not change the denominator. For detailed comparison of epsilon squared versus omega squared, see Glass and Hakstian (1969), and Carroll and Nordholm (1975).

drawing samples from the population and calculating $MS_{between}$ and MS_{within} for each sample, resulting in two sampling distributions of these two *MS* values (Myers & Well, 2003, p. 203). The means of these sampling distributions are the E(*MS*) and they play an important role in ANOVA. Myers and Well (2003) present the following formula for ω^2 (p. 208; see also our Table 2.1):

$$\omega^2 = \frac{\sigma_A^2}{\sigma_A^2+\sigma_{error}^2} \tag{6}$$

The numerator (σ_A^2) represents the variance of the treatment effect, while the denominator ($\sigma_A^2 + \sigma_{error}^2$) represents total population variance. Each of these components can be replaced with formulas for the E(*MS*). By using the E(*MS*) formulas, the formula for ω^2 can be rewritten as the estimate of the population parameter. In the general formula for omega squared we replace "between" with "effect", and add a hat (circumflex) on top of the omega, to indicate that it is an estimator. Myers and Well (2003, p. 209) rewrite the formula as follows:

$$\widehat{\omega}^2 = \frac{[(a-1)/a](1/n)(MS_{effect}-MS_{within})}{[(a-1)/a](1/n)(MS_{effect}-MS_{within})+MS_{within}} \tag{7}$$

where *a* is the number of groups/conditions in Factor A. Using algebra, this formula can be rewritten more simply, using components from the ANOVA table:

$$\widehat{\omega}^2 = \frac{SS_{between}-(df_{effect}\times MS_{within})}{SS_{total}+MS_{within}} \tag{8}$$

Let us first consider the denominator, which contains the total variance (SS_{total}), as well as the mean square of the error variance (MS_{within}). By adding MS_{within}, the denominator is enlarged compared to the formula for η^2 (Equation 4). It may be counterintuitive that a measure for error variance is added, while SS_{total} already contains SS_{within}. This

can be explained by the formulas for E(*MS*), as they follow the argument that the *F*-ratio ($MS_{\text{between}}/MS_{\text{within}}$) equals 1 when the null hypothesis is true (Keppel & Wickens, 2004, p. 36). The idea is that, when the null hypothesis is true, MS_{between} and MS_{within} are both estimates of σ^2_{error}, the error variance (Myers & Well, 2003, p. 203). Therefore, it is argued that we should "choose an error term such that its E(*MS*) and the E(*MS*) of the term to be tested are identical when the null hypothesis is true" (p. 204). The E(*MS*) formula for σ^2_A in the context of a one-way between-subjects ANOVA is:

$$\sigma^2_A = n\theta^2_A + \sigma^2_{\text{error}} \tag{9}$$

This means that σ^2_{error} is part of the E(*MS*) formulas for σ^2_A as well as added (again) to the total population variance (Glass & Hakstian, 1969, p. 406). MS_{within} is considered to be a good estimator for σ^2_{error}.[11]

Now let us consider the numerator. Since the numerator of eta squared (SS_{between}) includes variance due to the independent variable A as well as variance due to error, we need to subtract out the error variance amount. The exact amount to be subtracted is worked out using the E(*MS*) and degrees of freedom of MS_{between} and MS_{within}. MS_{within} represents the amount that each group mean in the sample is expected to vary from its respective mean in the population. The number of times we subtract MS_{within} is the degrees of freedom of our IV: the number of independent observations (i. e., number of conditions in the IV) minus one for our estimation of the overall population mean from sample data.

The formula for standard omega squared for one-way between-subjects designs can also be rewritten as a function of the values of *F*,

[11] For more in-depth information on E(*MS*) and their use in ANOVA and $\hat{\omega}^2$, we recommend: Carroll & Nordholm, 1975; Dodd & Schultz, 1973; Glass & Hakstian, 1969; Hays, 1963; Maxwell et al., 1981; Myers & Well, 2003; Vaughan & Corballis, 1969.

df_{effect}, and N (Carroll & Nordholm, 1975; Keppel & Wickens, 2004; Maxwell et al., 2018):

$$\widehat{\omega}^2 = \frac{df_{\text{effect}}(F_{\text{effect}}-1)}{df_{\text{effect}}(F_{\text{effect}}-1)+N} \tag{10}$$

Conceptually, this formula may be the most clear in how $\widehat{\omega}^2$ corrects for the sampling error which is present in η^2. As stated before, when there is no treatment effect and the null hypothesis holds true, the F-value will approximate 1, as MS_{between} and MS_{within} in that case are both estimates of σ^2_{error} (Keppel & Wickens, 2004, p.164). By subtracting 1 from F, the sampling error is corrected. Because of this correction, $\widehat{\omega}^2$ is always smaller than η^2.

To sum up, effect size for ANOVA consists of the proportion of total variance in a DV that is attributable to an IV. Eta squared calculates this very simply from sample data. Omega squared corrects for the bias in eta squared by adjusting both the numerator and the denominator, providing a better estimate of the effect size in the broader population. Both eta squared and omega squared use values that are readily available in an ANOVA summary table.

Underuse of Omega Squared

Bias in eta squared was described over 80 years ago (T. L. Kelley, 1935) and the less-biased alternative omega squared was proposed 60 years ago (Hays, 1963). Still, omega squared is rarely used. There are several explanations for this. Traditionally, there has been a heavy emphasis on significance testing, while behavioral scientists were educated far less on effect sizes and power analysis (APA, 2020; Cohen, 1994; Hyde, 2001; Keppel, 1991). The emphasis on the importance of effect size is relatively new. An explicit recommendation to utilize effect sizes was added to the fourth publication manual of the American Psychological Association (APA) in 1994, but this did not result in increased use of effect sizes in the

following years (Hyde, 2001; Thompson, 1999a, 1999b; L. Wilkinson, 1999). This was perhaps due to the large number of editions of the publication manual. The following editions have paid increasingly more attention to effect sizes. The fifth edition of the publication manual mentioned that "it is almost always necessary to include some index of effect size or strength of relationship in your Results section" (APA, 2001, p. 26). In the sixth and seventh editions, effect sizes have been added as a requirement for all publications, and guidance was added on what types of effect size should be reported (APA, 2010, 2020). The reporting of effect sizes has increased considerably since the late nineties; research over the past two decades shows that approximately half of the reported ANOVA tests are accompanied by any measure of effect size (Alhija & Levy, 2009; Fritz et al. , 2012; Peng et al., 2013; Sun et al., 2006; Zhou & Skidmore, 2017).

The shortcomings of (partial) eta squared have not been discussed in the publication manual (APA, 2020). Despite the overall increase in reporting of effect sizes, eta squared and partial eta squared have continued to dominate, as reviewed by Peng et al., 2013 (see also: Alhija & Levy, 2009; Barry et al., 2016; Fritz et al., 2012; Kirk, 1996). For example, Fritz et al. (2012) examined articles published in the *Journal of Experimental Psychology: General* in 2009 and 2010, and out of all the articles that reported ANOVA results, they found only one use of omega squared, compared to 32 uses of eta squared (either standard or partial). Although not as precise as the manual counting done for review articles like those cited here, we conducted a search of PsycINFO for publications in 2019 or 2020, and we found 454 results that contained any of the following search terms: η^2, $\eta\hat{}2$, η_p^2, or $\eta p\hat{}2$. By contrast only 18 results contained any of the following search terms: ω^2, $\omega\hat{}2$, ω_p^2, or $\omega p\hat{}2$. Thus, the underuse of omega squared continues.

The positive bias in eta squared and partial eta squared may actually be one of the reasons why they are preferred over other effect size measures (Fritz et al., 2012). Other reasons for the high prevalence of eta squared and partial eta squared are likely familiarity and convenience.

Partial eta squared can be automatically produced by the most used statistical packages, like SPSS. This easy accessibility promotes its use, even when it is not appropriate (Kirk, 1996; Zhou & Skidmore, 2017). Levine and Hullett (2002) further documented problems with effect sizes reported from SPSS. Omega squared was not included at all in SPSS until the 27th version in 2020 (Mathew, 2020), and then only for the one-way between-subjects ANOVA design.

The formulas for omega squared and partial omega squared differ across designs, which makes them more cumbersome to calculate than eta squared and partial eta squared. Finding the right formulas, and guidance on how to make calculations, is difficult. On the one hand, much information is published only in statistical textbooks. This content is often not indexed in commonly used internet search engines. On the other hand, and perhaps most importantly, there is no definite consensus on what formulas to use to calculate omega squared (Maxwell et al., 2018). This holds especially true for multifactor ANOVAs with one or more within-subjects factors. While some authors argue that it is impossible to calculate omega squared in these models (Keppel & Wickens, 2004), others recommend it (Gaebelein & Soderquist, 1978). Moreover, most publications only supply formulas for one- and two-way ANOVA designs, due to the large number of possible formulas (for an exception, see Dodd & Schultz, 1973; who also present formulas for three-way designs). Lastly, different authors often use different symbols and subscripts to describe the same components; while in other instances one term is used to describe opposing constructs. This can be very confusing, especially for inexperienced researchers. To aid in demystifying omega squared, we have included a disambiguation table where we clarify the different notations used in a selection of papers (see Appendix A), and we have done the legwork of gathering and verifying the formulas to use for the most common ANOVA designs.

Formulas for Omega Squared

In this section we will present formulas for omega squared and partial omega squared for ANOVA designs with fixed factors. For designs with random factors, we recommend readers to consult Dodd and Schultz (1973), Olejnik and Algina (2000), and Vaughan and Corballis (1969). A quick overview of the formulas can be found in Appendix B. For each design, we will show an example of output from SPSS (version 28), highlighting where each component of the relevant formula can be found. We recommend calculating the formulas using widely available software such as Microsoft Excel, and we provide examples of this in the supplemental materials (see Appendix D).

Between-Subjects Designs

One-Way Between-Subjects Designs

Conceptually, the effect size in the population is estimated with the following formula (see also Table 2.1):

$$\omega^2 = \frac{\sigma^2_{\text{A}}}{\sigma^2_{\text{A}}+\sigma^2_{\text{error}}} \qquad \text{(6, repeated)}$$

For any size between-subjects design with fixed factors, the formula for standard omega squared can be expressed in terms found in the ANOVA table (Dodd & Schultz, 1973):

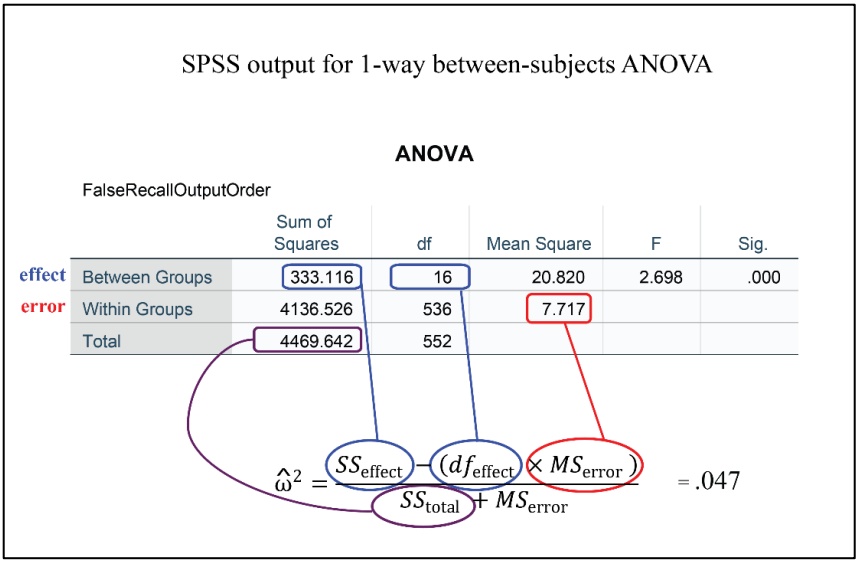
$$\widehat{\omega}^2 = \frac{SS_{\text{effect}}-(df_{\text{effect}}\times MS_{\text{error}})}{SS_{\text{total}}+MS_{\text{error}}} \qquad (11)$$

In the one-way between-subjects ANOVAs, the subscripts *between* and *within* are often used (see Formula 8). In other designs, these subscripts are replaced with *effect* and *error* respectively. For consistency, we will use *effect* and *error* in the formulas for all designs. See Figure 2.1 for an example of SPSS output from a one-way between-subjects ANOVA, showing the components used for omega squared. In this example,

$\widehat{\omega}^2$ = .047. By comparison, η^2 = .075. Notice that $\widehat{\omega}^2 < \eta^2$, as expected. Note also that we report omega squared to three decimal places, and never include a leading zero.

Figure 2.1

The Components for Standard Omega Squared in SPSS Output For a One-Way Between-Subjects Design.



Note. Data are from Finley et al. (2017), available in the supplemental materials. SPSS version 28.

Multi-Factor Between-Subjects Designs and Partial Omega Squared

For multi-factor between-subjects designs (aka higher order designs; two-way, three-way, etc.) it can be argued that it is inappropriate to use the standard omega squared formula (Keppel & Wickens, 2004). This is because the estimated total variance ($\widehat{\sigma}^2_{total}$, see Table 2.1) varies across designs. So, in a one-way design, the formula for the effect size of Factor A would be:

$$\widehat{\omega}^2 = \frac{\widehat{\sigma}_A^2}{\widehat{\sigma}_A^2 + \widehat{\sigma}_{error}^2} \qquad \text{(6, estimated version)}$$

And in a two-way design, where Factor B is added, the formula would be:

$$\widehat{\omega}^2 = \frac{\widehat{\sigma}_A^2}{\widehat{\sigma}_A^2 + \widehat{\sigma}_B^2 + \widehat{\sigma}_{AB}^2 + \widehat{\sigma}_{error}^2} \qquad (12)$$

This means that even when the estimated variance components for Factor A and the random error are identical in both designs, this will result in a different value for the estimated effect size. Therefore, *partial omega squared* is proposed, which always consists of the same components as the one-way design, regardless of the total number of factors in the design (Keppel, 1991). The effects of other factors are *partialled out* (Keren & Lewis, 1979) and can be written as:

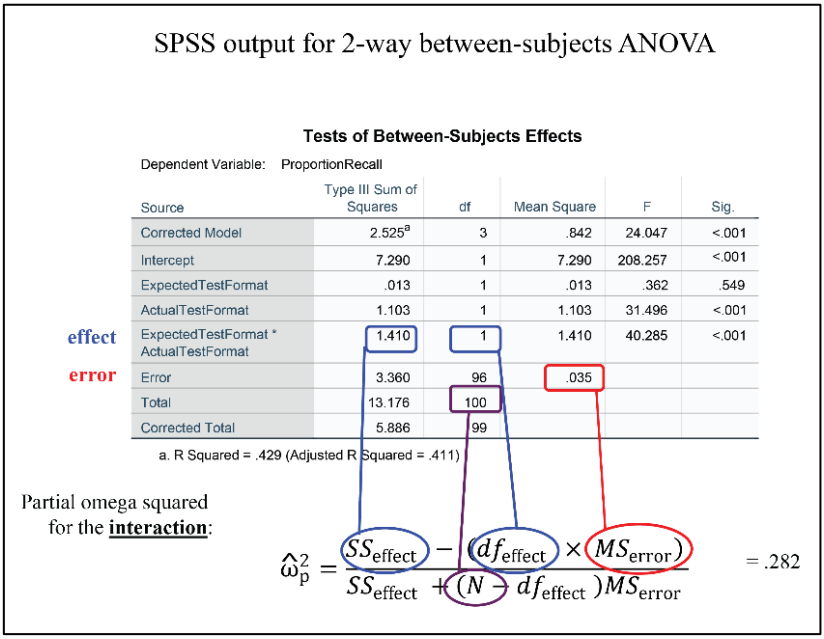
$$\widehat{\omega}_p^2 = \frac{SS_{effect} - (df_{effect} \times MS_{error})}{SS_{effect} + (N - df_{effect})MS_{error}} \qquad (13)$$

Standard omega squared represents the proportion of the total variance explained by the effect of one factor. Partial omega squared represents the proportion of variance explained *exclusively* by one factor that is *not* explained by other factors in the model. As the denominator for partial omega squared is smaller, this will always result in a higher value. The standard effect size is useful to compare the effect of the various factors within the design. This is impossible with the partial effect size, as they do not share the same denominator (Sechrest & Yeaton, 1982). The partial effect size is suitable for power analysis (Keppel & Wickens, 2004) and usually more suited to make comparisons across designs (Fritz et al., 2012; Keppel & Wickens, 2004; Levine & Hullett, 2002). We will go into the matter of partialling out factors and comparability further in the section on obstacles in using omega squared.

See Figure 2.2 for an example of SPSS output from a two-way between-subjects ANOVA, showing the components used for partial omega squared.

Figure 2.2

The Components for Partial Omega Squared in SPSS Output for a Two-Way Between-Subjects Design.



Note. Data are from Finley and Benjamin (2012), available in the supplemental materials. SPSS version 28.

The formulas for omega squared and partial omega squared in between-subject designs can be rewritten to be computed from the *F*-test statistic, *df*, and *N* (Maxwell et al., 2018). This is especially straightforward for partial omega squared (which is the same as the formula for standard omega squared in one-way designs, see Formula 10).

$$\widehat{\omega}^2 = \frac{df_{\text{effect}}(F_{\text{effect}}-1)}{\sum_{\text{all effects}}(df_{\text{effect}}F_{\text{effect}})+df_{\text{error}}+1} \quad (14)$$

$$\widehat{\omega}_{\text{p}}^2 = \frac{df_{\text{effect}}(F_{\text{effect}}-1)}{df_{\text{effect}}(F_{\text{effect}}-1)+N} \quad (15)$$

As the sample size and levels of the factors should be reported in any research paper, this means that omega squared and partial omega squared in between-subjects designs can often be calculated even in absence of an ANOVA table. For standard omega squared this requires the *F*-values and *df* for all factors in the design.

Within-Subjects Designs

About Within-Subjects Factors

When all subjects receive all levels of a treatment, the treatment is a within-subjects factor (Keppel & Wickens, 2004, p. 347). As the subjects are measured more than once, they are a potential source of variance (Tabachnick & Fidell, 2007a, p. 249). A one-way within-subjects design can therefore be thought of as a two-factor design, often indicated as an $\overline{\text{A}}\times\text{S}$ design. Factor $\overline{\text{A}}$ is the fixed within-subjects factor (a line is added above the letter to signify a within-subjects factor) and the subjects form the second, random between-subjects Factor S (Dodd & Schultz, 1973; Keppel & Wickens, 2004; Olejnik & Algina, 2000). The presence of Factor S complicates the conceptualization of standard and partial omega squared in ANOVAs that include within-subjects designs. This has led some authors to claim that it is impossible (Keppel, 1991; Keppel & Wickens, 2004) or at least problematic (Gaebelein & Soderquist, 1978; Vaughan & Corballis, 1969) to calculate omega squared for within-subjects designs. We will describe these complications and subsequently argue that omega squared can indeed be calculated.

When we insert a one-way within-subjects design into a table, this table holds one observation in each cell (Keppel & Wickens, 2004; Olejnik & Algina, 2000; Tabachnick & Fidell, 2007a), as shown in Table 2.2.

Table 2.2

One-way Within-Subjects ($\bar{A}\times$ S) ANOVA Design With a = 3 Conditions and N = 3 Subjects.

	Conditions of Factor A		
Subjects	a1	a2	a3
s1	Y_{11}	Y_{12}	Y_{13}
s2	Y_{21}	Y_{22}	Y_{23}
s3	Y_{31}	Y_{32}	Y_{33}

In a one-way within-subjects design, there are two identifiable sources of variance: variance due to treatment $\overline{A}$ ($\hat{\sigma}^2_{\text{effect}}$) and variance due to the systematic differences between subjects ($\hat{\sigma}^2_{\text{subject}}$). The remaining variance in the model is indicated with $\hat{\sigma}^2_{\text{effect}\times\text{subject}}$. Ideally, $\hat{\sigma}^2_{\text{effect}\times\text{subject}}$ is made up solely of random variance (error). This is the strength of within-subjects designs compared to between-subjects designs (Keppel & Wickens, 2004; Lakens, 2013; Loftus & Masson, 1994). In a between-subjects design, no distinction can be made between random error and systematic differences between subjects, and the total variance is therefore simply defined as:

$$\hat{\sigma}^2_{\text{total}} = \hat{\sigma}^2_{\text{effect}} + \hat{\sigma}^2_{\text{error}} \tag{16}$$

While in a within-subjects design, $\hat{\sigma}^2_{\text{error}}$ is replaced with $\hat{\sigma}^2_{\text{subjects}} + \hat{\sigma}^2_{\text{effect}\times\text{subject}}$. There are two ways to define the total variance for within-subjects designs (Keppel, 1991):

$$\hat{\sigma}^2_{\text{total}} = \hat{\sigma}^2_{\text{effect}} + \hat{\sigma}^2_{\text{subjects}} + \hat{\sigma}^2_{\text{effect}\times\text{subject}} \tag{17}$$

And

$$\hat{\sigma}^2_{\text{total}} = \hat{\sigma}^2_{\text{effect}} + \hat{\sigma}^2_{\text{effect} \times \text{subject}} \quad (18)$$

Excluding the variance due to subjects from the total variance leads to higher effect sizes and power, which are important reasons for authors to adopt a within-subjects design (Keppel, 1991; Keppel & Wickens, 2004; Lakens, 2013). This can be seen as partialling out the random Factor S, similar to how one would calculate partial effect sizes in a multi-factor between-subjects design. However, there are three important arguments *not* to partial out the systematic differences between subjects. Firstly, because of the viewpoint that effect sizes are meant to provide a standard metric that can be used across designs (Lakens, 2013; Maxwell et al., 2018). As the systematic differences between subjects are not partialled out in between-subjects designs, they should not be partialled out in within-subjects designs either, for comparability purposes. The higher effect sizes found in within-subjects designs when the systematic differences between subjects are partialled out, are therefore often seen as overestimations of the actual effect size (Lakens, 2013; Maxwell et al., 2018; Olejnik & Algina, 2003). Secondly, systematic differences between individuals are seen as part of the total population variance (Maxwell et al., 2018). It appears illogical to disregard them as such in within-subjects designs. The third argument relates to the assumption of additivity, which we will discuss below. We will then present formulas in which the variance due to subjects is not partialled out. In the section on obstacles for using omega squared we will discuss alternatives in which the subject variance can be partialled out.

Non-Additivity

The third argument to not partial out the variance due to subjects from the total variance relates to (non-)additivity. ANOVAs for designs that

include one or more within-subjects factors have to meet the assumption of additivity, which is a component of the assumption of sphericity (Tabachnik & Fidell, 2007). The assumption of sphericity holds true when the variances across conditions, as well as the covariances between pairs of conditions, are equal (Field, 2017, p. 654). The F test is not robust against violation of the assumption of sphericity, and violations increase the chance of Type I error (Maxwell et al., 2018). Departure from sphericity can be measured. It is denoted with a lower case epsilon (ε) and when the assumption of sphericity is met perfectly, it has a value of 1 (Maxwell et al. 2018). There are three well-known estimators for epsilon: Greenhouse-Geisser, Huynh-Feldt, and the lower-bound estimate of sphericity. In SPSS output for an ANOVA that includes within-subjects factors, those estimators are listed in a table titled "Mauchly's Test of Sphericity". When the assumption of sphericity is false, the estimators will be smaller than 1 and a correction should be made to the degrees of freedom. Mauchly's test itself is problematic, especially for small samples (Field, 2017, p. 655; Maxwell et al., 2018, p. 629), but the general idea is that if the test is statistically significant, then corrected *df* should be used.[12] The corrected *df* are listed as separate rows in the "Tests of Within-Subjects Effects" table of the SPSS output. While using the corrected degrees of freedom increases the *p*-value, it does not impact omega squared. This is because adjustments are made to all components used to calculate omega squared, so while the absolute values are changed, the ratio between the numerator and the denominator remains the same. The bottom line is that for calculating omega squared, it will not matter whether or not you use *df* corrected for violation of sphericity.

Additivity and sphericity are regularly discussed as if they are interchangeable, but in fact additivity is a component of sphericity. For a

[12] For guidelines on how to assess sphericity and which estimator to choose, see Maxwell et al. (2018, pp. 627-634) and Tabachnick and Fidell (2007, pp. 284-288).

more thorough explanation of the relation between sphericity and additivity see Tabachnick and Fidell (2007, pp. 247-248, pp. 284-288). As a general term, additivity signifies that there is no interaction between factors. Take for instance the effect of fertilizer and picking the bugs off our fictional rosebushes. We have tried these measures separately and know our roses grow 3cm a week because of fertilizer and 2cm because we remove the bugs. Now we decide to treat our bushes to both of the treatments at the same time. If the model is additive, the roses grow 5cm in total. But if they suddenly grow 7cm the model is non-additive, the bug-picking and fertilizer *interact* and contribute to an additional 2cm of growth.

In within-subjects designs, a distinction can be made between additive and non-additive models (Dodd & Schultz, 1973; Keppel, 1991; Tabachnick & Fidell, 2007a; Vaughan & Corballis, 1969; Winer, 1962). This does not refer to the possible interaction of the factors we are investigating, but to the possible interaction of the random subjects factor S with the IVs. Recall that in the within-subjects design we have one variance component that simultaneously indicates random variance (error) and the interaction between the treatment and the subjects: $\widehat{\sigma}^2_{\text{effect} \times \text{subject}}$ (Tabachnick & Fidell, 2007a). Ideally, we would have an additive design, where there is no interaction between the treatment and the subjects. When there is no interaction, $\widehat{\sigma}^2_{\text{effect} \times \text{subject}}$ represents only random error and this gives us the possibility to partial out the systematic differences between subjects ($\widehat{\sigma}^2_{\text{subject}}$). However, in any design that includes at least one within-subjects factor with more than two levels it is unreasonable to assume that the model is additive (Keppel, 1991; Tabachnick & Fidell, 2007a). In the social and behavioral sciences it is practically impossible to imagine a repeated measure that will not interact with the individual (Tabachnick & Fidell, 2007a, p. 248). It is often impossible to differentiate which part of $\widehat{\sigma}^2_{\text{effect} \times \text{subject}}$ is due to random error and which is due to the interaction between the subjects and the IV (Dodd & Schultz, 1973; Gaebelein & Soderquist, 1978; Keppel & Wickens, 2004; Vaughan &

Corballis, 1969). A test for non-additivity was devised by Tukey (Tukey, 1949; Winer, 1962) and is included in statistical packages like SPSS (Myers & Well, 2003) and SAS (Zambarano, 1992). However, this test is only suitable to assess one specific type of non-additivity, while other kinds exist (Dodd & Schultz, 1973; Zambarano, 1992). Zambarano (1992) provides some guidance on other methods to assess additivity, but this requires advanced statistical knowledge and still does not lead to a conclusive answer whether or not a model is additive. In practice, additivity is not often tested for; tests for sphericity and homogeneity of variance are seen as sufficient (Tabachnick & Fidell, 2007a). When a model is completely additive, this automatically means that the assumption of sphericity is met. However, a model can be non-additive while still meeting the assumption of sphericity through compound symmetry, and without inflating the *F* test. Thus, meeting the assumption of sphericity does not ensure additivity. A model where non-additivity may cause problems can be improved by adding a between-subjects "blocking" factor (Tabachnick & Fidell, 2007a, p. 285). For instance, if it is expected that native speakers recall more words in a memory test than non-native speakers, mother tongue can be added as a between-subjects factor. The interaction of mother tongue with the factor is then removed from the error term. Still, other sources of non-additivity could remain. It is also possible to transform the data to an additive scale (Myers & Well, 2003), although it can be challenging to find guidance on how to exactly transform the data. Some authors conclude that it is impossible to calculate omega squared for within-subjects designs, as it is impossible to estimate each source of error variance independently (Gaebelein & Soderquist, 1978; Keppel & Wickens, 2004).

In summary, non-additivity is a complicated matter. In practice, tests for the assumption homogeneity of variance and sphericity are seen as sufficient when carrying out ANOVAs (Tabachnick & Fidell, 2007a), but this does not solve the conundrum of choosing which formula to use to calculate the effect size. We argue that using the formulas presented in

this paper circumvents the problem by not partialling out variance caused by systematic differences between subjects. At the end of this section on designs that include within-subjects factors, we will discuss alternatives proposed by other authors. As there are many other options, we recommend to always report the formula used when reporting omega squared, especially when within-subjects factors are involved, and to include the full ANOVA table.

One-Way Within-Subjects Designs

For within-subjects designs the formula for standard omega squared is similar to the between-subjects design. The differences lie in the use of the two variance components (Dodd & Schultz, 1973; Keppel & Wickens, 2004; Maxwell et al., 2018):

$$\widehat{\omega}^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{effect}\times\text{subject}})}{SS_{\text{total}} + MS_{\text{subject}}} \tag{19}$$

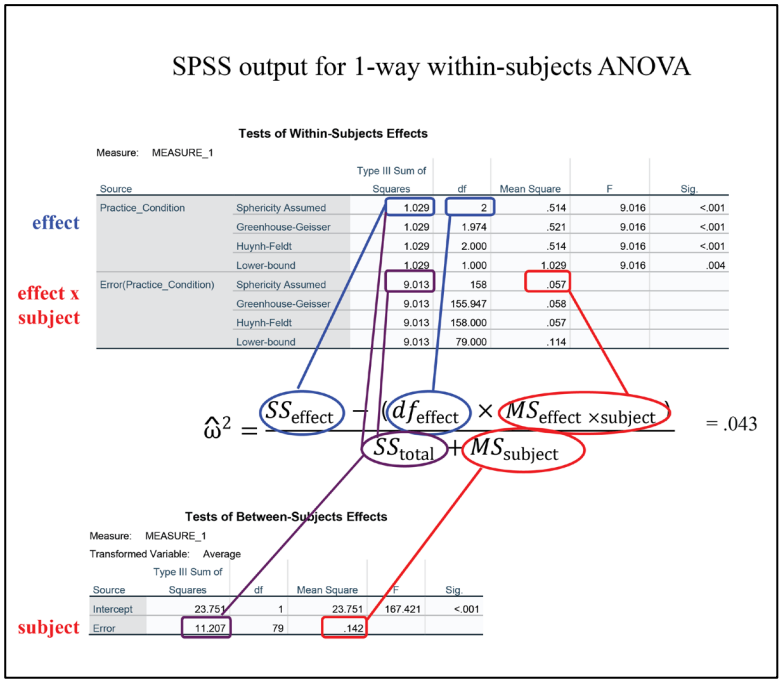
See Figure 2.3 for an example of SPSS output, showing the components used for omega squared. SPSS does not output the total sum of squares for within-subjects designs, and thus it is necessary to calculate SS_{total} from three values, as shown in Figure 2.3. It is necessary to use values from the "Tests of Between-Subjects Effects" table, even though this is a within-subjects design, because that is where the subject factor values can be found (labeled "error" by SPSS). The "intercept" factor in that table should be ignored for our purposes.

The exclusion of SS_{total} from SPSS output may relate to the discussion on whether SS_{subject} should be included in SS_{total}. For formulas that do partial out the systematic differences between subjects we recommend Olejnik & Algina (2000) and Keppel & Wickens (2004). It should be noted that these authors present different formulas that yield quite different results. This is because Olejnik and Algina (2000) add the product of the

error term and the number of *subjects* to the denominator, while Keppel and Wickens (2004) add the number of *observations*.

Figure 2.3

The Components for Standard Omega Squared in SPSS Output for a One-Way Within-Subjects Design.



Note. Data are from Finley et al. (2011), available in the supplemental materials. SPSS version 28.

Multi-Factor Within-Subjects Designs

For multi-factor within-subjects designs (two-way, three-way, etc.) partial omega squared can be calculated. Maxwell et al. (2018, p. 674) state: "We generally believe that the value of omega squared we calculate for a main effect in a factorial design should be identical to the value we would have obtained for that effect in a single-factor design." This reasoning follows the same logic as the partial omega squared formula for higher-order between-subjects designs. For partial omega squared the total sum of squares in the formula of standard omega squared is replaced with the components that form the total sum of squares in the one-way design:

$$\widehat{\omega}_{\mathrm{p}}^{2} = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{effect}\times\text{subject}})}{SS_{\text{effect}}+SS_{\text{effect}\times\text{subject}}+SS_{\text{subject}}+MS_{\text{subject}}} \tag{20}$$

See Figure 2.4 for an example of SPSS output from a two-way within-subjects ANOVA, showing the components used for partial omega squared. Again, the "Tests of Between-Subjects Effects" table is necessary to find the subject values (labeled "error" by SPSS), and the "intercept" factor should be ignored.

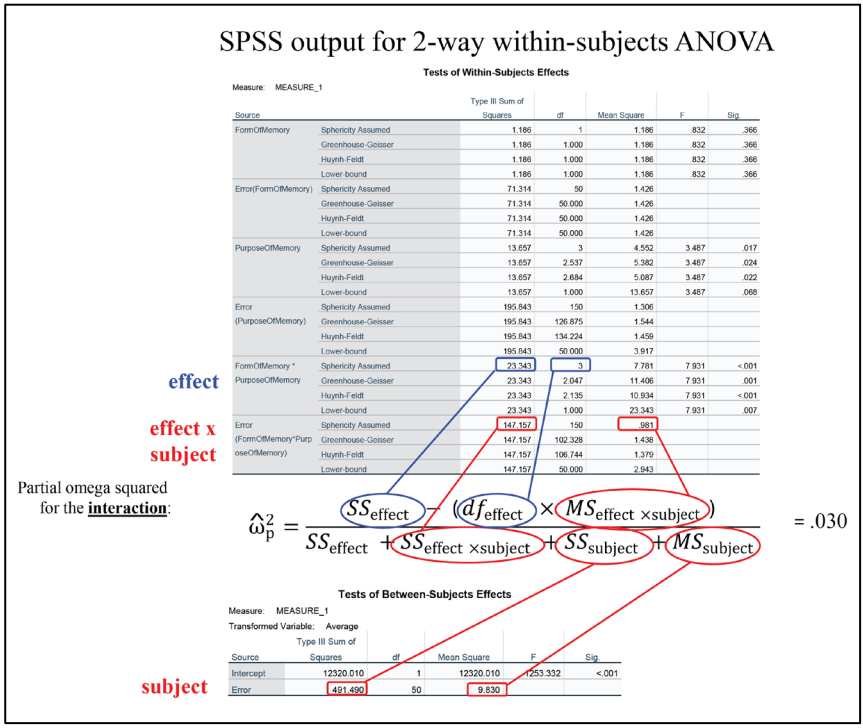
As with the one-way within-subjects design, the systematic differences between subjects are not partialled out, but all other factors are. Because partial eta squared does partial out the variance between subjects, Maxwell et al. (2018) argue that referring to this statistic as partial omega squared may lead to confusion and they refrain from doing so (p. 725). However, we argue that *not* making a distinction between this formula and the formula for standard omega squared, is still confusing. As we follow the idea that formulas for omega squared and partial omega squared should be comparable across designs, it is justifiable to define this formula as partial omega squared.

For between-subjects designs the formula for omega squared can be rewritten as a function of the *F* test statistic, *df*, and *N* (Formulas 14 and 15). This is possible for between-subjects designs because omega squared

is calculated with variance components from the effect and from subject error, both of which are also part of the calculation of *F*.

Figure 2.4

The Components for Partial Omega Squared in SPSS Output for a Two-Way Within-Subjects Design.



Note. Data are from Finley and Naaz (2023), available in the supplemental materials. SPSS version 28.

But rewriting the formula in such a way is impossible for designs that include within-subjects factors. For designs including any within-subjects factors (including split-plot designs), *F* is calculated without the variance due to subjects. It only includes the variance due to the interaction between the effect and the subjects ($\bar{A}$×S). As our formulas for omega

squared include the variance due to subjects, omega squared cannot be deduced from *F* for designs that include within-subjects factors. Some authors do suggest formulas to calculate omega squared from *F*. For the one-way within-subjects design, Keppel and Wickens (2004, p. 362) suggests using the same formula as for between-subjects designs (Formula 14 in this paper). As this excludes the variance due to subjects, this yields a higher value for omega squared than the formulas we have presented. For multifactor-designs, Keppel and Wickens (2004) present two formulas based on *F* to calculate a range in which omega squared falls. This often results in a rather large range, which is "too broad to be useful" (p. 427). It should be noted that the formula that they propose for the lower end of the range still yields a higher value than the formula we have presented, while the higher end of the range yields a lower value than partial eta squared. The bottom line is that there is no certain way to calculate omega squared from *F* for any within-subjects design.

Split-Plot Designs

A split-plot design is one in which there is at least one between-subjects factor and at least one within-subjects factor. These designs do include estimates of random error due to the within-subjects factor ($\widehat{\sigma}^2_{\text{effect} \times \text{subject}}$), as well as an error term that includes systematic differences due to the between-subjects factor ($\widehat{\sigma}^2_{\text{error}}$). For standard omega squared, the formula is as follows (Dodd & Schultz, 1973):

$$\widehat{\omega}^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{appropriate term}})}{SS_{\text{total}} + MS_{\text{subject/A}}} \tag{21}$$

The appropriate error term for the numerator ($MS_{\text{appropriate term}}$) depends on the factor of interest. To clarify, we will use an example with between-subjects Factor A and within-subjects Factor $\overline{\text{B}}$. For the main effect of Factor A the appropriate term is $MS_{\text{subject/A}}$; for the main effect of Factor $\overline{\text{B}}$ and for the interaction A$\overline{\text{B}}$ the appropriate error term is

$MS_{\text{B}\times\text{subject/A}}$ (Maxwell et al., 2018). For calculating standard omega squared in split-plot designs with more factors (and thus with more possible interactions), the appropriate error term for main effects and interaction effects of between-subjects factors will always be $MS_{\text{subject/X}}$, where X stands for all the between-subjects factors in the design. When the effect of interest includes one or more within-subjects factors, the error term will be based on that factor(s). For example in an $AB\bar{C}S$ model, when investigating the $A\bar{\text{C}}$ interaction, the appropriate error term is $MS_{\text{C}\times\text{subject/AB}}$. In an $\text{A}\bar{\text{B}}\bar{\text{C}}\text{S}$ model, when investigating the $\text{A}\bar{\text{B}}\bar{\text{C}}$ interaction, the appropriate error term is $MS_{\text{BC}\times\text{subject/A}}$.

When calculating partial omega squared for a split-plot design the formula also depends on what type of factor is used. If the effect of interest includes one or more within-subjects factors, we use the following formula (see Figure 2.5a):

$$\widehat{\omega}_{\text{p}}^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{effect}\times\text{subject/A}})}{SS_{\text{effect}} + SS_{\text{effect}\times\text{subject/A}} + SS_{\text{subject/A}} + MS_{\text{subject/A}}} \tag{22}$$

Note that even an interaction between a within-subjects factor and a between-subjects factor should use the within-subjects formula.

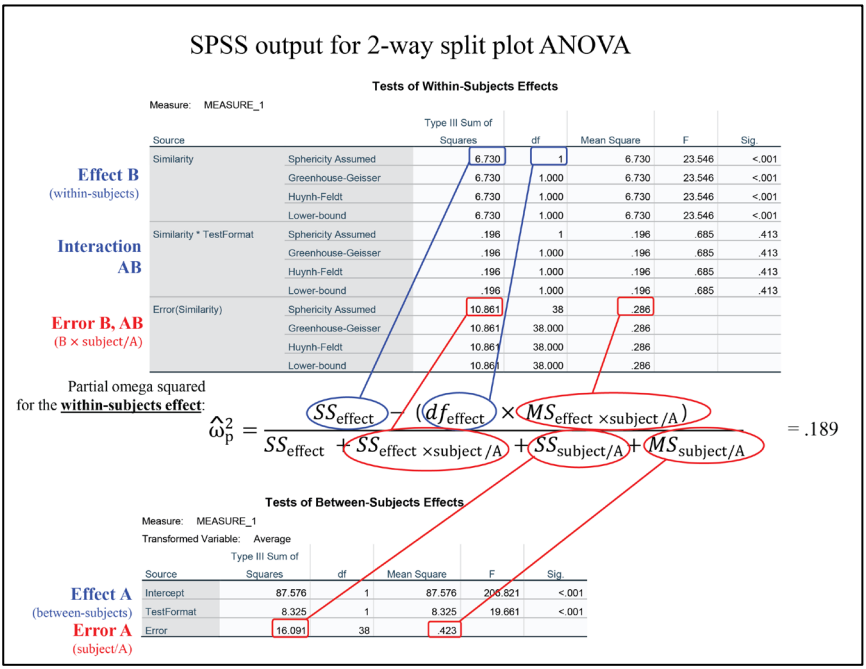
If the effect of interest concerns only between-subjects factors, we use the following formula (see Figure 2.5b):

$$\widehat{\omega}_{\text{p}}^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{subject/A}})}{SS_{\text{effect}} + SS_{\text{subject/A}} + MS_{\text{subject/A}}} \tag{23}$$

See Figures 2.5a and 2.5b for an example of SPSS output from a two-way split-plot ANOVA, showing the components used for partial omega squared for the within-subjects factor (2.5a), and for the between-subjects factor (2.5b).

Figure 2.5a

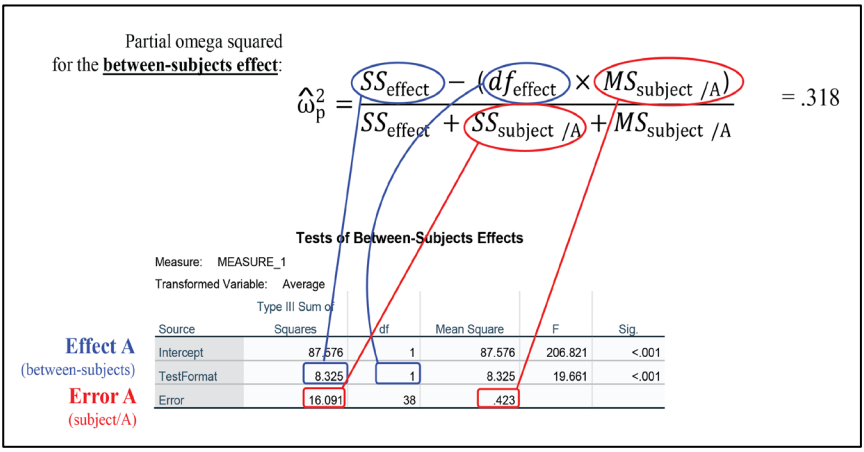
The components for partial omega squared in SPSS output for the effect of the within-subjects factor in a two-way split-plot design.



Note. Data are from Finley et al. (2015), available in the supplemental materials. SPSS version 28.

Figure 2.5b

The components for partial omega squared in SPSS output for the effect of the between-subjects factor in a two-way split-plot design.



Note. Data are from Finley et al. (2015), available in the supplemental materials. SPSS version 28.

In theory, Formulas 21, 22 and 23 can be used for split-plot designs with any number of factors, as long as the right error terms are selected. Olejnik and Algina (2000) present other formulas for split-plot designs (Table 18, pp. 278-279). For the effects that include within-subjects factors, the systematic differences are partialled out. However, readers should use those formulas with caution, as they apparently contain mistakes for the within-subjects factors; the examples provided by Olejnik and Algina in Table 18 of their paper are calculated differently from what the formulas prescribe.

Alternatives for Dealing with Non-Additivity

We have presented formulas in which the variability due to subjects is not partialled out. Alternatives to this method exist. Some authors have proposed ways to calculate a possible range for omega squared (Myers & Well, 2003; Tabachnick & Fidell, 2007a). Other authors propose specific

formulas in which the denominator is increased by adding an error term for each non-additive effect. Interested readers can find these formulas for one, two and three-way designs in Dodd and Schultz (1973); and Vaughan and Corballis (1969). It should be noted that Vaughan and Corballis exclude SS_{subjects} from SS_{total}, while Dodd and Schultz do not specify whether they include SS_{subjects} in SS_{total}. We argue that by not partialling out the variance caused by systematic differences between subjects, the problem of non-additivity is avoided altogether. The papers that provide formulas that correct for non-additivity do not provide formulas for partial omega squared (Dodd & Schultz, 1973; Vaughan & Corballis, 1969). Following the logic that the partial effect size should be equivalent to the effect size that would be obtained in a one-way design (Keppel & Wickens, 2004; Maxwell et al., 2018), one can use the correction provided by Dodd & Schultz (1973) for the one way design by adding $(N \times MS_{\text{effect}\times\text{subject}})$ to the denominator:

$$\widehat{\omega}_{\text{p}}^2 = \frac{df_{\text{effect}}(MS_{\text{effect}} - MS_{\text{error}})}{SS_{\text{effect}} + SS_{\text{effect}\times\text{subject}} + SS_{\text{subject}} + MS_{\text{subject}} + (N \times MS_{\text{effect}\times\text{subject}})} \quad (24)$$

Another solution to the problem of non-additivity is to investigate the contrasts for each level of the within-subjects factor of interest (Boik, 1981). When investigating the 1 *df* contrasts, the assumptions of additivity and sphericity are no longer in effect. Lastly, it is important to note that using (partial) eta squared does not circumvent the problems caused by non-additivity. Eta squared "treats the ambiguity in the error variances in a way that gives it the largest possible value" (Keppel & Wickens, p.428), giving a result that is positively biased.

Remaining Obstacles in Using Omega Squared

Partialling Out Factors

Partial eta squared is the most popular effect size (Alhija & Levy, 2009; Fritz et al., 2012; Peng et al., 2013; Sun et al., 2006; Zhou & Skidmore, 2017)

and some authors recommend always reporting a partial measure of effect (Keppel, 1991; Keppel & Wickens, 2004). However, partial effect sizes can be misleading and harder to interpret than standard effect sizes (Cohen, 1973; Levine & Hullett, 2002; Olejnik & Algina, 2003). The more independent causes there are for an effect, the smaller the effect size is for any of the individual causes. This is disregarded when using a partial effect size (Levine & Hullett, 2002). Identifying causes in a model can reduce the unexplained variance by adding them as a factor (Tabachnick & Fidell, 2007a). However, it is likely that any factor that is added to an ANOVA design can simultaneously reduce error while instilling additional variance (Levine & Hullett, 2002). This makes it hard to decide whether a factor can be partialled out. To use partial omega squared it is necessary that the other factor(s) in the design can be seen as extrinsic (Maxwell et al., 2018). It has to be reasonable to partial out Factor B when investigating the effect of Factor A. Cohen (1973) advised that manipulated variables and variables that are held constant can usually be safely partialled out. When it is likely that an additional factor has an influence on Factor A, the factor is called intrinsic and it should not be excluded when calculating an effect size (Maxwell et al., 2018). When the effect size is inflated because intrinsic factors are partialled out, it is no longer comparable to studies that do not include these factors in their models (Olejnik & Algina, 2003). Therefore, when the other factors are (possibly) intrinsic, it is better to provide a non-partial effect size (Levine & Hullett, 2010). It is of course quite possible that a design contains both extrinsic and intrinsic factors, which complicates the choice between the standard and partial effect size. A solution is offered in generalized omega squared ($\widehat{\omega}_G^2$), which is explained by Olejnik and Algina (2003). They go into many different designs (including random and fixed factors), where some factors are regarded as extrinsic and others as intrinsic. As this results in a myriad of possible combinations, this goes beyond the scope of this paper. However, by sharing the full ANOVA tables, researchers can enable others to calculate generalized omega squared. In their paper on generalized effect

sizes, Olejnik and Algina also describe generalized eta squared ($\hat{\eta}^2_G$). Similar to generalized omega squared, this effect size offers solutions to the problem of extrinsic and intrinsic factors. Additionally, its calculation is less complicated than the calculation of generalized omega squared. However, like its non-generalized counterpart, generalized eta squared is a descriptor of the sample and not of the population, and it does not compensate for the overestimation of the variance due to treatment (Olejnik & Algina, 2003, p. 441).

It can be hard to decide whether a factor should be regarded as intrinsic or extrinsic. Another solution can be to report both partial and standard effect sizes (Cohen, 1973). We recommend reporting explicitly which factors are considered intrinsic and which extrinsic. Again we advise reporting the formulas you use for the effect size, as well as including an ANOVA table.

Confidence Intervals for Omega Squared for Between-Subjects Designs

The APA recommends reporting confidence intervals for statistics, including effect sizes (7th ed., section 3.7; see also Fritz et al., 2012). As with any statistic, the effect size we calculate from our data is an *estimate* of the true effect size in the population. This is why we put the hat symbol over the Greek letter omega. How confident should we be in this estimate of effect size? That is what the confidence interval tells us. The wider the interval, the less precise is our estimate; the narrower the interval, the more precise is our estimate. A 95% confidence interval around a sample statistic means that if we were to endlessly resample the population with the same sample size each time, and put that same confidence interval width around the calculated sample statistic each time, that interval would contain the true population parameter 95% of the time. Unfortunately, determining confidence intervals is not straightforward for effect sizes such as eta squared and omega squared (Fidler & Thompson, 2001; K. Kelley, 2007; Thompson, 2007). In fact, as of this

writing there is no consensus on how to correctly calculate omega squared confidence intervals for within-subjects factors (K. Kelley, personal communication, March 17, 2022). However, it can be done for between-subjects factors.

For statistics such as the mean, a confidence interval can be obtained using a single simple formula. This is not possible for statistics such as eta squared or omega squared. But there is a roundabout way to do it, which we will now explain.

Remember that each effect size is tied to an effect. In ANOVA designs, each effect is tested using an *F* test statistic. *F* is a ratio (e.g., MS_{effect}/MS_{error} in a one-way between-subjects design, also called $MS_{between}/MS_{within}$). When there is no effect, the numerator and denominator are equal, yielding $F = 1$. If there is truly no effect in the population, then we would expect a sampling distribution of *F* values that is centered on $F = 1$, and skewed with a long tail to the right, since a ratio cannot go below 0. This is called the "central" *F* distribution. When we conclude that an effect is statistically significant, we are saying that the *F* test statistic that we obtained from our sample has less than a 5% probability of coming from that central *F* sampling distribution expected from the null hypothesis.

So then, what sampling distribution *did* our *F* statistic most likely come from? A different *F* distribution that is shifted over to the right by a *noncentrality parameter*, which is called lambda (λ). For a particular *F* statistic, lambda can be estimated iteratively by computer software. Software packages essentially use very advanced lookup tables. As there is an enormous number of possible lambda tables, finding the right lambda is virtually impossible without specialized software. Furthermore, the same methods can give us confidence intervals for lambda.[13]

[13] For more detailed discussion of noncentral distributions and their role in determining confidence intervals, see: Cumming and Finch (2001), Smithson (2001), and Steiger and Fouladi (1997).

Why is this relevant to omega squared? Because, for between-subjects designs, omega squared is related to lambda by a simple formula:

$$\omega^2 = \frac{\lambda}{\lambda + N_{\text{total}}} \tag{25}$$

Thus, we can calculate the confidence intervals for lambda, then convert those to the confidence intervals for omega squared.[14]

The steps for this procedure are laid out in Steiger (2004, p. 168), including additional details such as determining when one or both limits of the CI would be set to zero. K. Kelley (2007, section 4) covers similar procedures for CIs for R^2 in a regression context (which is eta squared in an ANOVA context). Unfortunately, as of this writing, such procedures for determining CIs are not easily available in popular statistical software such as SPSS.

Ambitious researchers may refer to Fidler and Thompson (2001, pp. 592-593) and Smithson (2001, Appendix), for syntax-based methods to calculate CIs in SPSS. Related procedures have been implemented in Excel (Cumming & Finch, 2001; J. B. Nelson, 2016), but do not provide a ready-made solution for researchers who just want to determine a CI with minimal struggle.

However, Ken Kelley has written a package of functions that can do what we need, called MBESS (K. Kelley, 2007, 2022), which can be used with the popular free statistical software called R. For the many researchers who are not well-versed in R, its use can be daunting and its documentation cryptic. We have faced our own confusion in using R, so we want to help readers by explaining the exact steps necessary to use R

[14] In addition to the noncentrality parameter approach to constructing confidence intervals, K. Kelley (2005) proposed two bootstrapping approaches, which were further studied by Algina et al. (2006) and Finch and French (2012). We do not address those approaches here.

to determine CIs for omega squared. This same procedure works for partial omega squared.

1. Go to https://cran.r-project.org/ and download R for the appropriate operating system for your computer (e.g., Mac OS, Windows). Use the downloaded file to install R.
2. Open R on your computer. You will see that it is a command line interface, in which you must type specific commands to tell the program what to do. Whenever you type a command, you must press the enter or return key to run the command. There are graphical user interfaces that can be installed for R, such as R Studio and R Commander, but they are not necessary.
3. Enter the following command: `install.packages("MBESS")`
   This command will retrieve the MBESS package of functions from R servers and then install that package. You only need to do this step once; the next time you use R, MBESS will already be installed.
4. Enter the following command: `library("MBESS")`
   This command will load the MBESS package so that its functions are ready to use. You will have to do this step every time you use R in order to use the MBESS package. Optionally, if you would like to read the manual for this package, you can visit https://cran.r-project.org/web/packages/MBESS/index.html or enter the following command in R: `help("MBESS")`
5. You will need to have the following values ready: F for the effect of interest, degrees freedom for the numerator of F (df1, aka df_{effect}), degrees freedom for the denominator of F (df2, aka df_{error}), and the total sample size of the design (N). Note that you do not even need the value of omega squared.

6. Now you will use the function ci.omega2 [added to the MBESS package in version 4.9.0]. Optionally, if you would like to read the manual for this function, enter the command:
   ```
   help(ci.omega2)
   ```
   To actually show you how to run the function, we will use values from the one-way between-subjects design example data (Finley et al., 2017) available in the supplemental materials of this paper (see Appendix D).
7. Enter the following command:
   ```
   ci.omega2(F.value=2.698, df.1=16, df.2=536,
   N=553)
   ```
8. The output gives separate values for the lower and upper limits of the confidence interval, and looks like this:
   ```
   $lower_limit_omega2
   [1] 0.0150022
   $upper_limit_omega2
   [1] 0.09235316
   ```
9. Omega squared from this example was separately calculated as .047 (Figure 2.1). We would report as follows: $\widehat{\omega}^2$ = .047, 95% CI [.0150, .092].

The confidence interval coverage is set to .95 (95%) by default for the `ci.omega2` function. If you wanted to instead calculate a 90% CI, you would add conf.level=.90 so that you would enter the following command:

```
ci.omega2(F.value=2.698, df.1=16, df.2=536,
N=553, conf.level=.90)
```

Note that if a value "NA" is output for one or both limits, just use 0 in place of NA. Here is an example using the main effect of Actual Test Format from our 2-way between-subjects example data (Finley & Benjamin, 2012):

Command:

```
ci.omega2(F.value=31.496, df.1=1, df.2=1, N=100)
```

Output:

```
$lower_limit_omega2
[1] NA
$upper_limit_omega2
[1] 0.6201419
```

Partial omega squared for this effect was separately calculated as .234 (see the online supplemental materials). We would report as follows: $\widehat{\omega}_p^2$ = .234, 95% CI [0, .620].

Confidence Intervals for Designs Including Within-Subjects Factors

When calculating confidence intervals, matters are more complicated for designs that include one or more within-subjects factors. As far as we know, no research into CIs for within-subjects and split-plot designs has been published. Whenever a within-subjects factor is included, omega squared is not directly related to *F*, as the error due to subjects is not part of the formula for *F*. Subsequently, omega squared is also not directly related to lambda. Perhaps future researchers could solve this problem by calculating a correction for lambda, or by developing a formula for omega squared that does derive from *F* even when within-subjects factors are present. Any such work that would provide more clarity on CIs for within-subjects designs would necessitate a Monte Carlo simulation study (K. Kelley, personal communication, March 21, 2022). This goes beyond the scope of this paper.

Quick Guide: Calculating, Reporting, and Interpreting Omega Squared

When reporting omega squared, we advise the following:

1. Use Appendix B to find the appropriate formula for your ANOVA design and the factor of interest. For each component in the formula, retrieve the value from the ANOVA table(s) output by the statistical software you are using. Examples of using SPSS output are shown in Figures 2. 1-5b, along with data and syntax in the supplemental materials. Use a tool such as Excel to perform the actual calculation of the formula using the component values; examples of this are also found in the supplemental materials.
2. Report omega squared with three decimals, as even a value as small as .008 can make a considerable difference in power analysis, especially for smaller effects (Lakens, 2015). Do not include a leading zero (e.g., 0.008), because that is never necessary for statistics that cannot exceed a value of one.
3. It is possible to find a negative effect size for omega squared. This happens when the within group error variance is so high that any treatment effects are either absent or impossible to detect (Keppel & Wickens, 2004). It should be noted that a negative value for omega squared does not indicate a negative effect, but signifies the absence of effect. It could be argued that when omega squared is negative, it should be set to zero. However, for clarity we advise reporting the actual value for omega squared, even if it is negative, accompanied by an interpretation in which the absence of effect is stated (Okada, 2017).
4. Always report the formula used to calculate the effect size, accompanied by a reference to the source of the formula (e.

g., cite this paper; Kroes & Finley, 2023). You can conveniently do this in the beginning of your Results section, or as a footnote.

5. Include a confidence interval for the effect size if possible (i. e., for between-subjects designs).
6. Include the complete ANOVA table(s) in your paper, Appendices, or supplementary materials to enable other researchers to calculate a different effect size when preferred. You may wish to reformat the tables output by SPSS to be consistent with APA style. Note that for within-subjects designs, SPSS outputs the subject factor (labeled "error") in a separate table called "Tests of Between-Subjects Effects" so be sure to include those values.

When interpreting the effect size, some guidelines exist on what constitutes a small, medium, and large effect (Kirk, 1996, p. 751). However, evaluating the effect size in such a manner should be done with much caution. Most importantly, these standardized classifications say little about the actual practical importance and practical significance of an effect (Keppel & Wickens, 2004; Maxwell et al., 2018). A small effect size could be extremely valuable if it, for instance, describes the effect of a life-saving drug. A large effect size could indicate that an effect is already commonly known and therefore not useful to investigate (Keppel & Wickens, 2004). Additionally, effect sizes vary across specific areas of research (Maxwell et al., 2018). What constitutes a groundbreaking discovery in one field, may be inconsequential in another. In most fields it is actually unclear what the smallest effect size of interest is (Lakens, 2022).

Conclusion and Discussion

Reporting measures of effect size has been a growing practice over the past decades, with eta squared and partial eta squared being reported most frequently for ANOVA designs (Alhija & Levy, 2009; Finley et al.,

2017; Fritz et al., 2012; Kirk, 2012; Zhou & Skidmore, 2017). Eta squared and partial eta squared are problematic because they are positively biased and thus tend to overestimate the population value (Albers & Lakens, 2018; Keppel & Wickens, 2004; Keselman, 1975; Lakens, 2015; Levine & Hullett, 2002; Yigit & Mendes, 2018). Omega squared and partial omega squared are less biased (Field, 2017; Keppel & Wickens, 2004; Lakens, 2013, 2015; Tabachnick & Fidell, 2007a; Yigit & Mendes, 2018) but are underused due to inconvenience, lack of guidance and unfamiliarity (Kirk, 1996; Zhou & Skidmore, 2017). There is lack of clarity about the varying formulas for different designs, especially for designs including within-subjects factors (Keppel & Wickens, 2004; Maxwell et al., 2018; Olejnik & Algina, 2000; Tabachnick & Fidell, 2007a).

To help with this lack of clarity we have provided formulas that can be used for between, within and split-plot ANOVA-designs with fixed factors (for a quick overview, see Appendix B). We have provided a guide to calculate confidence intervals for between-subjects designs, and we encourage researchers to investigate confidence intervals for designs that include within-subject factors. Disagreement exists whether variance due to systematic differences between subjects should be partialled out in designs that include within-subjects factors (Maxwell et al., 2018; Tabachnick & Fidell, 2007a). We agree with Maxwell et al. (2018) that it seems inadvisable to disregard variance caused by subjects, and we have presented formulas that include these differences. We argue that this offers an acceptable solution to the problem of non-additivity. Another problem is deciding whether factors in a model should be seen as intrinsic or extrinsic, and whether standard or partial omega squared should be used. A solution could be to present both (Cohen, 1973) or to calculate an alternative, like generalized omega squared (Olejnik & Algina, 2003).

SPSS and other statistical software packages have a great influence on what effect sizes are being used (Fritz et al., 2012; Kirk, 1996; Levine & Hullett, 2002). Statistical software developers can play a helpful role by including more options for omega squared, especially for the fixed

between designs, where there is no controversy over the formulas. Furthermore, the software should state the exact formulas being used. Since the APA started requiring reporting effect size, the use of effect sizes in general has increased considerably (Alhija & Levy, 2009; Finley et al., 2017; Peng et al., 2013; Zhou & Skidmore, 2017). The APA could play a facilitating role by offering more guidance about effect sizes, and by encouraging publishing ANOVA tables. We recommend researchers to always report the formulas used to calculate the reported effect sizes, and if possible to share the ANOVA table(s). This provides others the possibility to partial out specific factors, and it facilitates power calculations and meta-analyses. Most importantly, this aids in cumulative science, as it clarifies the methods used and offers researchers opportunities to make replications and gain deeper understanding of investigated effects.