Moderator Effects Differ on Alternative Effect-Size Measures

Michael Smithson* and Yiyun Shou

Research School of Psychology, The Australian National University

*Corresponding author: Michael.Smithson@anu.edu.au; Research School of Psychology,

Bldg 39, The Australian National University, Canberra, ACT 2601, Australia.

Phone: 61 2 6125 8356

Abstract

This paper discusses largely ignored issues regarding moderation of effect-sizes. We show that, under commonly-occurring conditions, popular alternatives for effect-size measures in ANOVA and multiple regression are not moderated identically across independent samples. Effects may appear to be unmoderated according to one effect-size measure but not according to another, or may even be moderated in opposite directions. We identify the conditions under which differential effect-size moderation can occur, and show that they are commonplace. We then review techniques for detecting and dealing with differential moderation of alternative effect-size measures. Finally, we discuss implications for research practice, reporting, replication, and meta-analysis.

Key Words

effect size, moderator, Cohen’s d, partial correlation, semi-partial correlation, regression coefficient, replication, ANOVA, regression

Introduction

Researchers in psychology nowadays are encouraged to report effect-sizes, replicate studies, and think meta-analytically (e.g., American Psychological Association, 2010, p. 34; Cooper & Patall, 2009; Cumming, 2013; Smithson, 2003, pp. 12-16). These reforms are laudable and long overdue. Nonetheless, they open up some largely undebated issues regarding moderation (and replication) of effect-sizes across samples and across studies. If these issues are ignored, then researchers may fall prey to difficulties in establishing when an effect has been moderated in a single study, when two or more studies can be said to be “replications” of one another, or whether a collection of studies’ effect-sizes is heterogeneous or not.

These difficulties arise because under commonly-occurring conditions, popular alternative effect-size measures in ANOVA and multiple regression are moderated differently across independent samples. Effects may appear to be unmoderated according to one effect-size measure but not according to another, or may even be moderated in opposite directions. Moderator effects are bread-and-butter in many areas of psychology, so differential effect-size moderation has important ramifications for research practice, reporting, replication, and meta-analysis. In this paper we address the following questions:

1. Under what conditions are alternative appropriate effect-size measures moderated differently?
2. How can we detect such differences?
3. How can we interpret differential effect-size moderation?

We begin by observing that when means are compared between two independent samples, Cohen’s d is the conventional effect-size employed, but in principle, either Cohen’s d or η^2 may be used. For a constant total sample size, Cohen’s d is partly determined by the inequality between cell sizes whereas η^2 is not. We describe the conditions, and provide examples, where one measure is moderated but the other is not, and where they are moderated in opposite directions.

We then turn to a well-documented but often ignored distinction between moderation of the “degree” and moderation of the “form” of the relationship between an independent variable (IV) and dependent variable (DV) (Zedeck, 1971; Arnold, 1982; DeShon & Alexander, 1996). In regression, “degree” moderation refers to moderation of association (i.e., correlation) and “form” moderation to moderation of slopes. In ANOVA, “degree” can refer to the differences between means and “form” to Cohen’s d. In another terminology, degree and form correspond to the moderation of standardized (scale-independent) and unstandardized (scale-based) effect-size measures.

Researchers may legitimately be interested only in moderation of degree or in moderation of form, or both. Arnold (1982) refers to debates among industrial and educational psychologists regarding “test fairness”, where “fairness” has at least two meanings. In one sense, a test is fair for all subpopulations (e.g., males and females) if its validity (the correlation between the test score and a criterion variable) is the same for these subpopulations. In another sense, a test is fair if a unit change in the criterion yields the same expected change in test score for all subpopulations. The first sense refers to degree, and the second to form. Arnold’s point is that each of these notions of fairness addresses a different kind of moderation.

However, the default assumption by researchers is that degree and form are moderated in the same way. In linear regression (and ANOVA), homoscedasticity (or homogeneity of variance) guarantees that this will be true. However, heteroscedasticity (or heterogeneity of variance, HeV) forces the two kinds of moderator effects to differ from one another.

Importantly, HeV in the independent variable (IV) can do this as well as HeV in the dependent variable (DV). Researchers seldom test for HeV in the IV, so most probably are unaware of this manifestation of the phenomenon. We describe the conditions under which form is moderated when degree is not (and vice-versa), and when they are moderated in opposite directions. We also discuss the important but often overlooked role of moderated scale reliability in generating HeV. This part of our paper overlaps with Smithson's (2012) treatment. However, that paper restricted its discussion to simple regression and ANOVA, i.e., the moderation of the effect of just one predictor. Our treatment extends the scope to include multiway ANOVA and multiple regression.

In multiway ANOVA and multiple regression, another important consideration about moderation effects needs to be taken into account, namely when a moderator variable affects more than one relationship between variables. We show that three popular alternative effect-size measures, semi-partial η^2 (a.k.a. semi-partial R^2), partial η^2 (a.k.a. partial R^2), and the standardized regression coefficient, may be moderated differently when other moderator effects are present. Importantly, the relevant moderator effects are not limited to moderations of the relationships between other predictors and the DV. Instead, they also include moderation of the relationships between other predictors and the IV whose relationship with the DV is under consideration, i.e., moderation of that IV's tolerance. Again, most researchers seem unaware or heedless of these phenomena. Indeed, to our knowledge, no systematic or comprehensive account of this issue exists in the published literature.

We then briefly review techniques for detecting and dealing with differential moderation of alternative effect-size measures. We reprise Smithson's (2012) approach to dealing with heteroscedasticity effects, and we review methods for evaluating moderation of tolerance and multiple R^2 in regression. Finally, we discuss implications for research practice, reporting, replication, and meta-analysis.

Effect-Sizes in ANOVA and Multiple Regression

ANOVA and linear multiple regression offer researchers a choice among effect-size measures. The most popular effect-size measures in ANOVA are differences between means (in the scale of the raw data), Cohen's d, partial η^2, and semi-partial η^2. The most popular effect-size measures in regression are semi-partial and partial correlations, and unstandardized and standardized regression coefficients. We briefly review these alternative measures and their interrelationships before proceeding to discuss moderator effects.

ANOVA

When means are compared between two independent samples, Cohen's d is the conventional "scale-free" effect-size employed. However, in principle, either Cohen's d or η^2 may be used. The formula linking the two measures can be written as

$$\eta^2 = \frac{t^2}{t^2 + N - 2} = \frac{d^2}{d^2 + 4\left(\bar{n} - 1\right)/\tilde{n}}, \qquad (1)$$

where t is the t-statistic, N is the total sample size, n_1 and n_2 are the number of observations in each sample, $\bar{n}$ is the arithmetic mean of n_1 and n_2, and $\tilde{n} = 2/\left(1/n_1 + 1/n_2\right)$ is their harmonic mean. Alternative forms of this equation are presented by McGrath and Meyer (2006), in their informative discussion of the differences between the correlation coefficient and d. We have chosen this version because of the role played by the ratio between the arithmetic and harmonic means in the right-hand part of equation (1).

In ANOVA for multi-way designs, two popular effect-size measures for main effects are semi-partial and partial η^2 (η_s^2 and η_p^2, respectively). These are identical to squared semi-partial and partial correlations in linear regression (see below). Useful formulas for the η^2 measures are as follows:

$$\eta_s^2 = \frac{SS_j}{SS_e + \sum_i SS_i} = \frac{SS_j}{SS_T}$$
$$\eta_p^2 = \frac{SS_j}{SS_e + SS_j} = \frac{SS_j}{SS_T - \sum_{i \neq j} SS_i} \qquad (2)$$

The SS_j term is the sum of squares for the j^{th} effect, SS_T is the total sum of squares, and SS_e is the error sum of squares. The pairs of formulas suggest two ways of understanding the difference between η_s^2 and η_p^2.

The middle pair of expressions in equations (2) is the “ANOVA” view, in which η_s^2 measures SS_j against the sums of squares for all effects plus SS_e, whereas η_p^2 measures SS_j against itself plus SS_e. Some methodologists (e.g., Tabachnick & Fidell, 2013, pp. 54-55) claim that η_s^2 is “flawed” because the j^{th} effect-size may appear smaller in more complex designs with more effects. In any case, it is best to consider η_s^2 and η_p^2 as addressing different questions about effects.

The right-hand pair of expressions in equation (2), with the SS_T terms, is what might be called the “regression” view (e.g., Tabachnick & Fidell, 2013, p. 145). Here, η_s^2 is viewed as the unique proportion of total variance explained by the j^{th} effect, while η_p^2 is the proportion of the variance left over, after the other effects have contributed their shares, explained by the j^{th} effect. A convenient summary of this distinction is

$$\eta_p^2 = \frac{\eta_s^2}{1 - \eta_{s(j)}^2}, \qquad (3)$$

where $\eta^2_{s(j)}$ denotes the proportion of variance explained by all of the effects except for the j^{th} effect. Equation (3) also makes clear the well-known inequality that $\eta_s^2 \leq \eta_p^2$.

Multiple Regression

In multiple regression, in addition to squared semi-partial and partial correlations, standardized regression coefficients often are used as indicators of the relative importance of predictors. This practice has long attracted criticism (e.g., Budescu, 1993). However, we include it in our discussion here, both because of its popularity and because it can be interpreted as an effect-size measure directly related to semi-partial correlations (e.g., Darlington, 1990, p.58).

To fix ideas, we need some alternative and additional notation. We will consider the effect of a predictor, X, on a dependent variable, Y, in the j^{th} sample (for $j = 1, 2, \ldots, J$). Let A denote the collection of all predictors other than X in the regression model. Let β_{xj}, R_{sxj} and R_{pxj} be the standardized regression coefficient, semi-partial and partial correlation (respectively) for X in the j^{th} sample. Finally, let R_{Ayj} denote the multiple correlation coefficient for the linear regression model predicting Y and containing all of the predictors in set A in the j^{th} sample, and R_{Axj} denote the multiple correlation coefficient for the linear regression model predicting X from all of the predictors in set A in the j^{th} sample.

The relationships among the standardized regression coefficient, semi-partial correlation, and partial correlation may be expressed as follows. First, we may rewrite equation (3) as

$$R_{pxj} = \frac{R_{sxj}}{\sqrt{1 - R_{Ayj}^2}}. \qquad (4)$$

It is also pertinent that the semi-partial correlation for a predictor is the correlation between the dependent variable and the residual of the predictor from a regression model predicting it from the other predictors in the model. The partial correlation, on the other hand, is the correlation between the residuals of both the dependent variable and the predictor, i.e., with the other predictors partialled out from both variables. In some statistical packages (e.g., SPSS), the *F*-test of significance which is based on the squared partial correlation is confusingly paired with output that reports the squared semi-partial correlation.

Second, the standardized regression coefficient is a function of the semi-partial correlation and “tolerance”, $1-R_{Axj}^{2}$, i.e.:

$$\beta_{xj} = \frac{R_{sxj}}{\sqrt{1-R_{Axj}^{2}}}. \tag{5}$$

Thus, as suggested earlier, the standardized regression coefficient also is an effect-size measure. It compares the semi-partial correlation for a predictor against the variation in that predictor that remains unexplained by the other predictors. The appropriate substitution from equation (4) yields the following relationship between the standardized regression coefficient and partial correlation:

$$\beta_{xj} = R_{pxj}\frac{\sqrt{1-R_{Ayj}^{2}}}{\sqrt{1-R_{Axj}^{2}}}. \tag{6}$$

As in the preceding material on ANOVA, it should be clear that these three alternative effect-sizes measure “effect-size” in ways that address different research questions.

Moderation of Alternative Effect-Sizes

Cohen’s d versus Partial η^{2}

Let us first examine the impact of unequal cell sizes on moderation of Cohen’s d versus $\eta_{p}{}^{2}$. For two independent samples, suppose that d is identical for both (i.e., unmoderated). When will the same be true of $\eta_{p}{}^{2}$? Equation (1) can be rewritten as

$$d = \frac{\eta_{p}}{\sqrt{1-\eta_{p}^{2}}} \times \sqrt{\frac{4(\bar{n}-1)}{\tilde{n}}}. \tag{7}$$

Thus, non-identical moderation of these two effect-size measures occurs when the $\left(\bar{n}-1\right)/\tilde{n}$ ratio differs between the samples. For example, suppose that sample 1 has n_{11} = n_{12} = 100 whereas sample 2 has n_{21} = 185 and n_{22} = 15. Then if $d_1 = d_2 = 0.9$, for sample 1 $\eta_{\mathrm{p}1}$ = .412 whereas for sample 2 $\eta_{\mathrm{p}2}$ = .232, so that the "effect-size" is moderated if we use partial eta but not if we use Cohen's d. Also, equation (7) implies that for constant d, the magnitude of η_{p} covaries negatively with the $\left(\bar{n}-1\right)/\tilde{n}$ ratio. Given constant total sample size, this ratio increases as sample sizes become more unequal.

McGrath and Meyer (2006) discuss the difference between the correlation and *d* from a somewhat different standpoint, characterizing unequal sample sizes as differing "base rates". Their conclusions parallel ours, although they do not discuss moderation per se. As they point out, base-rate sensitivity implies that for *d* power is influenced by inequality in sample sizes, whereas for η_p it is not. Equation (7) reveals the observation made by Rosnow, Rosenthal, and Rubin (2000) that power is inversely related to the $\left(\bar{n}-1\right)/\tilde{n}$ ratio.

Can Cohen's d and η_p be moderated in opposite directions? Let $\left(\bar{n}-1\right)/\tilde{n}$ be denoted by Q, and suppose that for sample 1 this ratio is Q, while for sample 2 the ratio is kQ, where k > 1. Now suppose that for sample 1 partial η^2 is $\eta_{\mathrm{p}1}{}^2$, whereas for sample 2 it is $\eta_{p2}{}^2 = c\eta_{p1}{}^2$, where c < 1 so that $\eta_{p1} < \eta_{p2}$. Then a straightforward algebraic argument shows that $d_1 > d_2$ iff $\left(kc-1\right)/\left(kc-c\right) > \eta_p^2$, which in turn requires that $kc > 1$. These conditions are by no means bizarre. For instance, suppose that sample 1 has n_{11} = n_{12} = 25 whereas sample 2 has n_{21} = 40 and n_{22} = 10, so that $Q_1 = 0.96$ and $Q_2 = 1.5$. Suppose also that for sample 1 $\eta_{p1}^2 = .33$ whereas for sample 2 $\eta_{p2}^2 = .25$. Then it follows that $k = 1.563$ and $c = 0.758$, so

$(kc-1)/(kc-c) = 0.431 > \eta_{p1}^2$, and therefore d_1 = 1.375 whereas d_2 = 1.414. Thus, Cohen's d and $\eta_p{}^2$ are moderated in opposite directions.

Form versus Degree Moderation

Suppose a linear relationship between two continuous random variables X and Y is moderated by a third variable, Z. The extent to which the correlation ρ is moderated by Z (moderation of degree) is equivalent to the extent to which the regression coefficients b_y and b_x are moderated by Z (moderation of form) iff the variance ratio σ_y^2 / σ_x^2 is constant over the range or states of Z. The same holds for moderation of the difference between means versus moderation of Cohen's d. Otherwise, moderation of slopes and of correlations (or of mean differences and Cohen's d) must diverge. Most of the literature on this issue focuses on tests for heterogeneity of variance (HeV) in Y, despite the fact that HeV in X also can render that variance ratio non-constant.

Consider the simplest case, where Z is a binary variable. For the i^{th} category of Z,

$$b_{yi} = \rho_i \frac{\sigma_{yi}}{\sigma_{xi}}. \tag{8}$$

A straightforward argument shows that if the $\sigma_{yi} / \sigma_{xi}$ ratio is not constant for $i = 1$ and $i = 2$ then $b_1 = b_2 \Rightarrow \rho_1 \neq \rho_2$, and likewise $b_1 \neq b_2 \Rightarrow \rho_1 = \rho_2$. More generally,

$$\frac{\sigma_{y1}\sigma_{x2}}{\sigma_{x1}\sigma_{y2}} > (<) 1 \Leftrightarrow \left|\frac{b_1}{b_2}\right| > (<) \left|\frac{\rho_1}{\rho_2}\right|. \tag{9}$$

The condition for correlations and slopes to be moderated in opposite directions follows immediately: We have $b_1 > b_2$ whereas $\rho_2 > \rho_1$ if, when $\rho_2 > \rho_1$, it is also true that

$$\frac{\sigma_{y1}\sigma_{x2}}{\sigma_{x1}\sigma_{y2}} > \frac{\rho_2}{\rho_1}. \tag{10}$$

The same implication holds if the inequalities are changed from > to <. Smithson (2012) argues that this condition is not unusual or extreme, and of course violations of homoscedasticity frequently occur in real data.

These results generalize to multiple regression, so that standardized and unstandardized regression coefficients may be moderated differently when the $\sigma_{yi} / \sigma_{xi}$ is not constant, because equation (8) becomes

$$b_{yi} = \beta_i \frac{\sigma_{yi}}{\sigma_{xi}}, \qquad (11)$$

where is the standardized regression coefficient.

Moderation of Reliability

It is common knowledge that the value of a sample correlation is influenced not only by the true population correlation value but also the reliability of the scales measuring the correlated constructs. Hunter and Schmidt's (1990) treatment of meta-analysis using correlation coefficients highlights this issue, but it is routinely ignored when researchers consider moderator effects. It is plausible that under many circumstances, scale reliability may be moderated. If so, then that may introduce artefacts into the assessment of moderator effects on correlation coefficients and other effect-size measures that are functions of correlations, such as Cohen's d and regression coefficients.

The observed squared correlation, $\tilde{\rho}^2$, is the product of the true squared correlation and the reliabilities of the scales being correlated:

$$\tilde{\rho}^2 = \rho^2 \rho_x \rho_y. \qquad (12)$$

Clearly, identical correlations in two samples may appear to be moderated because the reliabilities of one or both scales differ between the samples. It also is possible for the true correlation to be moderated in the opposite direction to the observed correlation. Letting

$C = \rho_x \rho_y$, if we have $C_2 = kC_1$, for $k > 1$, and $\rho_2^2 = c\rho_1^2$, for $c < 1$, then $\tilde{\rho}_2^2 > \tilde{\rho}_1^2$ iff $kc > 1$. Suppose, for instance, that for sample 1 $\rho_1^2 = .33 > \rho_2^2 = .25$, whereas the reliabilities for the scales in sample 1 both are .7 and in sample 2 both are .9. Then c = 0.758 and k = 1.653, so kc = 1.252 and thus $\tilde{\rho}_1^2 = .162 < \tilde{\rho}_2^2 = .203$, i.e., moderation in the opposite direction to that for the true correlations. We note in passing that researchers typically use Cronbach's alpha as a lower bound estimate of population reliability, despite the fact that other reliability estimates are arguably more accurate and useful than alpha (Dunn, Baguley, & Brunsden, 2014; Sijtsma, 2009)..

Semi-Partial versus Partial Correlations versus Standardized Regression Coefficient

We now turn to the three effect-size measures available in regression, two of which are also employed in ANOVA. We first need to establish when these effect-size measures have been moderated identically. It should be evident from equations (3), (4), and (5) that the best way to assess moderation of these parameters between independent samples is via their ratios rather than their differences. From equation (3) we have

$$\eta_{p1} - \eta_{p2} = \frac{\eta_{s1}}{\sqrt{1-\eta_{s(j)1}^2}} - \frac{\eta_{s2}}{\sqrt{1-\eta_{s(j)2}^2}}. \tag{13}$$

Even if $\eta_{s(j)1}^2 = \eta_{s(j)2}^2$, when they are not 0 then it still is the case that $\eta_{p1} - \eta_{p2} \neq \eta_{s1} - \eta_{s2}$ unless $\eta_{p1} - \eta_{p2} = 0$. On the other hand, from equation (3) if $\eta^2_{s(j)1} = \eta^2_{s(j)2}$ then $\eta_{s1}/\eta_{s2} = \eta_{p1}/\eta_{p2}$. Equivalently, from equation (4) if $R_{Ay1} = R_{Ay2}$ then $R_{sx1}/R_{sx2} = R_{px1}/R_{px2}$. Finally, from equation (5), if $R_{Ax1} = R_{Ax2}$ then $R_{sx1}/R_{sx2} = \beta_{x1}/\beta_{x2}$. In this paper, we therefore operationalize "identical moderation" of two effect-size measures across two samples as equal ratios for both parameters. Thus, for instance, $R_{sx1}/R_{sx2} = \beta_{x1}/\beta_{x2}$ is taken to mean that the semi-partial correlation and standardized regression coefficient have been identically moderated across samples 1 and 2.

Ratio comparisons provide practical guidelines for judging when effect-sizes of these kinds have been moderated identically (or replicated) between studies. For the moment, suppose we have an agreed-upon criterion for deciding when each of these effect-size measures has been moderated or not (be it a traditional significance test for their difference, an appropriate Bayes factor, or some other alternative). Then the following three propositions hold.

1. If the multiple correlations R_{Ay1} and R_{Ay2} are unmoderated ($R_{Ay1} = R_{Ay2}$) then partial and semi-partial correlations are moderated identically, whereas the corresponding standardized regression coefficients may be moderated differently.
2. If the multiple correlations R_{Ax1} and R_{Ax2} are unmoderated ($R_{Ax1} = R_{Ax2}$) then semi-partial correlations and standardized regression coefficients will be moderated identically, but partial correlations may be moderated differently.
3. If both pairs of multiple correlations are moderated, all three effect-size measures are moderated differently from one another.

How likely is differential moderation of these alternative effect-size measures? Partial and semi-partial η^2 are very likely to be moderated differently across samples. Equations (4) and (5) reveal that, for any two experiments with identical designs, if $\eta_{s1}^2 = \eta_{s2}^2$ then $\eta_{p1}^2 = \eta_{p2}^2$ if and only if $\sum_i SS_{i1} = \sum_i SS_{i2}$, and vice-versa. This strong constraint is seldom likely to be realized in research, even in carefully controlled experiments. A similar argument follows regarding the differential moderation of partial and semi-partial correlations for non-experimental studies involving multiple covariates.

Somewhat ironically, the magnitude of the differential moderation of alternative effect-sizes may increase with better multivariate models. That is, the larger the squared multiple correlation coefficients, the larger the discrepancy between effect-size measures. In two studies with the same multiple regression model, suppose that one predictor has R_{sx1}^{2} = R_{sx2}^{2} = .1 in both studies, so that the semi-partial correlations are perfect replicates, i.e., unmoderated. Suppose that for the other predictors in the model, R_{Ay1}^{2} = .2 and R_{Ay2}^{2} = .5. Then R_{px1}^{2} = .112 and R_{px2}^{2} = .141, so R_{px2}^{2}/R_{px1}^{2} = 1.265 which indicates moderation of R_{px}. But now suppose R_{Ay1}^{2} = .5 and R_{Ay2}^{2} = .8, so that the difference between them is the same as before but the model fits the data much better. Then R_{px1}^{2} = .141 and R_{px2}^{2} = .224, so R_{px2}^{2}/R_{px1}^{2} = 1.581, a greater degree of moderation of the partial correlations. Finally, suppose that the *ratio*, R_{Ay1}^{2}/R_{Ay2}^{2}, remains the same, with R_{Ay1}^{2} = .32 and R_{Ay2}^{2} = .8. Then R_{px1}^{2} = .121 and R_{px2}^{2} = .224, so R_{px2}^{2}/R_{px1}^{2} = 1.844, an even greater moderator effect.

As we have seen earlier in comparisons between alternative effect-size measures, it is possible for moderation to run in opposite directions for these alternative measures. Suppose that for two independent samples, $\eta_{s1} < \eta_{s2}$ so that $\eta_{s1}/\eta_{s2} = w < 1$. Then from equation (3), if η_{s1} and η_{s2} have the same sign, $\eta_{p1}^{2} > \eta_{p2}^{2}$ when $\sqrt{1-\eta_{s(j)1}^{2}} \Big/ \sqrt{1-\eta_{s(j)2}^{2}} < w.$ Similarly, from equation (5) it is clear that the semi-partial correlation and standardized regression coefficients can be moderated in opposite directions. Suppose that we have two independent samples with multiple regression models containing the same predictors, and $\beta_{x1}/\beta_{x2} = w < 1$, so that $\beta_{x1} < \beta_{x2}$. Then if β_{x1} and β_{x2} have the same sign, $R_{sx1} > R_{sx2}$ when $\sqrt{1-R_{Ax1}^{2}} \Big/ \sqrt{1-R_{Ax2}^{2}} < w$. Both of these reversals require that the “other” predictors’ effects are moderated in the opposite direction from the predictor whose effect’s moderation we are investigating. That is, in the first case, where $\eta_{s1} < \eta_{s2}$, we require that $\eta_{s(j)2}^{2} < \eta_{s(j)1}^{2}$. In the second case, where $\beta_{x1} < \beta_{x2}$, we require $R^2_{Ax2} < R^2_{Ax1}$. Neither requirement is outlandish, although instances of the first case probably are rarer than instances of the second (which involves two different dependent variables). However, the second case is less likely to be investigated by researchers for the same reason that, as Smithson (2012) observes, researchers seldom concern themselves with heteroscedasticity in a predictor.

The take-home lesson is that in a multiple linear regression model, moderation of alternative effect-size measures for any single predictor is partly determined by what else is being (un)moderated in the model. Replication or moderation of one effect-size measure across samples is no guarantee of replication or identical moderation of an alternative effect-size measure across the same samples.

Detecting and Dealing with Differential Moderator Effects

Cohen's d and Partial η^2

If the $\left(\bar{n}-1\right)/\tilde{n}$ ratio varies across samples, then there are differentially unequal sample sizes, but unfortunately the converse does not hold. For example, two independent samples with cell sizes of {40, 10} and {10, 40} will yield a significant chi-square test for unequal proportions (χ^2(14) = 36.00, p < .0005), but identical ratios, $\left(\bar{n}-1\right)/\tilde{n} = 1.5$. A reasonable procedure is to first test for unequal proportions across studies, and then "align" the highest and lowest cell frequencies and re-test for unequal proportions.

In our earlier example, sample 1 had $n_{11} = n_{12} = 25$ whereas sample 2 had $n_{21} = 40$ and $n_{22} = 10$. Here, there is no need to align the highest and lowest cell frequencies because one pair of them is identical (the test for equal proportions gives χ^2(1) = 9.890, p = .0017). Suppose instead that the first sample had cell sizes n_{11} = 20 and n_{12} = 30. Now the chi-square test yields χ^2(1) = 16.667 (p < .0005). If we align the cells so that we have {30, 20} and {40, 10}, then the chi-square test yields $\chi^2(1) = 4.762$ ($p = .0291$), still significant at the .05 level but reduced due to the alignment of the larger and smaller cell frequencies. Note that we also still observe differential moderation of d and η. As before, $\eta^2_{p1} = .33$ whereas $\eta^2_{p2} = .25$, and now $d_1 = 1.404$, nearly equal to $d_2 = 1.414$.

Can this kind of discrepancy identified in equation (7) occur in a collection of studies? Table 1 shows effect-sizes from Feingold's (1994) meta-analysis of studies comparing male and female samples' means on personality measures, in this case the subset comparing them on measures of assertiveness. Six studies (1, 9, 10, 12, 14, 15) have very unequal sample sizes (n_1 = number of males and n_2 = number of females). Applying the procedure described above, a chi-square test for equal proportions across studies yields $\chi^2(14) = 809.61$ ($p < .0005$) and a chi-square test for "aligned" pairs of sample sizes still is very large, with $\chi^2(14) = 457.25$ ($p < .0005$). We may conclude that the $(\bar{n}-1)/\tilde{n}$ ratios vary across studies, with Study 1 a clear outlier in this regard.

As a result, the unequal sample sizes in the studies result in differential moderation of d and η across the studies. Study 1 has $d = 0.26$, twice that of studies 14 and 15; but the three corresponding η values are almost identical (.069, .060, and .062, respectively). Study 9 also has $d = 0.26$, identical to study 1, but its $\eta = .117$, much larger than study 1. Studies 4 and 5, both with d less than d for study 1, have η values greater than η for study 1, thus showing moderation of the two effect-size measures in opposite directions.

Table 1. Studies with male-female comparisons on assertiveness

Study	n_1	n_2	N	d	η	$\bar{n}$	$\tilde{n}$	$4(\bar{n}-1)/\tilde{n}$
1	1024	86	1110	0.26	0.069	555.00	158.67	13.97
2	55	75	130	0.00	0.000	65.00	63.46	4.03

3	41	50	91	0.90	0.412	45.50	45.05	3.95
4	143	177	320	0.17	0.084	160.00	158.19	4.02
5	60	63	123	0.19	0.095	61.50	61.46	3.94
6	83	120	203	0.23	0.113	101.50	98.13	4.10
7	25	36	61	-0.18	-0.090	30.50	29.51	4.00
8	103	100	203	0.35	0.173	101.50	101.48	3.96
9	45	116	161	0.26	0.117	80.50	64.84	4.90
10	544	335	879	0.04	0.019	439.50	414.65	4.23
11	174	195	369	-0.06	-0.030	184.50	183.90	3.99
12	21	39	60	-0.02	-0.010	30.00	27.30	4.25
13	32	32	64	0.16	0.081	32.00	32.00	3.88
14	50	114	164	0.13	0.060	82.00	69.51	4.66
15	57	109	166	0.13	0.062	83.00	74.86	4.38
Totals	2457	1647	4104					

Differential Form versus Degree Moderation

Smithson (2012) presents a parametric test for the between-sample equality of the variance ratio σ_y^2 / σ_x^2 (EVR) based on the log-likelihood of a bivariate normal distribution for X and Y conditional on a categorical moderator Z, employing submodels for the standard deviations using a log link. He reports evidence supporting the Type I error-rate accuracy of this test and reasonable power for moderate departures from normality in X and Y. He also extends this test to the case where Z is a continuous moderator, along with simulation studies examining its Type I error-rates and power. Scripts for maximum likelihood estimation in R, SPSS and SAS are available via the link provided in Smithson (2012).

For a categorical moderator, Smithson (2012) discusses incorporating the EVR test in a structural equations model (SEM) approach that enables researchers to test simultaneously for EVR, HeV in the IV and DV, homogeneity of error variance, moderation of correlations, and moderation of slopes. He provides examples in two SEM packages that can fit these models: lavaan (Rosseel, 2012) and MPlus (Muthén & Muthén, 2010). Readers may consult Smithson (2012) for further details, examples, and a link to worked examples in both environments.

Multiple Regression and ANOVA: Comparing Squared Multiple Correlations

Because detecting differential moderation of alternative effect-size measures in multiple regression and multi-way ANOVA hinges on detecting the moderation of squared multiple correlations, we require methods for estimating confidence intervals around differences between squared multiple correlations. We survey five methods: Asymptotic, “modified asymptotic”, transformations to normality, bootstrapping, and estimation via structural equations models.

Olkin and Finn (1995) describe asymptotic methods for constructing confidence intervals for the difference between two squared multiple correlations. Briefly, $\left(R_1^2 - R_2^2\right) - \left(\rho_1^2 - \rho_2^2\right) \sim N\left(0, \sigma_\infty^2\right)$, where $\sigma_\infty^2 = \left(4/n_1\right) R_1^2 \left(1 - R_1^2\right)^2 + \left(4/n_2\right) R_2^2 \left(1 - R_2^2\right)^2$, with n_j denoting the sample sizes. However, as Algina and Keselman (1999) observe, this approach does not work well unless sample sizes are very large, and so we do not consider it further here.

Zou (2007) presents a “modified asymptotic” approach to constructing confidence intervals for the difference between two correlations or between two squared correlations. For two independent squared multiple correlations, R_1^2 and R_2^2, his procedure is as follows. First, use a scaled noncentral F approximation to the distribution of the squared multiple correlation to obtain confidence intervals around each of them, [l_1, u_1] and [l_2, u_2], respectively. Then, compute the lower and upper limits of the confidence interval around $R_1{}^2 - R_2{}^2$ by these formulas:

$$\begin{aligned} L &= R_1^2 - R_2^2 - \sqrt{\left(R_1^2 - l_1\right)^2 + \left(u_2 - R_2^2\right)^2} \\ U &= R_1^2 - R_2^2 + \sqrt{\left(R_2^2 - l_2\right)^2 + \left(u_1 - R_1^2\right)^2} \end{aligned}. \tag{14}$$

Zou demonstrates that this approach outperforms asymptotic methods in the accuracy of confidence interval coverage-rates for moderate sample sizes. However, a major limitation of this method is that it does not generalize to more than two samples.

Algina and Keselman (1999) investigated a variance-stabilizing transformation of the squared multiple correlation to normality proposed by Olkin and Finn, reporting minimum sample sizes required for adequately accurate confidence interval coverage-rates under a variety of conditions. The transformation is

$$z = \log\left(\frac{1+\sqrt{R^2}}{1-\sqrt{R^2}}\right) \tag{15}$$

with asymptotic variance 4/*n*. Thus, a confidence interval around the difference between z_1 and z_2 is approximated by $z_1 - z_2 \pm t_{\alpha/2}\sqrt{4/n_1 + 4/n_2}$. This is not the only such transformation (see, e.g., Hodgson, 1968), but in simulations it performs as well as or better than the other proposals (details are available from the first author), so we do not consider the others here.

An advantage of the transformation in equation (15) is that its approximation to the normal distribution allows a generalization to comparisons among more than two squared multiple correlations. An overall measure of the heterogeneity of *K* squared multiple correlations is obtained via the standard chi-square statistic:

$$V = \frac{\sum_{i=1}^{K} \left(z_i - z^+\right)^2 \Big/ \sigma_i^2}{\sum_{i=1}^{K} 1\big/\sigma_i^2}, \tag{16}$$

where z^+ is the weighted mean of the z_i, with weights defined as $1\big/\sigma_i^2$, and $\sigma_i^2 = 4\big/n_i$.

Asymptotically, $z_i - z^+ \sim N\left(0, \sigma_i^2\right)$, so that when the null hypothesis is true, $V \sim \chi^2_{K-1}$. Otherwise, for a fixed-effects model, *V* has a noncentral chi-square distribution. Its noncentrality parameter is the sum of squared standardized effects (Smithson, 2003: 43), and it can be converted to a squared partial correlation coefficient that can be used as an effect-size measure in this context. A confidence interval around the noncentrality parameter therefore can be transformed to a confidence interval around this effect-size. Denoting the noncentrality parameter by ν, the transformation to a squared partial correlation is

$$\eta^2 = \frac{\nu}{\nu + N - 1}, \tag{17}$$

where *N* is the sum of the sample sizes.

Chan (2009) presents a bootstrap method for comparing two squared multiple correlation coefficients. Let $\mathbf{X}$ be a vector of predictors of Y, and suppose there are two independent samples of these variables, S_1 and S_2 with sizes N_1 and N_2, from populations whose squared multiple correlation coefficients are ρ_1^2 and ρ_2^2, respectively. Chan's bootstrap procedure is as follows.

Now suppose we take *B* bootstrap samples. For $b = 1, 2, \ldots, B$:

1. Randomly select N_1 and N_2 cases, $(\mathbf{x}_{i1b}, y_{i1b})$ and $(\mathbf{x}_{j2b}, y_{j2b})$ respectively for $i = 1, 2, \ldots, N_1$ and $j = 1, 2, \ldots, N_2$, with replacement from S_1 and S_2.

2. Compute the predicted values $\hat{y}_{i1b}$ and $\hat{y}_{j2b}$, from these and their sample means compute the sample squared coefficients R^2_{1b} and R^2_{2b}, and then obtain $d_b = R^2_{1b} - R^2_{2b}$.

The bootstrap standard error (BSE) is then

$$\hat{\sigma}_B = \sqrt{\frac{\sum_{b=1}^{B}\left(d_b - \bar{d}\right)^2}{B-1}}. \tag{17}$$

The bootstrap confidence interval (BCI) then is

$$\left[\bar{d} - \hat{\sigma}_B z_{1-\alpha/2}, \bar{d} + \hat{\sigma}_B z_{1-\alpha/2}\right], \tag{18}$$

and the bootstrap percentile interval is the appropriate percentiles of the bootstrap cumulative distribution of the rank-ordered d_b.

Finally, Kwan and Chan (2014) propose a two-stage structural equations model (SEM) approach for comparing squared multiple correlations across groups. Unlike an earlier "phantom variable" SEM method for comparing squared multiple correlations (Cheung, 2009), their approach is not limited to comparing two groups. In the first stage, the original multi-group model is transformed into a model such that the squared multiple correlation coefficient becomes a free model parameter in the transformed model. In the second stage, the squared multiple correlations in the groups are compared by imposing linear between-group constraints on the parameters of interest in the transformed SEM, and model comparisons (e.g., between a null-hypothesis model where the squared correlations are identical versus the alternative model in which they differ) are performed via likelihood ratio tests.

Examples

For illustrative purposes, we present two examples, one using ANOVA and another with multiple regression. For simplicity, we restrict this presentation to three techniques: The

Olkin-Finn transformation to normality, the Zou's confidence intervals, and the Chan's bootstrap. We also do not illustrate form versus degree moderation; for illustrations thereof we refer the reader to Smithson (2012).

Our first example is an artificial 2 x 2 x 2 between-subjects factorial experimental design, with factors A, B, and C, and 20 observations in each cell (data and details of analyses are available from the first author). Table 2 shows the sample sums of squares, partial eη^2, 95% confidence intervals for partial η^2, and semi-partial η^2 values. There is a moderate main effect for factor A, a strong main effect for C, a strong A*C interaction effect, and a strong 3-way interaction effect.

Table 2. Three-way ANOVA example

				95% CI		
Factor	SS	df	η_p^2	lower	upper	η_s^2
A	10.374	1	.074	.008	.179	0.020
B	0.537	1	.004	.000	.057	0.001
C	208.148	1	.616	.503	.692	0.400
A * B	0.355	1	.003	.000	.052	0.001
A * C	66.871	1	.340	.203	.456	0.129
B * C	0.996	1	.008	.000	.068	0.002
A * B * C	103.144	1	.443	.306	.547	0.198
Error	129.943	112				
Total	520.369	119				

Suppose that we wish to interpret the interaction effects by using factor C as a stratifying moderator and computing the resulting simple effects for each panel of C. Table 3 displays the results of these analyses. The ratios of η_p and η_s for factor A are similar, 1.939 and 2.147 respectively, and their ratio is 1.107. However, the ratios for factor B are 0.177 and 0.143 respectively, giving a ratio of 1.238. Likewise, the ratios of η_p and η_s for the A*B effect are 1.011 and 0.830 giving a ratio of 1.218. The ratio of the $\sqrt{1-\eta^2_{s(j)}}$ terms for B is 0.806 and the ratio of the $\sqrt{1-\eta^2_{s(j)}}$ terms for A*B is 0.821. This latter ratio is smaller than the corresponding η_s ratio (0.830), so that η_p and η_s are moderated in opposite directions.

Table 3. Three-way ANOVA simple effects

Factor C: level 1

Factor	SS	df	η_p^2	η_s^2	$\eta^2_{s(j)}$
A	64.961	1	.5365	.3894	.2741
B	0.035	1	.0006	.0002	.6633
A * B	45.695	1	.4488	.2739	.3896
Error	56.128	56			
Total	166.819	59			

Factor C: level 2

Factor	SS	df	η_p^2	η_s^2	$\eta^2_{s(j)}$
A	12.284	1	.1427	.0845	.4079
B	1.498	1	.0199	.0103	.4820
A * B	57.805	1	.4392	.3976	.0948
Error	73.814	56			
Total	145.402	59			

The 95% confidence intervals around the differences between the $\eta^2_{s(j)}$ terms suggest that the semi-partial and partial correlations are moderated differently for the A*B effect.

The Chan 95% BCa bootstrap intervals are [-0.102, 0.340] for factor A, [-0.023, 0.388] for factor B, and [0.047, 0.499] for A*B. The Zou 95% intervals are [-0.098, 0.391] for factor A, [-0.033, 0.385] for factor B, and [0.120, 0.548] for A*B, reasonably similar to the bootstrap results. The Olkin-Finn technique agrees qualitatively with these assessments, yielding 95% confidence intervals of [-0.374, 1.072] for factor A, [-0.155, 1.291] for factor B, and [0.104, 1.550] for A*B. We may conclude that the partial and semi-partial correlations for the A*B effect are moderated differently from each other, with the semi-partial correlations being moderated more strongly and (slightly) in the opposite direction. The squared partial correlations for A*B are quite similar, at .449 and .439 for levels 1 and 2 on factor C, whereas the squared semi-partial correlations differ substantially, at .274 and .398.

Our final example is a multiple regression model with data from a study by Shin (2014), which focuses on risk-taking and psychological resilience. The dependent variable (*Y*) is the score on a risk-taking disposition scale (Blais & Weber, 2006), with predictors consisting of participants' gender (*G*) and two covariates, a measure of psychological resilience (X_1, Smith, et al., 2008) and a measure of ruminative thinking (X_2, Brinker & Dozois, 2009). The model is

$$Y_i^{'} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 G + \beta_4 X_2 G, \qquad (15)$$

so *G* takes the role of moderating the effect of ruminative thinking on risk-taking disposition.

The top part of Table 4 displays the unstandardized regression coefficient estimates and standard errors, and the standardized coefficients for this model. The remaining two parts of Table 4 show the simple-effects regression models for males and females. We now consider whether the partial correlations, semi-partial correlations, or standardized regression coefficients for X_2 have been moderated differently by gender. From Table 4, the standardized regression coefficients are .421 for the males and .193 for the females, and their ratio is 2.186. The corresponding partial correlations turn out to be .417 and .165 and their ratio is 2.529, while the semi-partial correlations are .402 and .160 and their ratio is 2.511, so the moderation effect appears to be stronger for both kinds of correlations than for the standardized regression coefficient.

Table 4. Regression model

				95% CI	for *b*	Partial	Semi-Partial
Covariate	*b*	*s.e*	*β*	lower	upper	Corr.	Corr.
intercept	31.940	12.013					
X_1	10.482	2.001	.369	6.540	14.425		
X_2	0.499	0.122	.521	0.259	0.739		
G	11.494	10.124	.201	-8.451	31.440		
X_2*G	-.305	.138	-.497	-.577	-.033		
Males				95% CI	for *b*		
Covariate	*b*	*s.e*	*β*	lower	upper		
intercept	13.248	20.050					
X_1	14.885	4.152	.393	6.604	23.165	-.415	-.400
X_2	0.544	0.142	.421	0.261	0.827	.417	.402
Females				95% CI	for *b*		
Covariate	*b*	*s.e*	*β*	lower	upper		
intercept	51.859	11.487					
X_1	8.650	2.235	.348	4.237	13.063	-.573	-.557
X_2	0.160	0.075	.193	0.012	0.307	.165	.160

Table 5 shows that the three methods of evaluating the differences between the relevant R_{Axj}^2 pair and between the R_{Ayj}^2 pair agree qualitatively. The confidence intervals for the difference between R_{Ax1}^2 and R_{Ax2}^2 contain only positive values, suggesting that the semi-partial correlation and the standardized regression coefficient are moderated differently. The confidence interval for the difference between R_{Ay1}^2 and R_{Ay2}^2 contains 0, so it is not clear whether the partial and semi-partial correlations are moderated differently from each other (although their ratios are very similar, so they probably are moderated similarly).

Table 5. 95% confidence intervals for differences between R_{Axj}^2 and R_{Ayj}^2 pairs

				Olkin-	Finn	Zou		Bootst.	
	Male	Female	Diff.	Lower	Upper	Lower	Upper	Lower	Upper
R_{Axj}^2	0.090	0.310	-0.220	-1.190	-0.085	-0.397	-0.090	-0.380	-0.042
R_{Ayj}^2	0.071	0.058	0.013	-0.497	0.609	-0.127	0.120	-0.103	0.182

A systematic comparison of alternative methods for detecting differences between squared multiple correlations has yet to be done, and this is an active topic of research. Nevertheless, the state of the art indicates that we have some serviceable methods for this purpose.

Conclusions and Recommendations

The conditions under which differential moderation of alternative effect-size measures can occur are quite likely to crop up in multivariate research. Differential moderation of alternative effect-size measures poses a problem for both meta-analysis and the interpretation of moderator effects within a study. A simple solution would be for all researchers to use just one effect-size measure and ignore the others (the partial correlation in preference to the semi-partial correlation, for example). However, Smithson’s (2012) review of the scattered literature on differential moderation of simple slopes and correlations identified contradictory published advice regarding whether tests of simple slopes should be preferred over tests of correlations or vice-versa. Smithson concludes that a superior approach would be to model both parameters, and the relevant variance ratios, and ascertain when and how these are moderated differently. McGrath and Meyer (2006: 398) provide a similar recommendation regarding the choice between η and Cohen's *d* (also see our summary discussion below).

Likewise, here we argue that a more adaptive response is to recognize that alternative effect-size measures can be moderated differently and to take this into account when addressing questions about moderator effects and/or replications of studies. The keys to doing this reside in recognizing that alternative effect-size measures convey different information about effects, bearing in mind that replication or moderation outcomes depend on the choice of an effect-size measure, undertaking to model more than one effect-size measure, and taking reliability into account where possible. The factors driving divergent moderation and replication outcomes for alternative effect-size measures are unequal sample sizes (or base-rates), moderated scale reliability, heterogeneity of variance, and multiple moderator effects involving the dependent variable and/or its predictors. We will conclude by briefly discussing the implications of each of these for research practice and reporting.

The discrepancy between moderation of d and η is driven by moderation of the ratio $\left(\bar{n}-1\right)/\tilde{n}$. As established by McGrath and Meyer (2006), the choice between d and η revolves around the issue of whether the researcher's purposes are best served by a base-rate sensitive measure (η) or a base-rate insensitive measure (d). If the moderation of $\left(\bar{n}-1\right)/\tilde{n}$ reflects a relevant phenomenon (e.g., different rates of a psychological disorder across subpopulations) then η might be preferred over d, whereas the converse would hold if moderation of $\left(\bar{n}-1\right)/\tilde{n}$ is due to an irrelevant happenstance. Where there are no clearcut reasons for preferring one statistic over the other, reporting both and assessing the moderation of sample sizes would be prudent.

The moderation of scale reliability can affect moderation of both d and η. It therefore stands as a potential explanatory factor for heterogeneity among effect-sizes in meta-analyses as well as among independent samples in the one study. Differential reliability across samples or studies clearly is important, both because of its implications regarding moderation and replication and because it is directly related to issues of measurement invariance.

Heterogeneity of variance drives the discrepancy between the moderation of unstandardized and standardized regression coefficients (or the special case of the simple regression coefficient versus correlation). We will not review the long-running debates regarding unstandardized vs standardized regression coefficients, but note that heterogeneity of variance is an additional factor for researchers to consider where moderation or replication is concerned. Above all, researchers should be aware that both are unlikely to be moderated identically, so a test for one is not a test for the other, and ideally they should examine variance heterogeneity in predictors as well as in the dependent variable. Unlike base-rate sensitivity of d vs η, it is not the case that one statistic is sensitive to variance heterogeneity whereas the other is not; instead both are differentially affected by it.

Finally, in multivariate studies, multiple moderator effects may cause discrepancies between the moderation of partial correlation, semi-partial correlation, and standardized regression coefficients. This is the case for moderator effects on the predictor under consideration as well as the dependent variable. Again, we will not enter debates such as whether to prefer partial over semi-partial correlations, but simply note that if researchers are going to choose just one of them then they should provide a clear rationale for doing so. Ideally, they should also report moderation of the relevant alternative measures when assessing moderator effects. If partial correlations are preferred, they are a function of the semi-partial correlation and R_{Ayj}, so it is wise to consider reporting moderator effects on those two statistics as well. Likewise, if standardized regression coefficients are preferred, then moderator effects on the semi-partial correlation and R_{Axj} would be relevant to report.

At the very least, researchers will be wise to exercise caution regarding claims about effect-size homogeneity or moderation in multivariate studies and meta-analyses, especially where questions of replication arise. Researchers who elect one effect-size measure should provide a rationale for that choice, and make it clear when claims about moderation or replication pertain only to that measure and not to alternative measures. It is essential to avoid the trap of believing that a test for moderation of one measure is a test for all. Ideally, future meta-analyses of multivariate studies should incorporate the techniques described in this paper for identifying and modeling differential moderation of alternative effect-size measures.

References

Algina, J. and Keselman, H.J. (1999). Comparing squared multiple correlation coefficients: Examination of a confidence interval and a test of significance. *Psychological Methods*, 4, 76-83.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Arnold, H. J. (1982). Moderator variables: A clarification of conceptual, analytic, and psychometric issues. *Organizational Behavior and Human Performance, 29*, 143–174. doi:10.1016/0030-5073(82)90254-9.

Blais, A.-R. & Weber, E.U., (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1, 33-47.

Brinker, J. & Dozois, D.J.A. (2009). Ruminative thought style and depressed mood. *Journal of Clinical Psychology*, 65, 1-19.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114, 542–551.

Chan, W. (2009) Bootstrap standard error and confidence intervals for the difference between two squared multiple correlation coefficients. *Educational and Psychological Measurement, 69*, 566-584. doi: 10.1177/0013164408324466.

Cooper, H., & Patall, E. A. (2009). The relative benefits of metaanalysis conducted with individual participant data versus aggregated data. *Psychological Methods*, 14, 165–176.

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*, 25, 7-29.

Darlington, R. (1990). *Regression and linear models*. New York, NY: McGraw-Hill.

DeShon, R. P., & Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods, 1*, 261–277. doi:10.1037/1082-989X.1.3.261.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. British Journal of Psychology, 105, 399–412. doi:10.1111/bjop.12046

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, 116, 429-256.

Hodgson, V. (1968). On the sampling distribution of the multiple correlation coefficient (abstract). *Annals of Mathematical Statistics*, 39, 307.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of Meta-Analysis*. Newbury Park, CA: Sage.

Kwan, J.LY. & Chan, W. (2014). Comparing squared multiple correlation coefficients using structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 21*,225–238. doi:10.1080/10705511.2014.882673.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: the case of r and d. Psychological methods, 11(4), 386.

Muthén,L., and Muthén,B. (2010). *Mplus User's Guide, 6thEdn*. Los Angeles, CA: Muthén and Muthén.

Olkin, I. & Finn, J.D. (1995). Correlations redux. *Psychological Bulletin*, 188, 155-164.

Rosnow, R. L., Rosenthal, R., & Rubin, D. R. (2000). Contrasts and correlations in effect-size estimation. Psychological Science, 11, 446–453.

Rosseel,Y. (2012). *lavaan: Latent Variable Analysis*. R package version 0.4–13. Available at: http://CRAN.R-project.org/package=lavaan.

Shin, J. (2014). [Uncertainty and Psychological Resilience]. Unpublished raw data.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika. 74, 107120. doi:10.1007/s1133600891010

Smith, B.W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard, J. (2008). The brief resilience scale: Assessing the ability to bounce back. *International Journal of Behavioral Medicine*, 15, 194–200.

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.

Smithson, M. (2012). A simple statistic for comparing moderation of slopes and correlations. *Frontiers*, 3, doi=10.3389/fpsyg.2012.00231

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Zedeck, S. (1971). Problems with the use of “moderator” variables. *Psychological Bulletin, 76*, 295–310. doi:10.1037/h0031543.

Zou, G.Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, 12, 399-413.