JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 540, 1627–1630
https://doi.org/10.1080/01621459.2022.2117703

Taylor & Francis
Taylor & Francis Group

Modeling and Learning From Variation and Covariation

Blakeley B. McShane[a], Ulf Böckenholt[a], and Karsten T. Hansen[b]

[a]Kellogg School of Management, Northwestern University, Evanston, IL; [b]Rady School of Management, University of California, San Diego, La Jolla, CA

We heartily thank editor Heping Zhang for according our article (McShane, Böckenholt, and Hansen 2022) discussion. We are extremely grateful for the opportunity to receive comments on our work from a set of distinguished discussants who possess a tremendous breadth and depth of knowledge and expertise, and we thank them profoundly for the great deal of time and effort they put into contemplating and responding to our article.

We were delighted that De Boeck, DeKay, and Xu (2022; hereafter DDX) appreciated our contributions, namely (i) our proposal to quantify the variation and covariation in variates and effects at all levels and (ii) our multilevel multivariate modeling framework for doing so. We were gratified that they recognized that our proposal addresses not only replicability but also generalizability and integrability; found that our modeling framework was easily adapted to and proved useful for their reanalysis of the data from the second metastudy of DeKay et al. (2022); and discussed how our framework would prove useful for the analysis of the data from potential future integrative metastudies.

Inspired by the discussion, we make five comments regarding modeling and learning from variation and covariation.

1. Simple Modeling of Variation and Covariation

As we discussed in our article, the typical approach to the analysis of the data from large-scale replication projects either foregoes quantifying the variation and covariation in variates and effects at the lab and subject levels or quantifies the variation in effects at the lab level and foregoes quantifying all other variation and covariation. What we did not discuss is that even highly simple and accessible models yield results that indicate substantial variation and covariation in our applications which in turn calls for more sophisticated modeling.

For example, consider the hierarchical (or multilevel) linear model with lab and subject terms—the most basic model possible that quantifies variation and covariation at the lab and subject levels. The model specification for this baseline model is given by

$$y_{i,v} = \alpha_v + \beta_{l[i]} + \gamma_i + \varepsilon_{i,v}$$

where $y_{i,v}$ denotes the observation for subject i and variate v; $l[i]$ denotes the lab l at which subject i was observed; the α_v are treated as fixed effects for each variate; the β_l are treated as random effects for each lab; the γ_i are treated as random effects for each subject; and the $\varepsilon_{i,v}$ are random errors for each subject and variate. The model assumes that the β_l are independent and identically distributed according to the normal distribution with mean zero and variance τ_β^2; the γ_i are independent and identically distributed according to the normal distribution with mean zero and variance τ_γ^2; the $\varepsilon_{i,v}$ are independent and distributed according to the normal distribution with mean zero and variance $\sigma_{d[v]}^2$ where $d[v]$ denotes the dependent measure associated with variate v; and there is zero covariation among the β_l, γ_i, and $\varepsilon_{j,v}$ for all l, i, j, and v.

This model can be viewed as the special case of our model that constrains the $\beta_{l,v}$ to be equal to β_l for all v and the $\gamma_{i,v}$ to be equal to γ_i for all v (or equivalently, $\mathbf{T}_{\boldsymbol{\beta}}$ to be equal to $\tau_\beta^2\mathbf{1}\mathbf{1}^{\mathrm{T}}$ and $\mathbf{T}_{\boldsymbol{\gamma}}$ to be equal to $\tau_\gamma^2\mathbf{1}\mathbf{1}^{\mathrm{T}}$ where $\mathbf{1}$ is the vector of ones of length equal to the number of variates). As such, it assumes equal variation and perfect covariation in variates at each level and zero variation and zero covariation in effects at each level.

In the Att-SNARC RRR, the estimates of the variation in variates at the lab and subject levels from this model (i.e., τ_β and τ_γ, respectively), which we present in the left panel of Figure 1, are 15 ms at the lab level and 44 ms at the subject level. These estimates are remarkably similar to those presented in the left panel of Figure 3 of our article to which these can be directly compared.

In the MLP, the estimates of the variation in variates at the lab and subject levels from the analogue of this model (i.e., τ_β and τ_γ, respectively), which we present in the upper left panel of Figure 2, are 0.11 σ_d at the lab level and 0.17 σ_d at the subject level. These estimates are smaller than those presented in the left panel of Figure 8 of our article to which these can be compared, but they are nonetheless substantial.

Further, when this model is extended to accommodate the repeated measures of the experimental conditions associated with the Anchoring phenomenon by introducing additional terms for them, the estimates of the variation in variates at the lab and subject levels, which we present in the lower left panel of Figure 2, are 0.15 σ_d for the variates not associated with the Anchoring phenomenon, 0.09 σ_d for the variates associated with the high anchor condition, and 0.29 σ_d for the variates associated with the low anchor condition at the lab level and 0.22 σ_d for the variates not associated with the Anchoring phenomenon,

CONTACT Blakeley B. McShane b-mcshane@kellogg.northwestern.edu Kellogg School of Management, Northwestern University, Evanston, IL.

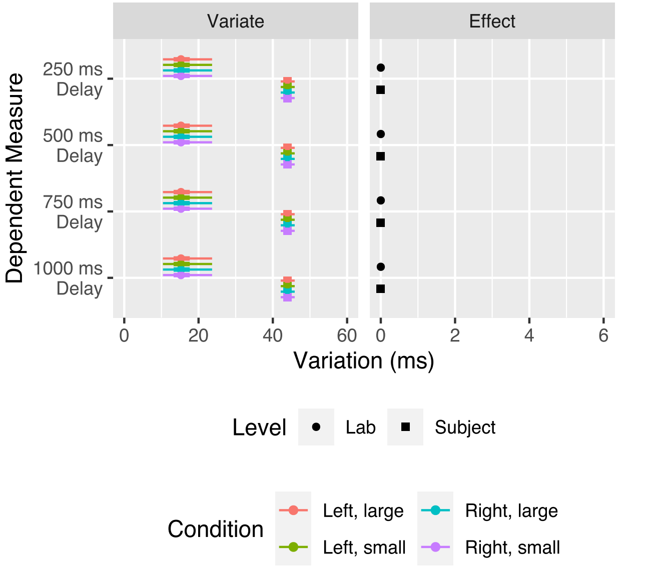

Figure 1. Att-SNARC RRR variation estimates from the baseline model. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively. The baseline model assumes equal variation in the 16 variates at each level and zero variation in the four effects at each level.

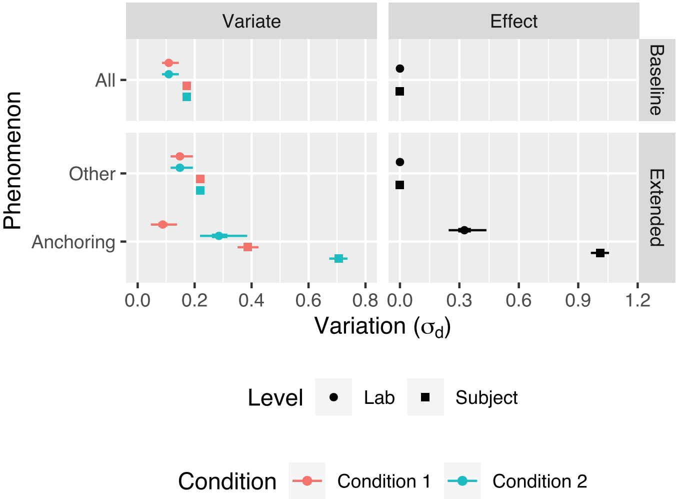

Figure 2. MLP variation estimates from the baseline and extended models. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively. The baseline model assumes equal variation in the 30 variates at each level and zero variation in the 15 effects at each level. The extended model assumes equal variation in the 22 variates not associated with the Anchoring phenomenon, equal variation in the four variates associated with the high anchor condition, and equal variation in the four variates associated with the low anchor condition at each level; it also assumes zero variation in the 11 effects not associated with the Anchoring phenomenon and equal variation in the four effects associated with the Anchoring phenomenon at each level.

0.39 σ_d for the variates associated with the high anchor condition, and 0.71 σ_d for the variates associated with the low anchor condition at the subject level. In addition, the estimates of the variation in the effects associated with Anchoring phenomenon at the lab and subject levels from this extended model, which we present in the lower right panel of the figure, are 0.33 σ_d at the lab level and 1.01 σ_d at the subject level.

In sum, even highly simple and accessible models yield results that indicate substantial variation and covariation in our applications which in turn calls for more sophisticated modeling (of course, their not having yielded results that indicate substantial variation and covariation would not have implied that more sophisticated models would also not have yielded results that indicate substantial variation and covariation and therefore it is necessary to consider such more sophisticated models— particularly to support claims of a lack of substantial variation and covariation).

2. Modeling Variation and Covariation in Effects

The models discussed above are obviously overly simplistic for modeling variation and covariation. Arguably, however, so too are the particular model specifications employed in our applications—especially for modeling variation and covariation in effects. For example, consider the three-factor structure employed to model $\mathbf{T}_{\beta}$ and $\mathbf{T}_{\gamma}$ in the Att-SNARC RRR. Because the observations $y_{i,v}$ are of variates and because the effects of interest are contrasts of a particular dependent measure as assessed across multiple experimental conditions, this three-factor structure implies a single-factor structure for effects, which is of course a highly—and arguably overly—constrained factor structure.

Yet, perhaps it is *not* overly constrained. As we discussed in our article, we expect the variation and covariation in variates to be higher than that in effects due to the zero-sum nature of contrasts. Further, it is arguably not implausible that the covariation in the variates associated with one dependent measure and those associated with another dependent measure would be similar (or even the same) across the experimental conditions associated with each of the dependent measures; this would in turn imply low (or zero) covariation in the effects associated with the dependent measures. Indeed, insofar as the covariation in variates is caused by or associated with some covariate or set of covariates and the covariate(s) have an effect on the variates associated with each dependent measure that is the same across the experimental conditions associated with the dependent measure—as is typically assumed in psychological research, for example, by ANCOVA models—that covariation in variates would be the same and thus the covariation in effects would be zero.

3. Modeling Variation and Covariation in Metastudies

DDX found that our modeling framework was easily adapted to and proved useful for their reanalysis of the data from the second metastudy of DeKay et al. (2022), which investigated a single phenomenon and involved a single lab. We discuss model extensions to accommodate metastudies that investigate multiple phenomena and involve multiple labs.

First, when a metastudy involves multiple labs, one approach is to simply view the lab as what DDX label a secondary design factor. In this case, the levels would be the microstudy and subject levels and our modeling framework could be applied with only minor modification (e.g., the microstudy level playing the role of the lab level). Then, the estimates could be examined *ex post* as a function of the secondary design factors as illustrated in Figure 1 of DDX's comment.

Alternatively, one can view the levels as the lab and subject levels and incorporate the secondary design factors into the factor structure. One manner of doing so at the lab level is to recognize that a metastudy is simply a study that follows a factorial design, albeit one with a larger number of factors than is typical, and to recall that when studies follow a factorial design, factors corresponding to this design can be introduced as we discussed in our article. Other possibilities include viewing the levels as the lab, microstudy, and subject levels, with either microstudies nested within labs or labs and microstudies crossed.

Second, when a metastudy investigates multiple phenomena, several of the model extensions we discussed in our article should prove useful. In particular, when a metastudy investigates multiple related phenomena and employs repeated measures for all phenomena under investigation (as do the integrative metastudies discussed by DDX), extending the factor structure to introduce a phenomenon factor as well as intercept, dependent measure, and variate factors for each phenomenon should prove particularly useful.

4. Learning from Variation and Covariation in Our Applications

DDX note that replicability, generalizability, and integrability are three important aspects of research and that our proposal to quantify the variation and covariation in variates and effects at all levels addresses all three. We discuss how it does so in our applications, noting that variation pertains more to replicability and generalizability and that covariation pertains more to integrability.

Replicability was covered in our article. Indeed, it was the focus of our article to the want of other considerations. This focus was inevitable due to design choices, specifically the choice of the Att-SNARC RRR and the MLP to—like all RRRs and MLPs conducted to date—use the same study materials within and across the labs involved in the projects. This choice severely constrains the degree of variation in variates and effects in these projects. Consequently, it is perhaps rather remarkable that our results indicated medium to high variation in variates and effects across labs in the MLP and thus a medium to low degree of replicability depending on the variate or effect.

In terms of generalizability, DDX comment that "[t]he kind of generalization that is of interest in such studies [as the Att-SNARC RRR and the MLP] is generalization across labs and their participant populations, using the exact same design and experimental materials in each lab." We are less sanguine that our results regarding the variation across labs in these projects can address such generalizability again due to design choices, namely the fact that the labs involved in these projects are a convenience sample of labs.

In terms of integrability, because the Att-SNARC RRR—like RRRs in general—investigated only a single phenomenon, it can address integrability to only a limited degree (i.e., across the four dependent measures associated with the phenomenon). However, because the MLP investigated multiple phenomena, it can address integrability to a greater degree yet again subject to design choices. Our results indicated covariation related to political ideology which suggested expanding the model to include political ideology which in turn suggested that political ideology associates with many of the variates.

Whether this reflects something fundamental or is an artifact of design choices remains to be seen. On one hand, political ideology is known to cause or associate with a host of psychological constructs (see, e.g., Jost et al. 2003). On the other hand, many of the dependent measures associated with the phenomena investigated by the MLP relate to political ideology, but for many of these phenomena, a dependent measure related to political ideology seems neither intrinsic to nor necessary for the phenomenon. For example, the dependent measure associated with the Allowed versus Forbidden phenomenon relates to freedom of speech, but the phenomenon could be investigated using alternative dependent measures, and in particular ones not so obviously related to political ideology. Insofar as covariation related to political ideology remains when such alternative dependent measures are used, this would suggest our results reflect something fundamental and are not an artifact of design choices. This seems worthy of examination in future research.

5. Learning from Variation and Covariation in Future Projects

In large-scale replication projects, replication has typically been evaluated based on a single replication study of some original study and dichotomously as successful or failed. Further, this dichotomization has typically been made based on criteria rooted in the null hypothesis significance testing paradigm. Finally, the replication study has typically been designed to reproduce the original study as closely as possible.

In addition, even large-scale replication projects such as RRRs and MLPs that have conducted multiple replication studies of some original study (i.e., one at each lab involved in the project) have typically evaluated replication dichotomously, based on criteria rooted in significance testing, and in a manner that could be considered to be based on a single study. Moreover, they have used the same study materials within and across the labs involved in the projects and these materials have typically been designed to reproduce the original study as closely as possible.

This state of affairs is perhaps curious as Rosenthal (1990), drawing on his work dating from the 1960s, already dismissed evaluations of replicability that are dichotomous and rooted in significance testing as "[t]he traditional, not very useful view of replication" and advocated evaluations of replicability that are continuous and rooted in effect sizes as "[t]he newer, more useful view of replication." He also dismissed evaluations of replicability that are based on a single study as inadequate and advocated evaluations of replicability that are based on multiple studies. He further advocated that these multiple studies "vary in [their] degree of similarity to the original study" to allow the studies to address generalizability.

Expanding on Rosenthal, we made four recommendations for future large-scale replication projects in our article, namely that they (i) quantify the variation and covariation in variates and effects at all levels, (ii) employ repeated measures for all phenomena under investigation, (iii) focus on a single phenomenon or a set of related phenomena, and (iv) vary the study materials within and across the labs involved in the projects. The fourth recommendation allows the project to address generalizability across differing implementations of the study, and we view the metastudy methodology discussed by DDX as an appealing means of adhering to this recommendation; indeed, DeKay et al. (2022) has already demonstrated the value of the metastudy methodology for the Gain versus Loss Framing phenomenon investigated by the MLP. Further, the second and third recommendations allow the project to address integrability across the various dependent measures associated with (i.e., operationalizations of) a phenomenon and the phenomena under investigation, respectively. Combined with the first recommendation, whether adhered to using our modeling framework or otherwise, projects designed and analyzed in accordance with these recommendations will be capable of providing not only evaluations of replicability that are based on multiple studies, continuous, and multi-faceted but also addressing generalizability across differing implementations of the study and integrability of the dependent measures and the phenomena under investigation.

ORCID

Blakeley B. McShane http://orcid.org/0000-0002-4839-266X
Ulf Böckenholt http://orcid.org/0000-0003-3663-8260

References

De Boeck, P., DeKay, M., and Xu, M. (2022), "The Potential of Factor Analysis for Replication, Generalization, and Integration," *Journal of the American Statistical Association*, this issue, DOI: 10.1080/01621459.2022.2096618. [1627]

DeKay, M. L., Rubinchik, N., Li, Z., and De Boeck, P. (2022), "Accelerating Psychological Science with Metastudies: A Demonstration using the Risky-Choice Framing Effect," *Perspectives on Psychological Science*, forthcoming. [1627,1628,1630]

Jost, J. T., Glaser, J., Kruglanski, A. W., and Sulloway, F. J. (2003), "Political Conservatism as Motivated Social Cognition," *Psychological Bulletin*, 129, 339–375. [1629]

McShane, B. B., Böckenholt, U., and Hansen, K. T. (2022), "Variation and Covariation in Large-Scale Replication Projects: An Evaluation of Replicability," *Journal of the American Statistical Association*, this issue, DOI: 10.1080/01621459.2022.2054816. [1627]

Rosenthal, R. (1990), "Replication in Behavioral Research," *Journal of Social Behavior and Personality*, 5, 1–30. [1629]