MULTIVARIATE BEHAVIORAL RESEARCH, 2018
VOL. 53, NO. 4, 571–573
https://doi.org/10.1080/00273171.2018.1462138

Routledge
Taylor & Francis Group

Addendum to: Heterogeneity Coefficients for Mahalanobis' *D* as a Multivariate Effect Size

Marco Del Giudice

Department of Psychology, University of New Mexico

ABSTRACT

In a previous paper (Del Giudice, 2017 [Heterogeneity coefficients for Mahalanobis' D as a multivariate effect size. *Multivariate Behavioral Research, 52*, 216–221]), I proposed two heterogeneity coefficients for Mahalanobis' *D* based on the Gini coefficient, labeled *H* and *EPV*. In this addendum I discuss the limitations of the original approach and note that the proposed indices may overestimate heterogeneity under certain conditions. I then describe two revised indices H_2 and EPV_2, and illustrate the difference between the original and revised indices with some real-world data sets.

KEYWORDS
Effect size; group differences; heterogeneity; Mahalanobis distance; multivariate

Multivariate effect sizes such as Mahalanobis' D raise the issue of heterogeneity in the contributions of individual variables to the overall effect. In Del Giudice (2017) I proposed a strategy to quantify heterogeneity: first, partition D^2 into a set of non-negative values that reflect the contributions of individual variables; second, apply the small-sample Gini formula to those values to obtain a heterogeneity coefficient ranging between 0 and 1. The critical step in this strategy is finding a suitable partition of D^2. In Del Giudice (2017) I used a simple approach to partition the multivariate D^2 into a weighted sum of the squared univariate effects (d_i^2). The resulting C_i values have two desirable properties. First, $C_i = 0$ if removing variable X_i from the set leaves D unchanged. For two variables X_1, X_2 with correlation r and effect sizes d_1, d_2 (the case I will use for illustration here):

$$C_1 = \frac{1}{1-r^2}\left(1 - r\frac{d_2}{d_1}\right) d_1^2 \tag{1}$$

and

$$C_2 = \frac{1}{1-r^2}\left(1 - r\frac{d_1}{d_2}\right) d_2^2. \tag{2}$$

It follows that $C_1 = 0$ when $d_1 = rd_2$, which is appropriate since in this case $D^2 = (C_1 + C_2) = d_2^2$. Conversely, $C_2 = 0$ when $d_2 = rd_1$, and $D^2 = d_1^2$. Second, when two highly correlated variables X_i and X_j are both in the set, their joint contribution is split between C_i and C_j, and is not partialed out as it would happen with methods for quantifying the contribution of individual variables that proceed by removing one variable at a time (e.g., Rencher, 1993).

In the original paper, I suggested that C_i values can be interpreted as "net contributions" to D^2. This is incorrect: C_i values can be negative, even though D^2 never decreases when more variables are added to the set. The fact that C_i values can be negative also makes them unsuitable for calculating the standard Gini coefficient. In Del Giudice (2017), I proposed an ad-hoc solution to this problem, namely, setting negative C_i values to zero and using the resulting C_i^* values to calculate two Gini-based heterogeneity coefficients, H and EPV. However, this approach is not ideal, because a negative C_i value still reflects a positive contribution of X_i to D (that is, D decreases if X_i is removed from the set). To illustrate, in the two-variable case with $r > 0$, negative values $C_i < 0$ occur whenever $0 < d_i < rd_j$; it is easy to show that under the same conditions $D^2 = (C_i + C_j) > d_j^2$, which implies that X_i makes a positive contribution to D. Negative C_i values are more likely to occur when d_i is small relative to d_j and the two variables are strongly correlated. In sum, setting negative C_i values to zero is an overly conservative approach—it tends to underestimate the contribution of some variables to D, and hence overestimate the amount of heterogeneity. The resulting distortion tends to become larger as collinearity among the variables increases.

A better solution to the problem of negative C_i values is to use the ordered absolute values $|C_1| \ldots |C_n|$ (where $|C_1| < |C_2| \ldots < |C_n|$, and $\overline{|C|}$ is their average) to

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

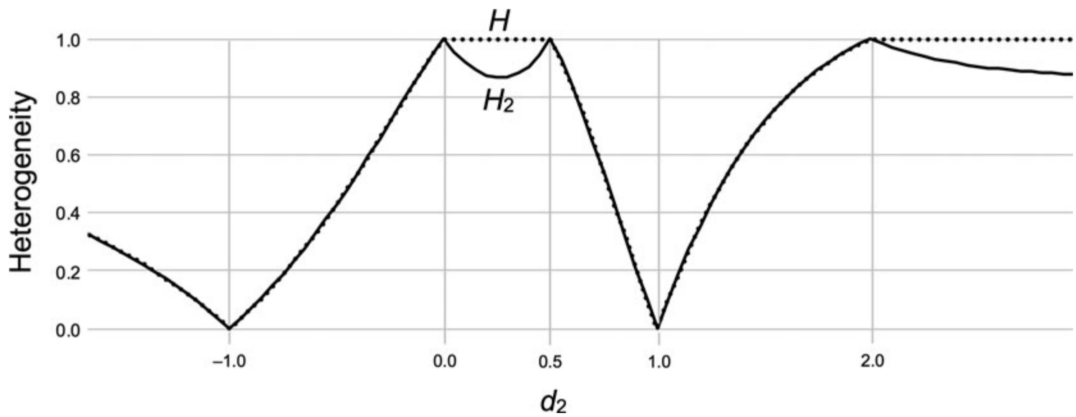

Figure 1. Behavior of heterogeneity coefficients H and H_2 in the case of two positively correlated variables. (In the example shown, $d_1 = 1.0$ and $r = .5$; the qualitative pattern does not depend on the choice of values.) The dotted line shows the original H coefficient. The solid line shows the revised H_2 coefficient discussed here. For both coefficients, heterogeneity is minimal when $d_2 = \pm d_1$ (in this example, $d_2 = \pm 1$) and maximal when $d_2 = 0$, when $d_2 = rd_1$ (in this example, $d_2 = 0.5$), and when $d_2 = d_1/r$ (in this example, $d_2 = 2$).

calculate heterogeneity. To avoid confusion, the resulting coefficients can be labeled as H_2 and EPV_2:

$$H_2 = \frac{(2/n)\sum_{i=1}^{n} i|C_i| - [(n+1)/n]\sum_{i=1}^{n}|C_i|}{(n-1)\overline{|C|}} \tag{3}$$

and

$$EPV_2 = 1 - \frac{n-1}{n}H_2. \tag{4}$$

Figure 1 illustrates the behavior of H and H_2 in the two-variable case (with $r, d_1 > 0$). For both $0 < d_2 < rd_1$ and $d_2 > d_1/r$ (equivalent to $0 < d_1 < rd_2$), coefficient H remains equal to 1 (maximum heterogeneity, wrongly indicating that only one variable contributes to D), whereas H_2 correctly decreases to reflect the nonzero contributions of d_2 and d_1, respectively.

To illustrate the difference between H and H_2 with some real-world examples, consider the empirical data sets analyzed in Del Giudice (2017). For the aggression data set from Del Giudice (2009), using H or H_2 makes no difference because there are no negative C_i values. For the personality data set from Del Giudice, Booth, and Irwing (2012), $H = .95$ ($EPV = .11$) while $H_2 = .90$ ($EPV_2 = .16$). After removing the "sensitivity" factor from the analysis, $H = .80$ ($EPV = .25$) while $H_2 = .76$ ($EPV_2 = .30$). For the brain anatomy data discussed in Del Giudice et al. (2016), H ranges from .44 to .70 (EPV from .36 to .58); the corresponding values of H_2 range from .32 to .58 (EPV_2 from .47 to .71). In these examples, H_2 does not dramatically change the picture, but it does suggest a somewhat more homogeneous contribution than indicated by H.

While coefficient H_2 improves on the original H, both have limitations owing to their reliance on the C partition. In particular, $C_i = 0$ whenever $d_i = 0$; however, a variable X_i may contribute to increase D even if $d_i = 0$, provided that it has nonzero correlations with the other variables. In the two-variable case, it is easy to show that if $d_i = 0$ and $r \neq 0$, then $D^2 > d_j^2$. Future research may show a way to partition D^2 so as to avoid this problem while maintaining the desirable properties of C. The currently available alternatives are not well suited for the task—for example, the partitioning method recently proposed by Garthwaite and Koch (2016) avoids the problem of negative values, but fails to identify cases in which one of the variables makes no contribution to D (in the two-variable case, when $d_2 = rd_1$ or $d_2 = d_1/r$). At present, H_2 offers a practical means to quantify heterogeneity and should prove useful in a variety of applications.

Article information

Conflict of Interest Disclosures: The author signed a form for disclosure of potential conflicts of interest. The author did not report any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The author affirms having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not funded or sponsored.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The author would like to thank Benjamin Reiser for his comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the author alone, and endorsement by the author's institution is not intended and should not be inferred.

References

Del Giudice, M. (2009). On the real magnitude of psychological sex differences. *Evolutionary Psychology*, *7*, 264–279. doi:10.1177/147470490900700209

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., Booth, T., & Irwing, P. (2012). The distance between Mars and Venus: Measuring global sex differences in personality. *PLoS ONE*, *7*, e29265. doi:10.1371/journal.pone.0029265

Del Giudice, M., Lippa, R. A., Puts, D. A., Bailey, D. H., Bailey, J. M., & Schmitt, D. P. (2016). Joel et al.'s method systematically fails to detect large, consistent sex differences. *Proceedings of the National Academy of Sciences USA*, *113*, E1965. doi:10.1073/pnas.1525534113

Garthwaite, P. H., & Koch, I. (2016). Evaluating the contributions of individual variables to a quadratic form. *Australian & New Zealand Journal of Statistics*, *58*, 99–119. doi:10.1111/anzs.12144

Rencher, A. C. (1993). The contribution of individual variables to Hotelling's T^2, Wilks' Λ, and R^2. *Biometrics*, *49*, 479–489. doi:10.2307/2532560