Are we comparing apples or squared apples?
The proportion of explained variance exaggerates differences between effects

Marco Del Giudice
University of New Mexico

Marco Del Giudice, Department of Psychology, University of New Mexico. Logan Hall, 2001 Redondo Dr. NE, Albuquerque, NM 87131, USA; email: marcodg@unm.edu

In this brief note, I wish to draw attention to a problem that has been discussed a number of times in the literature, but whose implications are still not widely appreciated. As a result, many researchers (present author included) keep making distorted inferences about the relative size and importance of certain effects, by directly comparing the proportions of variance they account for. This can be especially consequential in disciplines where variance components are routinely used as indices of effect size, such as behavior genetics. To illustrate, the narrow-sense heritability of depression (i.e., the variance in the risk for the disorder explained by additive genetic effects) is 30-40%, whereas that of bipolar disorder is at least 60% and perhaps as high as 80% (Johansson et al., 2019; Knopik et al., 2017). These figures seem to indicate that the influence of genetic factors on the risk of developing a condition is approximately twice as large in bipolar disorder as in depression. But this interpretation is incorrect; as I discuss below, genetic factors play a much more similar role in the two disorders than suggested by this comparison.

The proportion of explained variance is a non-intuitive and often misleading index of effect size. Variances are mathematically convenient because they combine additively; however, they are not expressed in the original units of the variable of interest—say income in dollars, intelligence in IQ points, or height in inches—but in *squared* units. Even when these variance units are not meaningless (as with squared dollars or squared IQ points), they still fail to measure the actual trait under consideration (e.g., square inches do not measure a person's height). In contrast, the correlation coefficient—the square root of the explained variance—quantifies the relation between two variables in terms of the (standardized) original units, and thus has a natural interpretation with respect to the size of the effect.[1] If the correlation between X and Y is .30, a change of one standard deviation in X predicts a change of 0.30 standard deviations in Y (and vice versa; here I do not distinguish between statistical prediction and genuine causality). The proportion of the variance of Y accounted for by X is just 9%, which makes the effect seem small and unimportant. But, as noted above, explained variance is expressed in *squared* units of Y, and relates to the real-world effect of X on Y in a highly nonlinear fashion.[2] Over the years, many have noted that the proportion of explained variance can lead researchers to dramatically underestimate the importance of certain effects, and have recommended the use of correlations (or other unsquared indices such as Cohen's d) to quantify and interpret effect sizes (e.g., Abelson, 1985; Beatty, 2002; Breaugh, 2003; D'Andrade & Dart, 1990; Funder & Ozer, 2019; Hunter & Schmidt, 1990, 2014; Rosenthal & Rubin, 1979).

An important corollary, but one that is seldom discussed explicitly, is that *comparing* effects based on their respective proportions of explained variance tends to exaggerate the

[1] In some scenarios, the *un*squared correlation between two variables measures the variance explained by a third variable of interest (see Johnson, 2011; Ozer, 1985). For example, the correlation between monozygotic twins reared apart is a direct estimate of trait heritability (i.e., the proportion of variance explained by additive genetic factors); the correlation between two parallel forms of a scale is a direct estimate of their reliability (i.e., the proportion of variance explained by the latent construct). In these scenarios, the effect of interest is *not* the association between the two measured variables, but that between each of them and a third, unobserved variable (the genetic factor; the latent construct). As usual, the effect of interest is quantified by the square root of the proportion of explained variance—in this case, the square root of the heritability or reliability.

[2] Throughout the paper, I use "real-world" as a shorthand for effects expressed in the original (standardized) units of the relevant variables, as contrasted with the squared units of variance.

differences among them—often by a large margin (Hunter & Schmidt, 1990, 2014). Consider a variable *Z* that correlates .60 with *Y*. A change of one standard deviation in *Z* predicts a change of 0.60 standard deviations in *Y*. That is, a given change in *Z* has twice the effect on *Y* than the same amount of change in *X*. But the variance explained by *Z* (36%) is *four times* as large as that explained by *X*—a ratio that grossly exaggerates the real-world difference between the respective effects of *X* and *Z* on *Y*, expressed in the scale of the original units of these variables.

More generally, the ratio between two correlations (henceforth the "effect ratio") is simply the square root of the ratio between the corresponding squared correlations (i.e., the proportions of explained variance). Of course, it is not always sensible to compare two standardized effect sizes, and—depending on context—*un*standardized effects can be more informative than standardized ones. But when it makes sense to compare proportions of explained variance in the context of continuous variables, the effect ratio provides a much more realistic index of the relative importance of the effects, which usually concerns the original units of the variables rather than the squared units of the variance. However, note that the ratio between correlations is not the same as the ratio between values of Cohen's *d*, because *d* is nonlinearly related to the correlation coefficient. Thus, when the focus of the analysis is the difference between two groups (i.e., the proportion of the total variance accounted for by a binary group variable, such as males vs. females), the simple effect ratio described here does not correspond to the ratio between *d* values, except in special cases.[3] In the rest of this paper, I only consider examples in which the relevant variables are continuous and the correlation coefficient is a natural effect size.

Going back to the case of depression and bipolar disorder, the ratio of the heritabilities of these disorders is about two; the square root of this ratio is about 1.41, meaning that genetic factors contribute about 40% more to the risk of bipolar disorder compared with that of depression (instead of twice as much, as suggested by the heritabilities). Indeed, one standard deviation increase in the genetic predisposition for bipolar disorder increases risk by $\sqrt{.60} \approx 0.77$ standard deviations, whereas one standard deviation increase in the genetic predisposition for depression increases risk by $\sqrt{.30} \approx 0.55$ standard deviations.

For another example, consider this quote from Plomin and von Stumm (2018): "One of the most interesting developmental findings about intelligence is that its heritability as estimated in twin studies increases dramatically from infancy (20%) to childhood (40%) to adulthood (60%)" (p. 152). Although the heritability increases threefold, the real-world impact of genetic factors on intelligence is only about 70% larger in adulthood than in infancy ($\sqrt{3} \approx 1.73$). Specifically, one standard deviation increase in the genetic score for intelligence can be expected to increase intelligence by $\sqrt{.20} \approx 0.45$ standard deviations in infancy, $\sqrt{.40} \approx 0.63$ standard deviations in childhood, and $\sqrt{.60} \approx 0.77$ standard deviations in adulthood. In the same paper, the authors predicted that genomewide polygenic scores "will explain substantially more than 10% of the variance in intelligence, which is more than 20% of the 50% heritability of intelligence", and commented "Nonetheless, 10% is a long way from the heritability estimate of

[3] For equal-sized groups, the conversion is $d = \frac{2r}{\sqrt{1-r^2}}$. Hence, the ratio between two *d* values is $\frac{d_1}{d_2} = \frac{r_1}{r_2}\sqrt{\frac{1-r_2^2}{1-r_1^2}}$. The ratio of *d*'s closely approximates the ratio of *r*'s only if the two correlations are both small or very similar to each other, so that the term inside the square root becomes approximately 1.

50% obtained from twin studies of intelligence" (p. 151). However, a polygenic score that explains "just" 10% of the variance can be expected to predict the actual value of the phenotype almost half as well as the full genotype ($\sqrt{.10/.50} \approx 0.45$), assuming that the estimate from twin studies is correct.

As an important caveat to the above paragraphs, I wish to stress that there are contexts in which directly comparing heritabilities has a natural, meaningful interpretation. For example, the response of a trait under selection can be predicted with the breeder's equation $R = h^2S$, where S is the selection differential (i.e., the within-generation change in the trait mean), R is the response to selection (i.e., the between-generation change in the trait mean), and h^2 is the narrow-sense heritability. In this particular context, the effect of interest—that is, the response to selection—is directly proportional to the heritability; for instance, doubling the heritability of a trait will double its response to the same amount of selection. More generally, I am emphatically *not* suggesting that the heritability is a meaningless quantity. The point is that the real-world effect of additive genetic factors on a trait (in terms of the trait's original units) is not quantified by the heritability but by its square root.

In quantitative genetics, the routine use of variance components as indices of effect size may have led researchers to underestimate the impact of shared environmental factors (i.e., those aspects of the environment that tend to increase the similarity between siblings). In his famous paper on the "three laws of behavior genetics", Turkheimer (2000) expressed a common perception in the field: "Although according to the second law shared environment accounts for a small proportion of the variability in behavioral outcomes, according to the third law, nonshared environment usually accounts for a substantial portion. So perhaps the appropriate conclusion is not so much that the family environment does not matter for development, but rather that the part of the family environment that is shared by siblings *does not matter*" (p. 162; emphasis mine).

Specific examples of deflationary interpretations of the shared environmental variance can be found in various sources, including Knopik et al.' s (2017) classic textbook of behavior genetics. After discussing results that estimate the heritability of verbal and spatial abilities at about 40-50%, the authors noted: "[H]owever, adoption designs show little influence of shared environment. For example, the correlations for adoptive siblings are only about 0.10, suggesting that only 10 percent of the variance of verbal and spatial abilities is due to shared environmental factors" (p. 174). But these figures correspond to an effect ratio of about 2, meaning that shared environmental factors are roughly half as influential as genetic ones. Or: "Large twin studies found similar results [heritability around 60%] in the early school years for both reading disability and reading ability. However, in all of these studies, shared environmental influence is modest, typically accounting for less than 20 percent of the variance" (p. 205). But even a "modest" 10% of variance would indicate a real-world effect 40% as large as that of genes ($\sqrt{.10/.60} \approx 0.41$).

In their comprehensive meta-analysis of 50 years of human twin studies (including cognitive and behavioral traits but also morphological, metabolic, reproductive traits, etc.), Polderman et al. (2015) estimated the mean heritability across traits at 48.8% and the mean shared environmental component at 17.4%. Taken at face value, these figures seem to suggest that additive genetic factors are almost three times as influential as the shared environment (explained variance ratio: 2.80); but in terms of real-world effects on the phenotype, the impact

of genes is only 67% larger than that of the shared environment (effect ratio: $\sqrt{2.80} \approx 1.67$). Equivalently, one could say that the impact of the shared environment is 60% as large as that of genes (the reciprocal of 1.67 is 0.60). Similarly, the heritable and shared environmental components of criminality and substance use have been estimated at about 50% and 20%, respectively (Kendler et al., 2016, 2019). Translated into real-world effects, this corresponds to a ratio of about $\sqrt{.50/.20} \approx 1.58$ (or its reciprocal 0.63), meaning that shared environmental factors are about 60% as influential as genetic factors.[4]

It is important to note that the discrepancy between the ratio of explained variances and the effect ratio becomes more pronounced as the effects being compared grow more different from each other. For example, consider a trait that is 50% heritable and has a 5% shared environmental component (ten times smaller than the heritability). Many would regard that 5% of variance as very small, or even practically negligible; but in fact, the effect ratio is only $\sqrt{10} \approx 3.16$ (reciprocal 0.32), meaning that the impact of the shared environment on the phenotype is about one third of that of genes (!). Even a shared environmental component of just 1% is not as tiny as it looks against a heritability of 50%. The effect ratio in this case is $\sqrt{50} \approx 7.07$, meaning that genetic factors are seven times more influential than the shared environment—a substantial difference, but not nearly as large as indicated by the size of the variance components.[5]

Psychometrics is another discipline in which variance components are routinely calculated and directly interpreted by researchers. In classical test theory, the reliability of a scale is the proportion of *true score variance* (i.e., variance shared with the latent construct being measured) on the total variance, with the remainder accounted for by measurement error. Thus, directly comparing the reliabilities of different scales may give a misleading impression of their relations with the latent construct of interest. If two scales have reliabilities of .60 and .80, the higher-quality scale correlates with the latent construct only 15% more strongly than the lower-quality one (effect ratio: $\sqrt{.80/.60} \approx 1.15$).

This phenomenon can become especially insidious when the variance of psychometric scales is parsed at a finer scale of analysis. For example, McCrae (2015, Table 1) reported that measured scores on the facets of the Big Five personality domains contain an average of 34% common trait variance (i.e., variance shared with the broader domain) and 9% facet-specific variance. (To illustrate: in this model of personality, the broad domain of Extraversion has six narrower facets: Warmth, Gregariousness, Assertiveness, Activity, Excitement seeking, and

[4] Hunter and Schmidt (1990, 2014) illustrated this point with a similar example from behavior genetics: if intelligence is 80% heritable and 20% environmental (which may be the case in older adults, at least in wealthier countries; see Plomin & Deary, 2015), the proper ratio between the real-world genetic and environmental effects is two, not four as suggested by the size of the variance components.

[5] Here I am assuming that the 1% or 5% of shared environmental variance in these examples represents a reliable effect, and not a spurious estimate resulting from sampling error or other forms of bias. At least in some cases, it is reasonable to treat small variance components as effectively zero; my argument only applies to genuine nonzero effects that happen to account for a small proportion of the variance. Note that shared environmental effects that account for a few percent of the variance cannot be reliably detected in twin studies unless sample size is quite large; hence, they are often dropped from the best-fitting model and set to zero (see e.g., Burt, 2014).

Positive emotions.) These figures seem to suggest that a person's true score on a given personality domain (e.g., Extraversion) contributes to their measured score on a facet of that domain (e.g., Assertiveness) almost four times as much as their true score on the facet itself (explained variance ratio: 3.78). However, the effect ratio is a markedly smaller $\sqrt{3.78} \approx 1.94$, meaning that personality domains contribute to scores about twice as much as facets (instead of almost four times as much).

Later in the same paper, McCrae (2015) estimated variance components for scores on single personality items. On average, the total score variance consisted of 12% common trait variance; 24% item-specific variance; 13% method variance (regarded as systematic error); and 51% random error. McCrae concluded: "The observed values are sobering: In the typical item, nearly two thirds of the variance is either random or systematic error […], which is why single items are notoriously unreliable; of the remaining true-score variance […], only a third is due to the common trait" (p. 106). However, effect ratios paint a less sobering picture: measurement error contributes only 33% more than true score variation ($\sqrt{(.51 + .13)/(.24 + .12)} \approx 1.33$), and the contribution of common trait variation to item scores is about 70% as large as that of item-specific variation ($\sqrt{.12/.24} \approx 0.71$).

The use of variance components as measures of effect size is especially widespread in behavior genetics and psychometrics, but by no means limited to these disciplines. An entire line of methodological research—under the rubric of "relative importance analysis" or "relative weight analysis"—seeks to supplement standard multiple regression coefficients with indices that quantify the amount of variance explained by each predictor in the model, in order to rank and compare predictors in terms of their importance (see Johnson & LeBreton, 2004; Tonidandel & LeBreton, 2011, 2015). Relative importance analysis is popular in various areas of the applied social sciences, including organizational, vocational, and business psychology. The rationale for relying on variance components is that, unlike regression coefficients, they are additive and sum to the total R^2 of the model. But while additivity is a convenient property, variance-based indices can dramatically magnify the apparent differences in importance among predictors, even when their real-world effects on the outcome are not very dissimilar.[6] In principal component analysis (PCA) and exploratory factor analysis (EFA), components and factors are routinely ranked based on the proportion of variance they explain in the original variables (as measured by the corresponding eigenvalues) ; this may easily inflate the perceived differences between "strong" and "weak" dimensions of variation in the data.

In sum: using the proportion of explained variance as an index of effect size does not just distort the real-world magnitude of individual effects, but also exaggerates the *differences* between effects, which may lead to strikingly incorrect judgements of relative importance. Luckily, a meaningful and interpretable "effect ratio" can be easily calculated as the square root of the ratio between proportions of explained variance. In a number of practical examples, effect ratios tell a different story than variance components, and might prompt one to rethink the

[6] A similar criticism applies to the summary indices I have proposed to quantify the heterogeneity in the contributions of individual variables to Mahalanobis' *D*, which is the multivariate generalization of Cohen's *d* (Del Giudice, 2017, 2018). Those indices rely on a decomposition of the total squared effect size into a weighted sum of squared univariate effect sizes, and arguably provide an inflated sense of heterogeneity across variables.

interpretation of certain canonical results (e.g., regarding the role of the shared environment in the development of psychological traits). This simple but consequential point should be understood more widely; with no pretense of originality, I hope that this note will contribute to raise awareness and prevent fallacious interpretations of research findings.

Acknowledgments

I wish to thank Mike Bailey, Steve Gangestad, and Emil Kirkegaard for their helpful comments on a previous draft of this manuscript.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129-133. doi: 10.1037/0033-2909.97.1.129

Beatty, M. J. (2002). Do we know a vector from a scalar? Why measures of association (not their squares) are appropriate indices of effect. *Human Communication Research, 28,* 605-611. doi: 10.1111/j.1468-2958.2002.tb00827.x

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29,* 79-97. doi: 10.1177/014920630302900106

Burt, A. S. (2014). The shared environment as a key source of variability in child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry, 55,* 304-312. doi: 10.1111/jcpp.12173

D'Andrade, R., & Dart, J. (1990). The interpretation of r versus r^2, or why percent of variance accounted for is a poor measure of size of effect. *Journal of Quantitative Anthropology, 2,* 47–59.

Del Giudice, M. (2017). Heterogeneity coefficients for Mahalanobis' D as a multivariate effect size. *Multivariate Behavioral Research, 52,* 216-221. doi: 10.1080/00273171.2016.1262237

Del Giudice, M. (2018). Addendum to: Heterogeneity coefficients for Mahalanobis' D as a multivariate effect size. *Multivariate Behavioral Research, 53,* 571-573. doi: 10.1080/00273171.2018.1462138

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2,* 156-168. doi: 10.1177/2515245919847202

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings* (1st ed.). Sage.

Hunter, J. E., & Schmidt, F. L. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Sage.

Johansson, V., Kuja-Halkola, R., Cannon, T. D., Hultman, C. M., & Hedman, A. M. (2019). A population-based heritability estimate of bipolar disorder–In a Swedish twin sample. *Psychiatry Research, 278,* 180-187. doi: 10.1016/j.psychres.2019.06.010

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods, 7,* 238-257. doi: 10.1177/1094428104266510

Johnson, W. (2011). Correlation and explaining variance: To square or not to square? *Intelligence, 39,* 249-254. doi: 10.1016/j.intell.2011.07.001

Knopik, V. S., Neiderhiser, J. M., DeFries, J. C., & Plomin, R. (2017). *Behavioral genetics* (7th ed.). Worth.

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review, 19,* 97-112. doi: 10.1177/1088868314541857

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315.

Plomin, R., & Deary, I. J. (2015). Genetics and intelligence differences: Five special findings. *Molecular Psychiatry, 20,* 98-108. doi: 10.1038/mp.2014.105

Plomin, R., & von Stumm, S. (2018). The new genetics of intelligence. *Nature Reviews Genetics, 19,* 148-159. doi: 10.1038/nrg.2017.104

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics, 47,* 702-709. doi: 10.1038/ng.3285

Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology, 9,* 395-396. doi: 10.1111/j.1559-1816.1979.tb02713.x

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology, 26,* 1-9. doi: 10.1007/s10869-010-9204-3

Tonidandel, S., & LeBreton, J. M. (2015). RWA web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. *Journal of Business and Psychology, 30,* 207-216. doi: 10.1007/s10869-014-9351-z

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science, 9,* 160-164. doi: 10.1111/1467-8721.00084