JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 540, 1605–1621: Application and Case Studies Discussion
https://doi.org/10.1080/01621459.2022.2054816

Taylor & Francis
Taylor & Francis Group

Variation and Covariation in Large-Scale Replication Projects: An Evaluation of Replicability

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

ABSTRACT

Over the last decade, large-scale replication projects across the biomedical and social sciences have reported relatively low replication rates. In these 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. However, evaluations of replicability that are based on a single study and are dichotomous are inadequate, and evaluations of replicability should instead be based on multiple studies, be continuous, and be multi-faceted. Further, such evaluations are in fact possible due to two characteristics shared by many large-scale replication projects. In this article, we provide such an evaluation for two prominent large-scale replication projects, one which replicated a phenomenon from cognitive psychology and another which replicated 13 phenomena from social psychology and behavioral economics. Our results indicate a very high degree of replicability in the former and a medium to low degree of replicability in the latter. They also suggest an unidentified covariate in each, namely ocular dominance in the former and political ideology in the latter, that is theoretically pertinent. We conclude by discussing evaluations of replicability at large, recommendations for future large-scale replication projects, and design-based model generalization. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2021
Accepted February 2022

KEYWORDS
Factor analysis;
Meta-analysis; Multilevel;
Multivariate; Psychology;
Replicability; Replication

1. Introduction

Over the last decade, researchers across the biomedical and social sciences have raised concern about the replicability of research studies. For example, they have reported replication rates of only about one-quarter for pharmaceutical development studies (Prinz, Schlange, and Asadullah 2011), one-tenth for cancer research studies (Begley and Ellis 2012), one-third for psychological research studies (Open Science Collaboration 2015), six-tenths for economics research studies (Camerer et al. 2016), and sixth-tenths for research studies across the social sciences (Camerer et al. 2018). In these 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, with the most popular criterion being that the replication study is considered to have successfully replicated the original study if either both failed to attain "statistical significance" or both attained "statistical significance" and were directionally consistent and to have failed to replicate the original study otherwise.

However, evaluations of replicability that are based on a single study and are dichotomous are inadequate, and evaluations of replicability should instead be based on multiple studies, be continuous, and be multi-faceted (McShane et al. 2019). Further, such evaluations are in fact possible due to two characteristics shared by many large-scale replication projects, including the various Registered Replication Reports (RRRs) and the various Many Labs Projects (MLPs).

The first characteristic is that many large-scale replication projects are multilevel in nature in that the same phenomenon (e.g., RRRs) or the same phenomena (e.g., MLPs) are investigated across multiple labs (i.e., the same study materials are administered in a coordinated but separate fashion at each lab to different subjects). The multilevel nature of these projects allows for an evaluation of replicability that is based on multiple studies and is continuous, in particular, a quantification of the variation in the estimates from lab to lab involved in the project.

The second characteristic is that many large-scale replication projects are multivariate in nature in that they feature multiple experimental conditions and often feature multiple dependent measures. The multivariate nature of these projects allows for an evaluation of replicability that is multi-faceted, in particular, a quantification of the variation in the estimates of each *variate* (i.e., a particular dependent measure as assessed in a particular experimental condition) as well as each *effect* of interest (i.e., a contrast of a particular dependent measure as assessed across multiple experimental conditions). The multivariate nature also allows for these quantifications at not only the lab level but

CONTACT Blakeley B. McShane b-mcshane@kellogg.northwestern.edu Kellogg School of Management, Northwestern University, Evanston, IL.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

also the subject level. The multivariate nature finally allows for the quantification of not only the variation in the estimates of variates and effects at these levels but also the covariation, which can facilitate the identification of covariates that cause or associate with the variation.

In sum, many large-scale replication projects not only allow for but demand eight quantifications relevant for evaluating replicability, namely that of the variation and covariation in variates and effects at the lab and subject levels. However, this opportunity for quantification has gone almost entirely unrealized because the typical approach to the analysis of the data from these projects is highly impoverished. Specifically, the typical approach has been to (a) collapse the subject level data to a single estimate of each effect for each lab and dependent measure and (b) analyze the estimates of each effect separately for each dependent measure via the basic fixed effects or random effects meta-analytic model. This approach provides no more than one of the eight quantifications relevant for evaluating replicability. In particular, analysis via the fixed effects meta-analytic model foregoes providing all eight quantifications while analysis via the random effects meta-analytic model provides a quantification of the variation in effects at the lab level and foregoes providing the other seven quantifications.

In this article, we provide all eight quantifications relevant for evaluating replicability for two prominent large-scale replication projects. Specifically, we provide them for an RRR which replicated the attentional spatial-numerical association of response codes (Att-SNARC) phenomenon (Colling et al. 2020), a phenomenon from cognitive psychology. Our results indicate very low variation and very high covariation in variates across both labs and subjects; trivial variation and covariation in effects across both labs and subjects; and a very high degree of replicability for all variates and effects. Further, the variation and covariation suggests an unidentified covariate, namely ocular dominance, that is theoretically pertinent.

We also provide them for the original MLP (Klein et al. 2014) which replicated 13 phenomena from social psychology and behavioral economics. Our results indicate medium to high variation and covariation in variates across both labs and subjects; medium to high variation in effects across labs; low to high variation in effects across subjects; relatively low covariation in effects across both labs and subjects; and a medium to low degree of replicability depending on the variate or effect. Further, the variation and covariation suggests an unidentified covariate, namely political ideology, that is theoretically pertinent. Finally, the variation and covariation also suggests three insights regarding Anchoring, one of the phenomena investigated by the MLP.

To quantify the variation and covariation in variates and effects at the lab and subject levels, we introduce a multilevel multivariate modeling framework for analyzing all of the subject level data from large-scale replication projects jointly in a single analysis. Our framework employs a factor analytic structure for the variance-covariance matrices at the lab and subject levels that is specially tailored to the design of these projects. Specifically, the factor analytic structure is constrained based on the design of these projects. This results in three distinct advantages. First, it is interpretable. For example, the estimates of the factor loadings facilitated the identification of the ocular dominance covariate in the Att-SNARC RRR and the political ideology covariate in the MLP. Second, it is adaptable. For example, it can accommodate settings where all of the dependent measures are repeated measures of the same phenomenon as in the Att-SNARC RRR as well as settings where most are single measures of distinct phenomena but several are repeated measures of the same phenomenon as in the MLP. Third, it is parsimonious. This is necessary because there is little data at each level relative to the size of the variance-covariance matrices (i.e., in large-scale replication projects, the number of labs and the number of observations per subject is small relative to the size of these matrices).

Our modeling framework builds on and extends prior work on multilevel multivariate meta-analytic models and factor analytic structures for the variance-covariance matrices of multilevel multivariate models (see, e.g., Kalaian and Raudenbush 1996; Berkey et al. 1998; and McShane and Böckenholt 2018 for the former and Muthén 1994 and Rabe-Hesketh, Skrondal, and Pickles 2004 for the latter). It is sufficiently general to accommodate an arbitrary number of phenomena, dependent measures, experimental conditions, levels, and covariates at any level. It will prove useful for the analysis of the data from past and future large-scale replication projects.

In the remainder of this article, we quantify the variation and covariation in variates and effects at the lab and subject levels for the Att-SNARC RRR and the MLP. We conclude by discussing evaluations of replicability at large, recommendations for future large-scale replication projects, and design-based model generalization.

2. Att-SNARC RRR

2.1. Description

The Att-SNARC RRR is a large-scale replication of the Att-SNARC phenomenon (Fischer et al. 2003), which purports that subjects react more quickly to targets that appear on the left when the targets are preceded by small numbers and react more quickly to targets that appear on the right when the targets are preceded by large numbers. The importance of the Att-SNARC phenomenon derives from two facts considered in tandem. First, one of the foundational issues in cognitive psychology is how individuals represent concepts, for which there are two broad accounts: (a) classical ones that view these representations as symbolic (i.e., the representations *do not* capture characteristics of what they represent) and distinct from sensorimotor processing (see, e.g., Fodor 1975 and Newell and Simon 1976) and (b) alternative ones—termed embodied, situated, or grounded—which view them as analogical (i.e., the representations *do* capture characteristics of what they represent) and intimately linked to sensorimotor processing (see, e.g., Wilson 2002; Gładziejewski and Miłkowski 2017; and Williams and Colling 2018). Second, numerical cognition is regarded as the "prime example of embodied cognition," and spatial-numerical associations including the Att-SNARC phenomenon provide the key evidence for that claim (Fischer and Brugger 2011). As such, the Att-SNARC phenomenon has been used as evidence for embodied number representations and to support strong claims—dating from at least as early as Galton (1880)—about the link between number and space (e.g., a mental number line).

The Att-SNARC RRR involved 48 researchers and 1105 subjects across 17 labs and featured four dependent measures and four experimental conditions for a total of 16 variates. The four dependent measures were all reaction times assessed in milliseconds (ms) and averaged over 40 trials; they differed with respect to the time delay between the removal of the number and the appearance of the target, with delays of 250, 500, 750, and 1000 ms, respectively. The experimental conditions followed a two-by-two design, with the target appearing on the left or the right and the number being small or large. Each subject was assigned to all four of the experimental conditions, and all four dependent measures were assessed for each experimental condition for each subject. The effects of interest were the interaction effect for each dependent measure.

2.2. Model

Let i index subjects; v index variates; $y_{i,v}$ denote the observation for subject i and variate v; $l[i]$ denote the lab l at which subject i was observed; and $d[v]$ denote the dependent measure associated with variate v. Our model specification for the $y_{i,v}$ is given by

$$y_{i,v} = \alpha_v + \beta_{l[i],v} + \gamma_{i,v} + \varepsilon_{i,v} \tag{1}$$

where the α_v are treated as fixed effects for each variate; the $\beta_{l,v}$ are treated as random effects for each lab and variate; the $\gamma_{i,v}$ are treated as random effects for each subject and variate; and the $\varepsilon_{i,v}$ are random errors for each subject and variate.

Letting $\boldsymbol{\beta}_l$ denote the vector of $\beta_{l,v}$ for each lab and $\boldsymbol{\gamma}_i$ denote the vector of $\gamma_{i,v}$ for each subject, we assume that the $\boldsymbol{\beta}_l$ are independent and identically distributed according to the multivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{T}_{\boldsymbol{\beta}}$; the $\boldsymbol{\gamma}_i$ are independent and identically distributed according to the multivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{T}_{\boldsymbol{\gamma}}$; the $\varepsilon_{i,v}$ are independent and distributed according to the normal distribution with mean zero and variance $\sigma^2_{d[v]}$; and there is zero covariation among the $\boldsymbol{\beta}_l$, $\boldsymbol{\gamma}_i$, and $\varepsilon_{j,v}$ for all l, i, j, and v.

To model $\mathbf{T}_{\boldsymbol{\beta}}$ and $\mathbf{T}_{\boldsymbol{\gamma}}$, we employ a three-factor structure for each. The first factor has loadings that are constrained to be equal for all variates; the second factor has loadings that are constrained to be equal for all variates associated with a given dependent measure (i.e., the experimental conditions associated with the dependent measure) but that can vary across variates associated with different dependent measures; and the third factor has loadings that are unconstrained and thus can vary across all variates. Consequently, we refer to these, respectively, as the intercept factor, the dependent measure factor, and the variate factor.

To be specific, our model specification for the $\beta_{l,v}$ is given by

$$\beta_{l,v} = \varphi^{\beta}_{l,1} + \varphi^{\beta}_{l,2}\lambda^{\beta}_{1,d[v]} + \varphi^{\beta}_{l,3}\lambda^{\beta}_{2,v} + \eta^{\beta}_{l,v}$$

where the $\varphi^{\beta}_{l,\cdot}$ are factor scores for each lab; $\lambda^{\beta}_{1,d}$ is a factor loading for each dependent measure; $\lambda^{\beta}_{2,v}$ is a factor loading for each variate; and $\eta^{\beta}_{l,v}$ is an idiosyncratic term for each lab and variate. Letting $\boldsymbol{\varphi}^{\boldsymbol{\beta}}_l$ denote the vector of $\varphi^{\beta}_{l,\cdot}$ for each lab and $\boldsymbol{\eta}^{\boldsymbol{\beta}}_l$ denote the vector of $\eta^{\beta}_{l,v}$ for each lab, we assume that the $\boldsymbol{\varphi}^{\boldsymbol{\beta}}_l$ are independent and identically distributed according to the multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{\Omega}_{\boldsymbol{\varphi},\boldsymbol{\beta}}$; the $\boldsymbol{\eta}^{\boldsymbol{\beta}}_l$ are independent and identically distributed according to the multivariate normal distribution with mean zero and variance-covariance matrix $\boldsymbol{\Omega}_{\boldsymbol{\eta},\boldsymbol{\beta}}$, which allows covariance only across variates associated with a given dependent measure but not across variates associated with different dependent measures; and there is zero covariation among the $\boldsymbol{\varphi}^{\boldsymbol{\beta}}_l$ and $\boldsymbol{\eta}^{\boldsymbol{\beta}}_m$ for all l and m. Consequently, $\mathbf{T}_{\boldsymbol{\beta}} = \boldsymbol{\Lambda}_{\boldsymbol{\beta}}\boldsymbol{\Omega}_{\boldsymbol{\varphi},\boldsymbol{\beta}}\boldsymbol{\Lambda}^{\mathrm{T}}_{\boldsymbol{\beta}} + \boldsymbol{\Omega}_{\boldsymbol{\eta},\boldsymbol{\beta}}$ where $\boldsymbol{\Lambda}_{\boldsymbol{\beta}}$ is the matrix with rows $[1, \lambda^{\beta}_{1,d[v]}, \lambda^{\beta}_{2,v}]$.

Our model specification for the $\gamma_{i,v}$ is nearly identical to the above. Specifically, it is given by

$$\gamma_{i,v} = \varphi^{\gamma}_{i,1} + \varphi^{\gamma}_{i,2}\lambda^{\gamma}_{1,d[v]} + \varphi^{\gamma}_{i,3}\lambda^{\gamma}_{2,v} \tag{2}$$

where all is *mutatis mutandis* as above but the analogue of $\eta^{\beta}_{l,v}$ is omitted because variates were observed no more than once for each subject. Consequently, $\mathbf{T}_{\boldsymbol{\gamma}} = \boldsymbol{\Lambda}_{\boldsymbol{\gamma}}\boldsymbol{\Omega}_{\boldsymbol{\varphi},\boldsymbol{\gamma}}\boldsymbol{\Lambda}^{\mathrm{T}}_{\boldsymbol{\gamma}}$.

We estimate our model in Stan (Carpenter et al. 2017) using the default settings, the default weakly informative priors given in the Stan User's Guide (Stan Development Team 2020), and redundant parameterization with identifiability constraints (e.g., factor loadings to have zero mean and unit standard deviation) imposed *ex post* (McCulloch and Rossi 1994; Gelman et al. 2008). We resolve reflection invariance using the method of Erosheva and Curtis (2017).

2.3. Results

2.3.1. Principal Results

Our focus is on quantifying the variation and covariation in variates and effects at the lab and subject levels. The variation and covariation in variates is modeled by $\mathbf{T}_{\boldsymbol{\beta}}$ at the lab level and by $\mathbf{T}_{\boldsymbol{\gamma}}$ at the subject level, and the variation and covariation in effects is modeled by these respective matrices at each level in conjunction with the contrast matrix corresponding to the effects. Consequently, we focus our discussion of results on estimates of these quantities noting that we expect the variation and covariation in variates to be higher than that in effects due to the zero-sum nature of contrasts.

We begin by discussing the estimates of the (scaled) factor loadings at the lab level (i.e., the elements of $\boldsymbol{\Lambda}_{\boldsymbol{\beta}}$ scaled by the square root of the associated diagonal element of $\boldsymbol{\Omega}_{\boldsymbol{\varphi},\boldsymbol{\beta}}$ so as to be in ms units), which we present in Figure 1. First, the intercept factor loading estimate indicates common covariation of 16 ms across all variates and subjects within a given lab.[1] To put this in context, 16 ms is about 1.21 times the error standard deviation σ_d (the estimates of which range from 12 to 14 ms across the four dependent measures). Because the subject populations did not differ across the labs involved in the Att-SNARC RRR (all were university students), we believe this reflects lab differences in equipment. Second, the dependent measure factor loading estimates have a monotone nature, which indicates that the degree of the covariation between the variates associated with one dependent measure and the variates associated with another

[1] All estimates discussed in the text are posterior median estimates rounded to the nearest integer for those given in ms and ms^2 units and to two decimal places for those given in all other units.

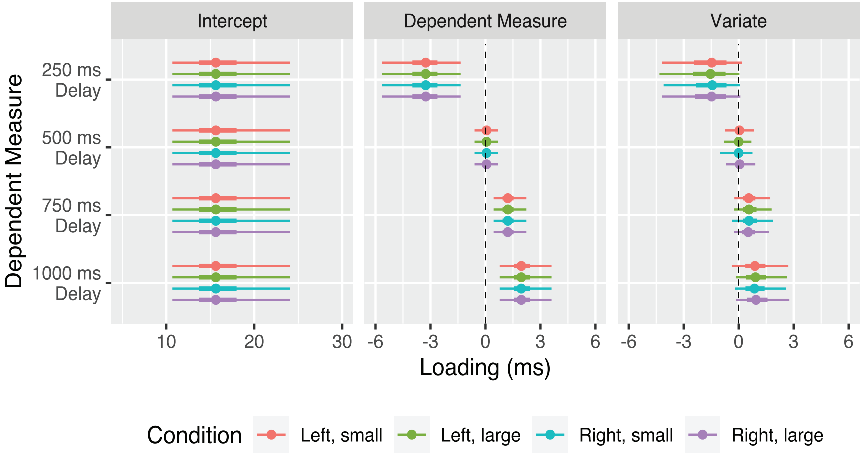

Figure 1. Att-SNARC RRR lab level factor loading estimates. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively.

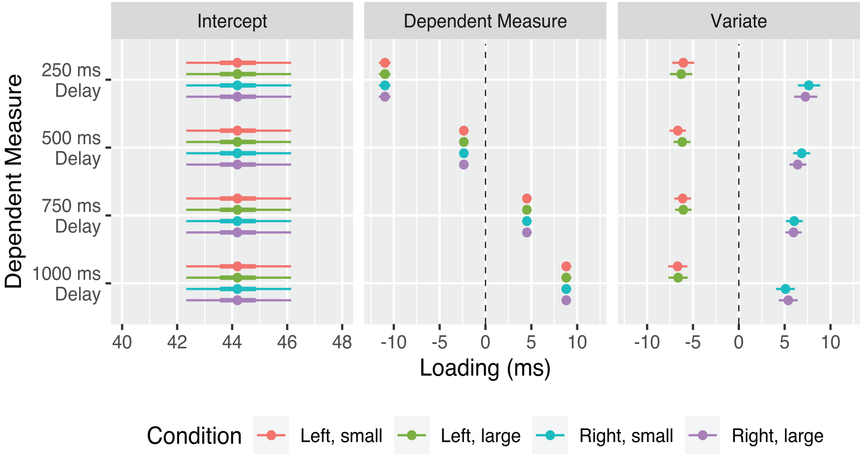

Figure 2. Att-SNARC RRR subject level factor loading estimates. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively.

dependent measure decreases as the magnitude of the difference of the time delays of the two dependent measures increases (i.e., because estimates of the same (different) sign indicate stronger (weaker) covariation). Third, the variate factor loading estimates are nearly identical for all variates associated with each dependent measure. Further, they vary across dependent measures in a manner that is similar to those of the dependent measure factor. Finally, they are very small in magnitude. These results indicate that this factor is not necessary here.

We now discuss the estimates of the variation and covariation in the idiosyncratic terms for each lab and variate (i.e., $\mathbf{\Omega}_{\eta,\beta}$). In short, these are trivial. Specifically, the estimates of the variation (i.e., the square roots of the diagonal elements of $\mathbf{\Omega}_{\eta,\beta}$) range from 0 to 2 ms across the 16 variates and the estimates of the covariation (i.e., the off-diagonal elements of $\mathbf{\Omega}_{\eta,\beta}$) range from 0 to 1 ms^2 across the pairs of variates.

We now discuss the estimates of the factor loadings at the subject level (i.e., $\mathbf{\Lambda}_\gamma$ scaled by $\mathbf{\Omega}_\gamma$ as done above at the lab level), which we present in Figure 2. First, the intercept factor loading estimate indicates common covariation of 44 ms across all variates for a given subject. To put this in context, 44 ms is about 3.41 times σ_d. We believe this reflects individual differences in reaction times (i.e., that some individuals generally react more quickly while others generally react more slowly). Second, the dependent measure factor loading estimates are again a monotone function of the dependent measures. Third, the variate factor loading estimates are nearly identical for the eight variates associated with experimental conditions with the target appearing on the left and so too for the eight variates associated with experimental conditions with the target appearing on the right; also, all 16 are nearly identical in magnitude. This indicates stronger (weaker) covariation in variates with targets appearing on the same (different) side. We believe this reflects individual differences in ocular dominance (i.e., eye preference or eyedness, the tendency to prefer visual input from one eye to the other; see Section 2.3.2).

We now discuss the estimates of the variation and covariation in variates at the lab and subject levels (i.e., $\mathbf{T}_\beta$ and $\mathbf{T}_\gamma$, respectively), which aggregate across the estimates discussed above. The estimates of the variation (i.e., the square roots of the diagonal elements of $\mathbf{T}_\beta$ and $\mathbf{T}_\gamma$), which we present in the left panel of Figure 3, are highly similar and very low at each level, ranging from 15 to 17 ms across the 16 variates at the lab level and from 44 to 48 ms across the 16 variates at the subject level. Further, because the intercept factor is the dominant factor at both levels (see Figures 1 and 2), the estimates of the covariation

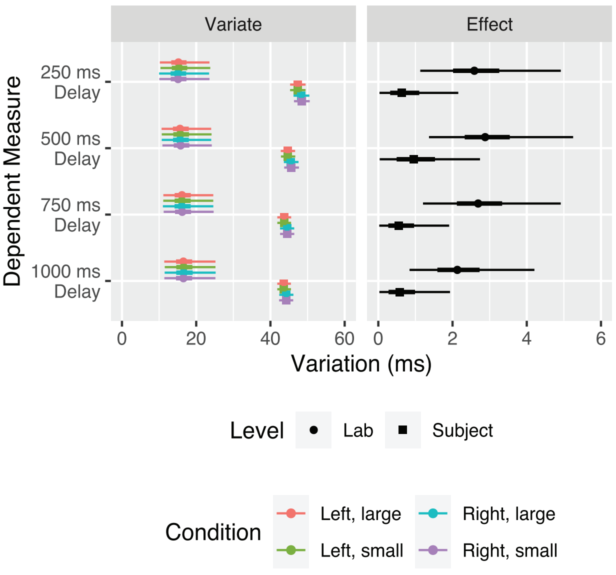

Figure 3. Att-SNARC RRR variation estimates. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively.

(i.e., the off-diagonal elements of $\mathbf{T}_{\beta}$ and $\mathbf{T}_{\gamma}$) are very high at each level, ranging from 225 to 274 ms^2 across the pairs of variates at the lab level and from 1824 to 2336 ms^2 across the pairs of variates at the subject level; this corresponds to estimates of the correlation which range from 0.91 to 1.00 across the pairs of variates at the lab level and from 0.86 to 1.00 across the pairs of variates at the subject level.

As noted above, the effects of interest in the Att-SNARC RRR were the interaction effect for each dependent measure. Thus, we now discuss the estimates of the variation and covariation in effects at the lab and subject levels (i.e., $\mathbf{C}\mathbf{T}_{\beta}\mathbf{C}^{\mathrm{T}}$ and $\mathbf{C}\mathbf{T}_{\gamma}\mathbf{C}^{\mathrm{T}}$, respectively, where $\mathbf{C}$ is the contrast matrix corresponding to the effects). The estimates of the variation (i.e., the square roots of the diagonal elements of $\mathbf{C}\mathbf{T}_{\beta}\mathbf{C}^{\mathrm{T}}$ and $\mathbf{C}\mathbf{T}_{\gamma}\mathbf{C}^{\mathrm{T}}$), which we present in the right panel of Figure 3, are highly similar and trivial at each level, ranging from 2 to 3 ms across the four effects at the lab level and all 1 ms for the four effects at the subject level. Given this trivial degree of variation, the estimates of the covariation (i.e., the off-diagonal elements of $\mathbf{C}\mathbf{T}_{\beta}\mathbf{C}^{\mathrm{T}}$ and $\mathbf{C}\mathbf{T}_{\gamma}\mathbf{C}^{\mathrm{T}}$) are also trivial, all 0 ms^2 for the pairs of effects at the lab level and all 0 ms^2 for the pairs of effects at the subject level.

2.3.2. Ocular Dominance Results

As noted above in our discussion of Figure 2, the variate factor loading estimates at the subject level appear to reflect individual differences in ocular dominance. To probe this would require an assessment of ocular dominance. While ocular dominance was not assessed in the Att-SNARC RRR, handedness—a proxy for ocular dominance (Bourassa, McManus, and Bryden 1996)—was. This suggests expanding the model to include handedness.

We do so by replacing the variate factor loading $\lambda^{\gamma}_{2,v}$ in Equation (2) with $x_{i,v}$, which is defined to be one for left-handed subjects for variates associated with experimental conditions with the target appearing on the left and for right-handed subjects for variates associated with experimental conditions with the target appearing on the right (i.e., subject handedness and variate target-side match) and negative one for left-handed subjects for variates associated with experimental conditions with the target appearing on the right and for right-handed subjects for variates associated with experimental conditions with the target appearing on the left (i.e., subject handedness and variate target-side do not match). We also, as is customary, add $\alpha x_{i,v}$ to Equation (1).

We note that $x_{i,v}$ in Equation (2) can be understood as a variate factor that has loadings that are fully constrained, namely to be equal to negative (positive) one for variates associated with experimental conditions with the target appearing on the left (right) for right-handed subjects and vice versa for left-handed subjects. With that understanding, we now discuss the estimates of the factor loadings at the subject level (i.e., $\mathbf{\Lambda}_{\gamma}$ scaled by $\mathbf{\Omega}_{\gamma}$ as done above), which we present in Figure 4. The estimates are remarkably similar to those presented in Figure 2 thereby supporting the notion that the variate factor loading estimates in that original figure reflect individual differences in ocular dominance.

We note that given this similarity, it is unsurprising that estimates of the variation and covariation in variates and effects also remain similar to those discussed above. Specifically, the

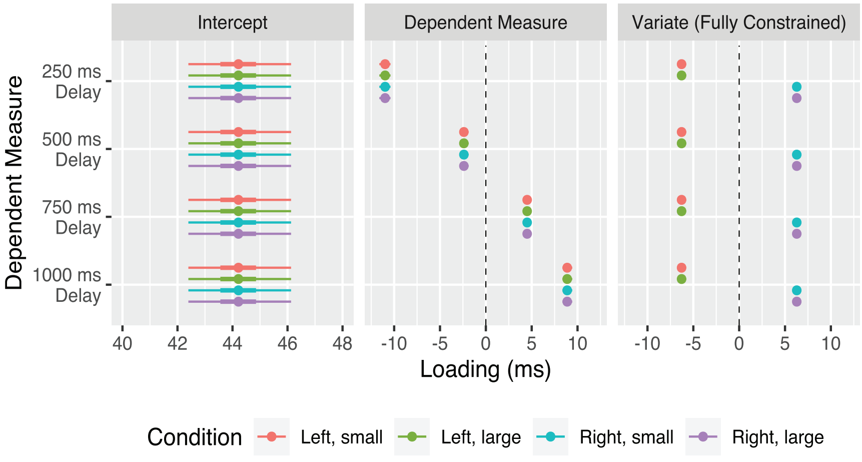

Figure 4. Att-SNARC RRR subject level factor loading estimates with fully constrained variate factor loadings. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively.

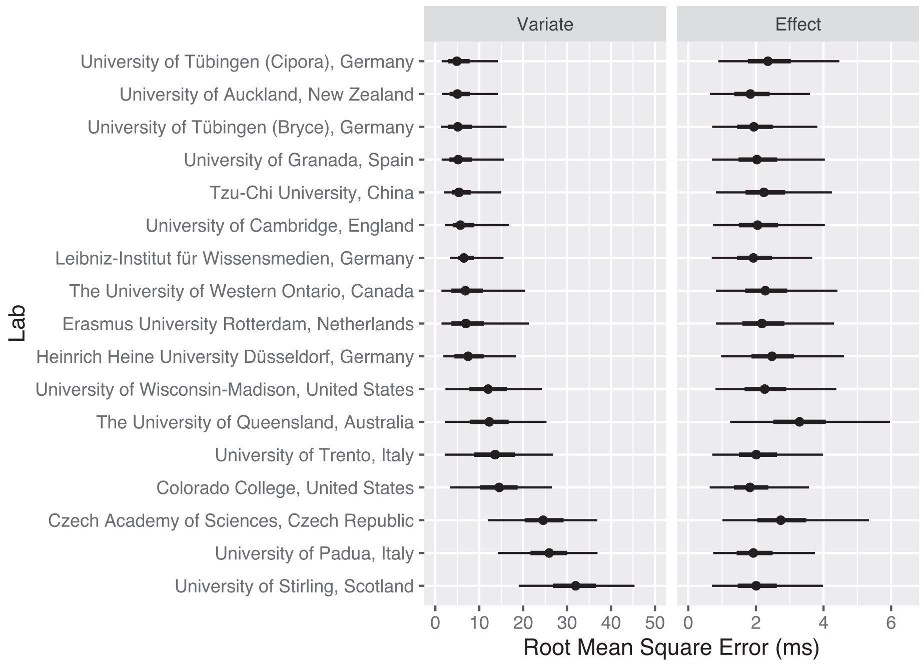

Figure 5. Att-SNARC RRR lab atypicality estimates. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively. Labs are sorted by the posterior median estimate of the variate atypicality.

analogue of the estimates of the factor loadings at the lab level remain similar to those presented in Figure 1 and the analogue of the estimates of the variation in variates and effects at the lab and subject levels remain similar to those presented in Figure 3. We also note that the estimate of α is -1 ms indicating that reaction times are trivially quicker when subject handedness and variate target-side match.

2.3.3. *Evaluation of Replicability*

The estimates of the variation in variates and effects at the lab level presented in Figure 3 provide a multi-study, continuous, multi-faceted evaluation of replicability that indicates a very high degree of replicability for all variates and effects. Specifically, the estimates of the variation in variates at the lab level are very low ranging from, as noted above, 15 to 17 ms across the 16 variates; further, because the estimates of the covariation in variates at the lab level are very high, the low variation is likely to be driven by a common cause such as lab differences in equipment as suggested above. Additionally, the estimates of the variation in effects at the lab level are trivial ranging from, as noted above, 2 to 3 ms across the four effects.

To further evaluate the variation in the estimates of variates and effects across labs, we examine the atypicality of the estimates from each lab. We estimate this atypicality by computing the root mean square error of the lab level estimates of the variates (i.e., the $\beta_{l,v}$) and the effects (i.e., contrasts of the $\beta_{l,v}$), which we present in Figure 5. The estimates suggest that no lab is atypical in terms of either variates or effects.

3. MLP

3.1. *Description*

The MLP is a large-scale replication of 13 phenomena from social psychology and behavioral economics. We present 12 of these phenomena along with their associated dependent measures and experimental conditions in Table 1.[2] These phenomena are among some of the most important in these fields. For example, the Currency Priming and Flag Priming phenomena are examples of social priming, one of the most prominent subdomains of social psychology. Similarly, the Sunk Costs phenomenon and the concomitant fallacy are foundational to rational decision-making in classical economics. Finally, the 2002 Sveriges Riksbank Prize was awarded for, among other things, work on the Gain versus Loss Framing and Anchoring phenomena.

The MLP involved 51 researchers and 6344 subjects across 36 labs and featured 15 dependent measures and two experimental conditions per dependent measure for a total of 30 variates. However, because the MLP investigated many phenomena whereas the Att-SNARC RRR investigated only a single phenomenon, there are several differences in the dependent measures and experimental conditions in the MLP as compared to those in the Att-SNARC RRR. First, whereas the dependent measures in the Att-SNARC RRR were conceptually similar and assessed in the same units, the dependent measures in the MLP were conceptually distinct and were assessed in a variety of units on a variety of scales, including four binary (i.e., the dependent measures associated with the Gain versus Loss Framing, Low versus High Category Scales, Allowed versus Forbidden, and Norm of Reciprocity phenomena) and two ordinal (i.e., the dependent measures associated with the Sunk Costs and Quote Attribution phenomena). Second, whereas the experimental conditions were the same across the dependent measures

[2] The MLP reused the dependent measure associated with the Sex Differences in Implicit Math Attitudes phenomenon in an analysis of a thirteenth phenomenon, Relations Between Implicit and Explicit Math Attitudes, that had no experimental condition associated with it; we do not consider that phenomenon here.

Table 1. MLP phenomena, dependent measures, and experimental conditions.

Phenomenon	Dependent measure	Experimental conditions
Gain versus Loss Framing	Binary choice of deterministic versus stochastic option	People will die, people will be saved
Retrospective Gambler Fallacy	Estimate of how many times a man had rolled dice	Two sixes, three sixes
Sex Differences in Implicit Math Attitudes	Implicit Association Test of attitudes toward math compared to arts	Female, male
Sunk Costs	Likelihood of attending a football game on an integer scale ranging from one to nine	Free, paid
Quote Attribution	Agreement with a quotation on an integer scale ranging from one to nine	Liked source, disliked source
Low versus High Category Scales	Binary coded report of daily television watching (greater than versus less than two and a half hours)	Low category scale, high category scale
Allowed versus Forbidden	Binary choice of whether the local country should allow versus forbid speeches against democracy	Forbidden, allowed
Currency Priming	Eight item system justification scale	No prime, money prime
Imagined Contact	Four item scale indicating interest and willingness to interact with Muslims	Contact, control
Norm of Reciprocity	Binary choice of whether the local country should allow versus forbid North Korean newspapers to come in and send back the news as they see it	Asked second, asked first
Flag Priming	Eight item questionnaire assessing views toward various political issues (e.g., abortion, gun control, affirmative action)	No prime, flag prime
Anchoring	Distance from San Francisco to New York City Number of babies born per day in the United States Population of Chicago Height of Mount Everest	High anchor, low anchor

NOTE: Phenomena are sorted as discussed in the caption to Figure 8.

in the Att-SNARC RRR, the experimental conditions varied—and were also conceptually distinct—across the dependent measures in the MLP. That said, four of the dependent measures in the MLP were repeated measures of the same phenomenon (i.e., Anchoring). Further, these four dependent measures were conceptually similar (although assessed in different units) and the experimental conditions were the same across them as in the Att-SNARC RRR.

There are two additional differences between the MLP and the Att-SNARC RRR. First, whereas each subject was assigned to all four of the experimental conditions and all four dependent measures were assessed for each experimental condition for each subject in the Att-SNARC RRR, each subject was randomly assigned to only one of the two experimental conditions associated with each dependent measure and thus the dependent measure was assessed for only this experimental condition for the subject in the MLP.[3] Second, whereas the effects of interest were the interaction effect for each dependent measure in the Att-SNARC RRR, the effects of interest were the simple effect for each dependent measure in the MLP.

3.2. Model

We extend our model specification for the Att-SNARC RRR to accommodate two prominent characteristics of the MLP, namely that (a) the dependent measures were assessed in a variety of units on a variety of scales and (b) four of the dependent measures were repeated measures of the same phenomenon (i.e., Anchoring).

To accommodate the variety of units, we standardize by σ_d; to accommodate the variety of scales, we employ the generalized linear model. Specifically, we introduce $y^\star_{i,v}$ and let $y_{i,v} = y^\star_{i,v}$ if the dependent measure associated with variate v is treated as continuous; $y_{i,v} = \mathbf{1}(y^\star_{i,v} > 0)$ and $\sigma^2_{d[v]} = 1$ if the dependent measure associated with variate v is binary; and $y_{i,v} = k$ if $c_{d[v],k-1} < y^\star_{i,v} \leq c_{d[v],k}$, $\sigma^2_{d[v]} = 1$, $c_{d[v],0} = -\infty$, and $c_{d[v],K_{d[v]}} = \infty$ where $K_{d[v]}$ is the maximum possible value of $y_{i,v}$ and the $c_{d,k}$ are treated as fixed effects for each dependent measure and value if the dependent measure associated with variate v is ordinal (i.e., binary and ordinal dependent measures are modeled according to the binary and ordinal probit specifications, respectively). Our model specification for the $y^\star_{i,v}$ is given by

$$y^\star_{i,v} = \sigma_{d[v]}(\alpha_v + \beta_{l[i],v} + \gamma_{i,v}) + \varepsilon_{i,v} \tag{3}$$

where everything is treated as in the Att-SNARC RRR model specification. We employ the two-parameter Box-Cox transformation for dependent measures treated as continuous (Box and Cox 1964). We also align the dependent measures by reverse-coding those associated with the Gain versus Loss Framing, Sex Differences in Implicit Math Attitudes, Quote Attribution, Imagined Contact, Norm of Reciprocity, and Anchoring phenomena.

To accommodate the repeated measures of the Anchoring phenomenon, we introduce additional factors. A natural manner of doing so would be to add to the three-factor structure used in the Att-SNARC RRR an analogous three-factor structure for the variates associated with the Anchoring phenomenon. This would imply employing a factor structure consisting of the baseline three factors that apply to all 30 variates (i.e., the intercept factor, the dependent measure factor, and the variate factor) plus an additional three factors that apply to only the

[3] There was one exception: the experimental condition for the dependent measure associated with the Sex Differences in Implicit Math Attitudes phenomenon was sex, which was of course not randomly assigned. We also note that the order in which the phenomena were presented was randomized, with the exception that the Sex Differences in Implicit Math Attitudes phenomenon was always presented last, and that the four dependent measures associated with the Anchoring phenomenon were always presented in the order distance from San Francisco to New York City, population of Chicago, height of Mount Everest, number of babies born per day in the United States with subjects randomly assigned to the high anchor or low anchor condition separately for each of the four dependent measures.

eight variates associated with the Anchoring phenomenon (i.e., an Anchoring intercept factor, an Anchoring dependent measure factor, and an Anchoring variate factor).

However, because the experimental conditions are the same across the four dependent measures associated with the Anchoring phenomenon, the Anchoring intercept factor can be extended so that the loadings vary across the experimental conditions. Therefore, we employ a factor structure consisting of the baseline three factors that apply to all 30 variates plus an additional four factors that apply to only the eight variates associated with the Anchoring phenomenon (i.e., a high anchor condition Anchoring intercept factor, a low anchor condition Anchoring intercept factor, an Anchoring dependent measure factor, and an Anchoring variate factor) at the lab level. Specifically, our model specification for the $\beta_{l,v}$ is given by

$$\beta_{l,v} = \varphi^{\beta}_{l,1} + \varphi^{\beta}_{l,2}\lambda^{\beta}_{1,d[v]} + \varphi^{\beta}_{l,3}\lambda^{\beta}_{2,v} + \big(\varphi^{\beta}_{l,4}\mathbf{1}(c[v]=1) \\ + \varphi^{\beta}_{l,5}\mathbf{1}(c[v]=2) + \varphi^{\beta}_{l,6}\lambda^{\beta}_{3,d[v]} + \varphi^{\beta}_{l,7}\lambda^{\beta}_{4,v}\big) \cdot \mathbf{1}_{\mathbf{A}}(v) + \eta^{\beta}_{l,v}$$

where $\mathbf{1}(x)$ is defined to be one if x is true and zero otherwise; $c[v]$ denotes the experimental condition associated with variate v; and $\mathbf{1}_{\mathbf{A}}(v)$ is defined to be one if variate v is associated with the Anchoring phenomenon and zero otherwise.

This factor structure is not identified at the subject level. Specifically, the additional four factors that apply to only the eight variates associated with the Anchoring phenomenon are not identified because there are only four observations of the phenomenon per subject (i.e., one for each of the four dependent measures associated with the phenomenon). Therefore, we instead employ a factor structure consisting of the baseline three factors that apply to all 30 variates plus an additional two factors that apply to only the eight variates associated with the Anchoring phenomenon (i.e., an Anchoring intercept factor and an Anchoring variate factor) at the subject level. Specifically, our model specification for the $\gamma_{i,v}$ is given by

$$\gamma_{i,v} = \varphi^{\gamma}_{i,1} + \varphi^{\gamma}_{i,2}\lambda^{\gamma}_{1,d[v]} + \varphi^{\gamma}_{i,3}\lambda^{\gamma}_{2,v} + (\varphi^{\gamma}_{i,4} + \varphi^{\gamma}_{i,5}\lambda^{\gamma}_{3,v}) \cdot \mathbf{1}_{\mathbf{A}}(v).$$

We estimate our model as in the Att-SNARC RRR.

3.3. Results

3.3.1. Principal Results

Because our focus is on quantifying the variation and covariation in variates and effects at the lab and subject levels, we focus our discussion of results on estimates of these quantities. We begin by discussing the estimates of the (scaled) factor loadings at the lab level (i.e., the elements of $\mathbf{\Lambda}_{\beta}$ scaled by the square root of the associated diagonal element of $\mathbf{\Omega}_{\varphi,\beta}$ so as to be in error standard deviation σ_d units), which we present in Figure 6. First, the intercept factor loading estimate indicates common covariation of 0.09 σ_d across all variates and subjects within a given lab; this is surprisingly large given that one of the criteria on which the phenomena investigated by the MLP were chosen was diversity and suggests that the variates associated with these phenomena may not in fact be so diverse. Second, the dependent measure factor loading estimates are most prominent for the dependent measures associated with the Currency Priming, Norm of Reciprocity, and Flag Priming phenomena; notably, all three of these dependent measures relate to political ideology (see Table 1 and Section 3.3.2). Third, the variate factor loading estimates show no clear pattern; this is perhaps not unsurprising given that the dependent measures and experimental conditions associated with the variates are conceptually distinct. Fourth, the Anchoring intercept factor loading estimates indicates additional common covariation of 0.08 (high anchor condition) and 0.09 (low anchor condition) σ_d across the variates associated with the Anchoring phenomenon and subjects within a given lab. Fifth, the Anchoring dependent measure factor loading estimates are all very small in magnitude thereby indicating this factor is not necessary here; insofar as the four dependent measures associated with the Anchoring phenomenon are in fact exchangeable measures of it, this is unsurprising because the Anchoring intercept factors would account for any common covariation. Sixth, the Anchoring variate factor loading estimates differ in a manner that is consistent with the experimental condition and those for the variates associated with the dependent measure regarding the height of Mount Everest are largest in magnitude.

We now discuss the estimates of the factor loadings at the subject level (i.e., $\mathbf{\Lambda}_{\gamma}$ scaled by $\mathbf{\Omega}_{\gamma}$ as done above at the lab level), which we present in Figure 7. First, the intercept factor loading estimate indicates common covariation of 0.24 σ_d across all variates for a given subject; again, this is surprisingly large given that one of the criteria on which the phenomena investigated by the MLP were chosen was diversity and suggests that the variates associated with these phenomena may not in fact be so diverse. It may also suggest individual differences in response styles. Second, the dependent measure factor loading estimates are entirely dominated by the dependent measure associated with the Flag Priming phenomenon; as noted above, this dependent measure relates to political ideology. Third, the variate factor loading estimates show no clear pattern although they are most prominent for the variates associated with the Imagined Contact, Norm of Reciprocity, and Flag Priming phenomena; notably, the dependent measures associated with all three of these phenomena relate to political ideology. Fourth, the Anchoring intercept factor loading estimate indicates additional common covariation of 0.32 σ_d across the variates associated with the Anchoring phenomenon for a given subject. Fifth, the Anchoring variate factor loading estimates differ in a manner that is consistent with the experimental condition and those for the variates associated with the dependent measure regarding the height of Mount Everest are largest in magnitude.

We now discuss the estimates of the variation and covariation in variates at the lab and subject levels (i.e., $\mathbf{T}_{\beta}$ and $\mathbf{T}_{\gamma}$, respectively). The estimates of the variation (i.e., the square roots of the diagonal elements of $\mathbf{T}_{\beta}$ and $\mathbf{T}_{\gamma}$), which we present in the left panel of Figure 8, are medium to high at each level, ranging from 0.11 to 0.69 σ_d across the 30 variates with a median of 0.24 σ_d at the lab level and ranging from 0.13 to 1.11 σ_d across the 30 variates with a median of 0.36 σ_d at the subject level. The estimates of the covariation (i.e., the off-diagonal elements of $\mathbf{T}_{\beta}$ and $\mathbf{T}_{\gamma}$) are also medium to high across the pairs of variates at each level. For the variates not associated with the Anchoring phenomenon, this covariation is driven by the fact that many of the factor loading estimates that were most prominent were for variates associated with dependent measures that

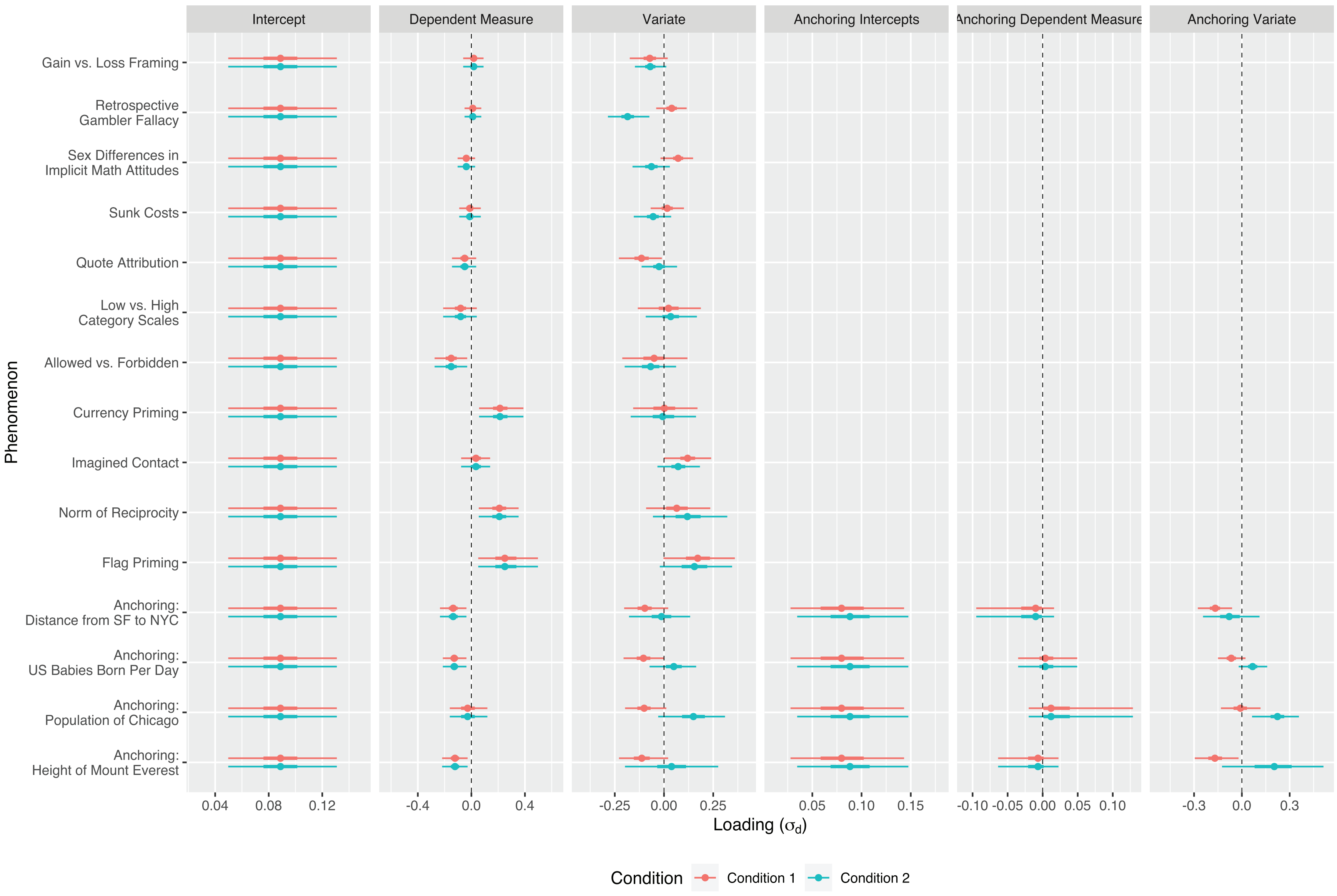

Figure 6. MLP lab level factor loading estimates. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively. Phenomena are sorted as discussed in the caption to Figure 8. Experimental conditions are as given in Table 1.

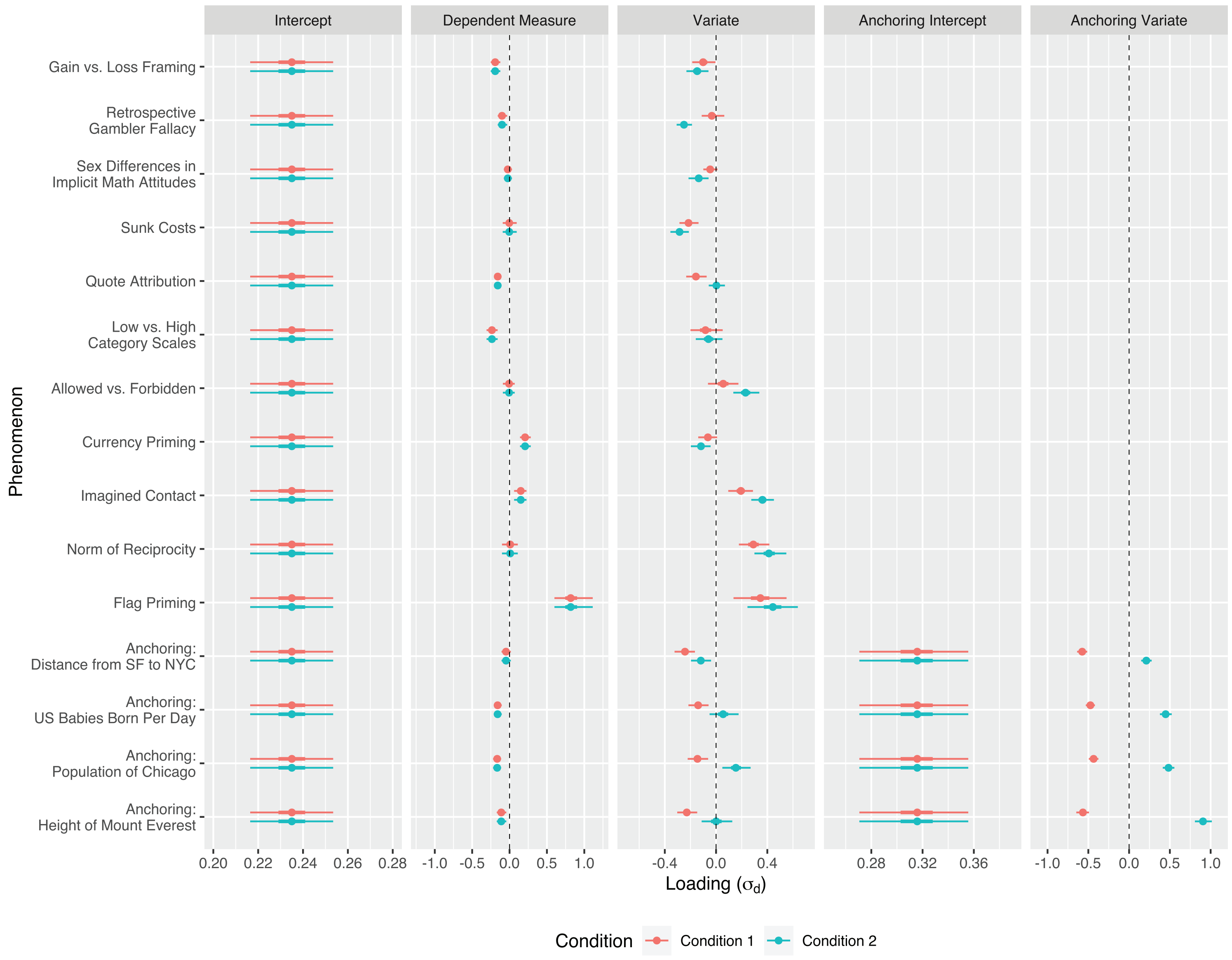

Figure 7. MLP subject level factor loading estimates. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively. Phenomena are sorted as discussed in the caption to Figure 8. Experimental conditions are as given in Table 1.

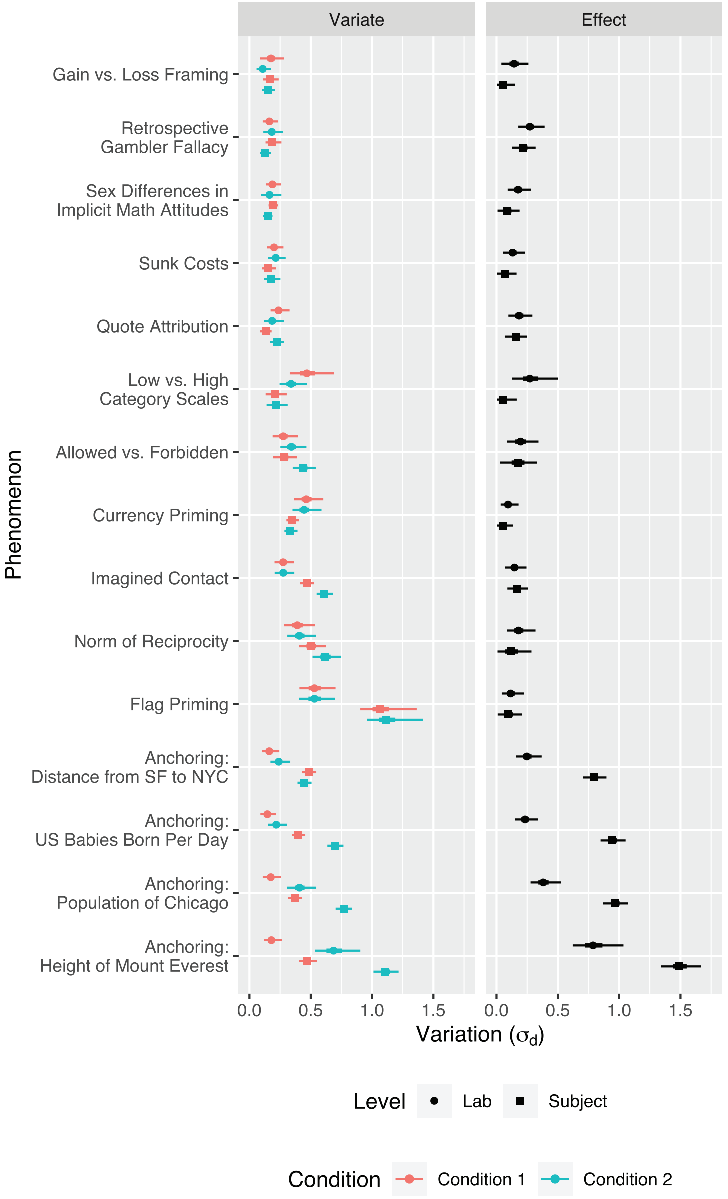

Figure 8. MLP variation estimates. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively. Phenomena other than the Anchoring phenomenon and the dependent measures associated with the Anchoring phenomenon are separately sorted by the posterior median estimate of the variation summed across the associated variates (i.e., across levels and experimental conditions). Experimental conditions are as given in Table 1.

relate to political ideology. For the variates associated with the Anchoring phenomenon, this covariation is of course driven by the Anchoring factor loading estimates.

As noted above, the effects of interest in the MLP were the simple effect for each dependent measure. Thus, we now discuss the estimates of the variation and covariation in effects at the lab and subject levels (i.e., $\mathbf{C}\mathbf{T}_\beta\mathbf{C}^\mathrm{T}$ and $\mathbf{C}\mathbf{T}_\gamma\mathbf{C}^\mathrm{T}$, respectively, where $\mathbf{C}$ is the contrast matrix corresponding to the effects). The estimates of the variation (i.e., the square roots of the diagonal elements of $\mathbf{C}\mathbf{T}_\beta\mathbf{C}^\mathrm{T}$ and $\mathbf{C}\mathbf{T}_\gamma\mathbf{C}^\mathrm{T}$), which we present in the right panel of Figure 8, are medium to high ranging from 0.09 to 0.79 σ_d across the 15 effects with a median of 0.18 σ_d at the lab level and low to high ranging from 0.05 to 1.49 σ_d across the 15 effects with a median of 0.16 σ_d at the subject level. The estimates of the covariation (i.e., the off-diagonal elements of $\mathbf{C}\mathbf{T}_\beta\mathbf{C}^\mathrm{T}$ and $\mathbf{C}\mathbf{T}_\gamma\mathbf{C}^\mathrm{T}$) are for the most part relatively low across the pairs of effects at each level, except among the four effects associated with the Anchoring phenomenon for which the estimates were high at each level.

3.3.2. Political Ideology Results

As noted above in our discussions of Figures 6 and 7, many of the factor loading estimates at the lab and subject levels that were most prominent were for variates associated with dependent measures that relate to political ideology. For example, at the lab level, the dependent measure factor loading estimates were most prominent for the dependent measures associated with the Currency Priming, Norm of Reciprocity, and Flag Priming phenomena. The dependent measure associated with the Currency Priming phenomenon is the eight item system justification scale which "measur[es] perceptions of the fairness, legitimacy, and justifiability of the prevailing social system" (Kay and Jost 2003); the dependent measure associated with Norm of Reciprocity phenomenon relates to freedom of the press; and the dependent measure associated with Flag Priming phenomenon is an eight item questionnaire assessing views toward various political issues (e.g., abortion, gun control, affirmative action).

Further, even several variates with factor loading estimates that were not prominent for any factor at any level also are associated with dependent measures that relate to political ideology, for example, the dependent measures associated with the Quote Attribution and Allowed versus Forbidden phenomena. The dependent measure associated with the Quote Attribution phenomenon assesses agreement with the quotation "I have sworn to only live free. Even if I find bitter the taste of death, I don't want to die humiliated or deceived" of Osama bin Laden when it is attributed to bin Laden (disliked source) or George Washington (liked source), and the dependent measure associated with the Allowed versus Forbidden phenomenon relates to freedom of speech.

Conveniently in light of this, political ideology was assessed in the MLP, specifically via a single item assessed on an integer scale ranging from one (strongly liberal) to seven (strongly conservative) as part of a six item demographic survey administered at the end of the MLP. The facts that (a) many of the factor loading estimates at the lab and subject levels that were most prominent were for variates associated with dependent measures that relate to political ideology and (b) several variates with factor loading estimates that were not prominent for any factor at any level also are associated with dependent measures that relate to political ideology suggest expanding the model to include political ideology.

We do so by replacing α_v in Equation (3) with $\alpha_{v,0} + \alpha_{v,1}x_i$ where x_i is the political ideology of subject i. This allows political ideology to associate not only with the variates but also the effects (i.e., because the effects in the MLP were the simple effect $\alpha_v - \alpha_{v'}$ for each dependent measure where v and v' are the two variates associated with the dependent measure, it allows political ideology to associate with the effects via $(\alpha_{v,0} - \alpha_{v',0}) + (\alpha_{v,1} - \alpha_{v',1})x_i$).

We now discuss the estimates of the association of political ideology with the variates and effects, which we present in Figure 9. We note that because of the presentation in σ_d units, the estimates are not meaningful in an absolute sense; however, they are meaningful in a relative sense, for example, by comparing the values for the two variates associated with a given dependent measure at a given level of political ideology. Broadly speaking, political ideology associates with many of the variates—sometimes more strongly than the experimental manipulation and sometimes interacting with (i.e., moderating) it. For example, the association of political ideology and the variates associated with the Currency Priming, Imagine Contact, and Flag Priming phenomena dwarfs that of the experimental manipulation. The association of political ideology and the variates associated with the Sex Differences in Implicit Math Attitudes, Sunk Costs, and Allowed versus Forbidden phenomena is comparable to that of the experimental manipulation. Further, political ideology appears to associate not only with the Quote Attribution variates but also with the Quote Attribution effect (i.e., political ideology appears to interact with (i.e., moderate) the Quote Attribution experimental manipulation). Finally, political ideology appears to have very little association with the variates associated with the Gain versus Loss Framing and Anchoring phenomena.

We note that expanding the model to include political ideology also alters the estimates of the variation and covariation in variates and effects. Specifically, the analogue of the estimates of the factor loadings at the lab level remain similar to those presented in Figure 6; this is unsurprising because political ideology is assessed at the subject level rather than the lab level. However, the analogue of the estimates of the factor loadings at the subject level change substantially from those presented in Figure 7; most notably, the intercept factor loading estimate decreases from 0.24 to 0.19 σ_d and the Flag Priming dependent measure factor loading estimate decreases from 0.82 to 0.03 σ_d. Finally, the analogue of the estimates of the variation in variates and effects at the lab and subject levels remain similar to those presented in Figure 8 with one major exception: the estimates of the variation in the variates associated with the Flag Priming phenomenon decrease from 0.53 and 0.53 to 0.34 and 0.30 σ_d, respectively, at the lab level and from 1.07 and 1.11 to 0.52 and 0.57 σ_d, respectively, at the subject level.

3.3.3. Anchoring Results

As noted above in our discussion of Figure 8, the estimates of the variation and covariation in the eight variates and the four effects associated with the Anchoring phenomenon were high at both the lab and subject levels, and this arose in large part from the estimates of the Anchoring factor loadings presented in Figures 6 and 7. Further, as noted above, the Anchoring phenomenon is unique among the phenomena investigated by the MLP because it is the only one with repeated measures. However, it is also unique for another reason: it is the only phenomenon for which there is a correct response (e.g., the height of Mount Everest is 29,029 feet) and thus for which the experimental manipulation should be entirely ineffective for subjects who know this correct response. Both of these facts likely contributed to the magnitude of the estimates presented in Figures 6–8.

The magnitude of these estimates prompted us to closely examine the study materials and data pertaining to the Anchoring phenomenon which in turn suggested three insights that relate back to the magnitude of these estimates. First, for three of the four dependent measures (all but that regarding the distance from San Francisco to New York City), the numeric value given in the low anchor condition was much farther from the correct response than that given in the high anchor condition. This is reflected in the fact that the estimates of the variation in the variates associated with the low anchor condition and these three dependent measures at the lab and subject levels presented in Figure 8 far exceed those associated with the high anchor condition.

Second, and unsurprisingly, there are individual differences in knowledge (i.e., some subjects know the correct response for one or more of the dependent measures while others do not). This is reflected in the magnitude of the Anchoring intercept and Anchoring variate factor loading estimates presented in Figure 7 and of the estimates of the variation in variates and effects at the subject level presented in in Figure 8.

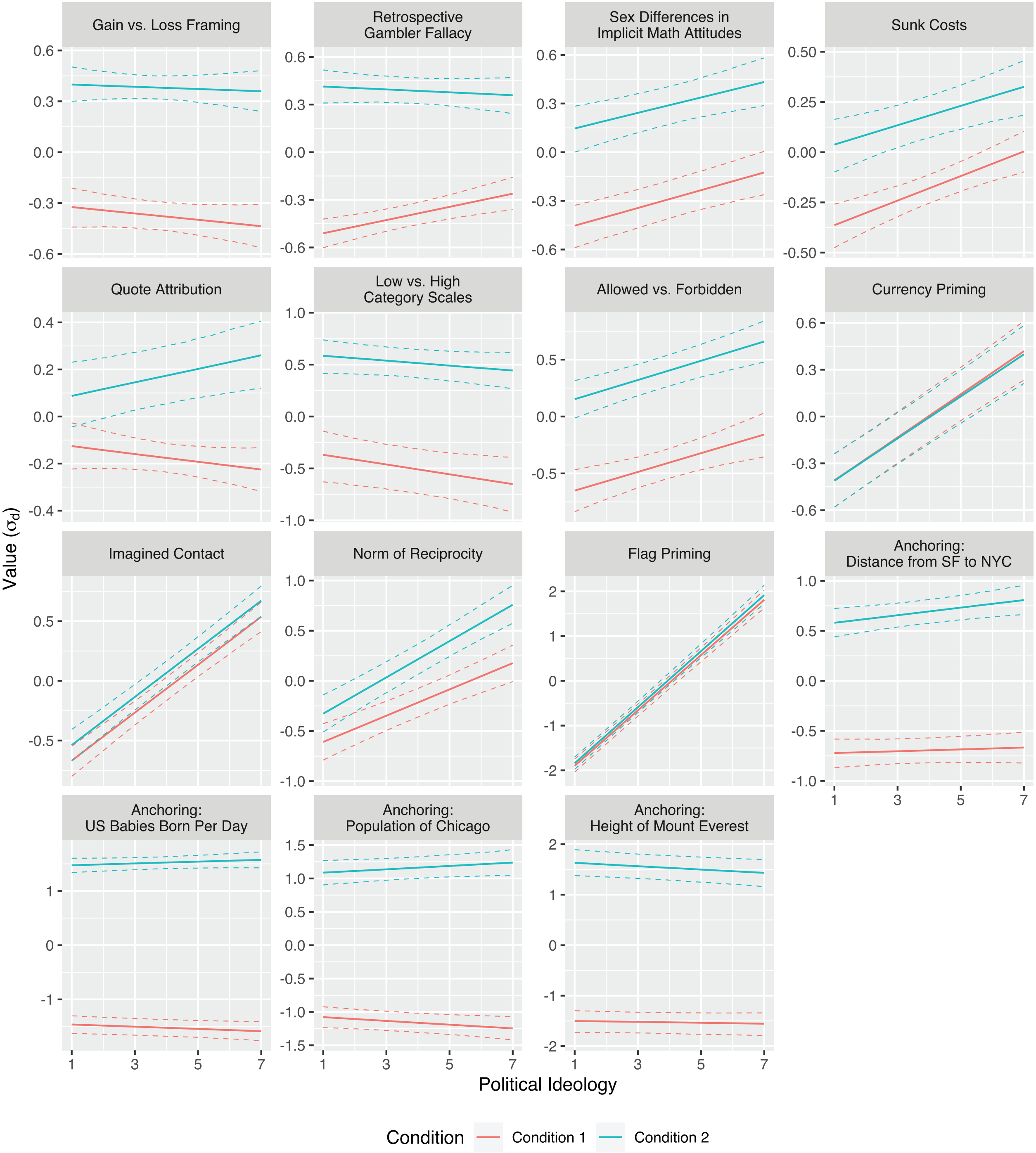

Figure 9. MLP political ideology estimates. Posterior median estimates are given by the solid lines; 95% equal-tailed posterior interval estimates are given by the dashed lines. Phenomena are sorted as discussed in the caption to Figure 8. Experimental conditions are as given in Table 1.

Third, and perhaps surprisingly, there are lab differences in knowledge, especially for the dependent measure regarding the height of Mount Everest. This is reflected in the magnitude of the Anchoring intercept and Anchoring variate factor loading estimates presented in Figure 6 and of the estimates of the variation in variates and effects at the lab level presented in Figure 8, especially those for the dependent measure regarding the height of Mount Everest.

3.3.4. Evaluation of Replicability

The estimates of the variation in variates and effects at the lab level presented in Figure 8 provide a multi-study, continuous, multi-faceted evaluation of replicability that indicates a medium to low degree of replicability depending on the variate or effect. Specifically, the estimates of the variation in variates at the lab level are medium to high ranging from, as noted above, 0.11 to 0.69 σ_d across the 30 variates with a median of 0.24 σ_d. Additionally, the estimates of the variation in effects at the lab level are medium to high ranging from, as noted above, 0.09 to 0.79 σ_d across the 15 effects with a median of 0.18 σ_d.

To further evaluate the variation in the estimates of variates and effects across labs, we examine the atypicality of the estimates from each lab. We estimate this atypicality by computing the root mean square error of the lab level estimates of the

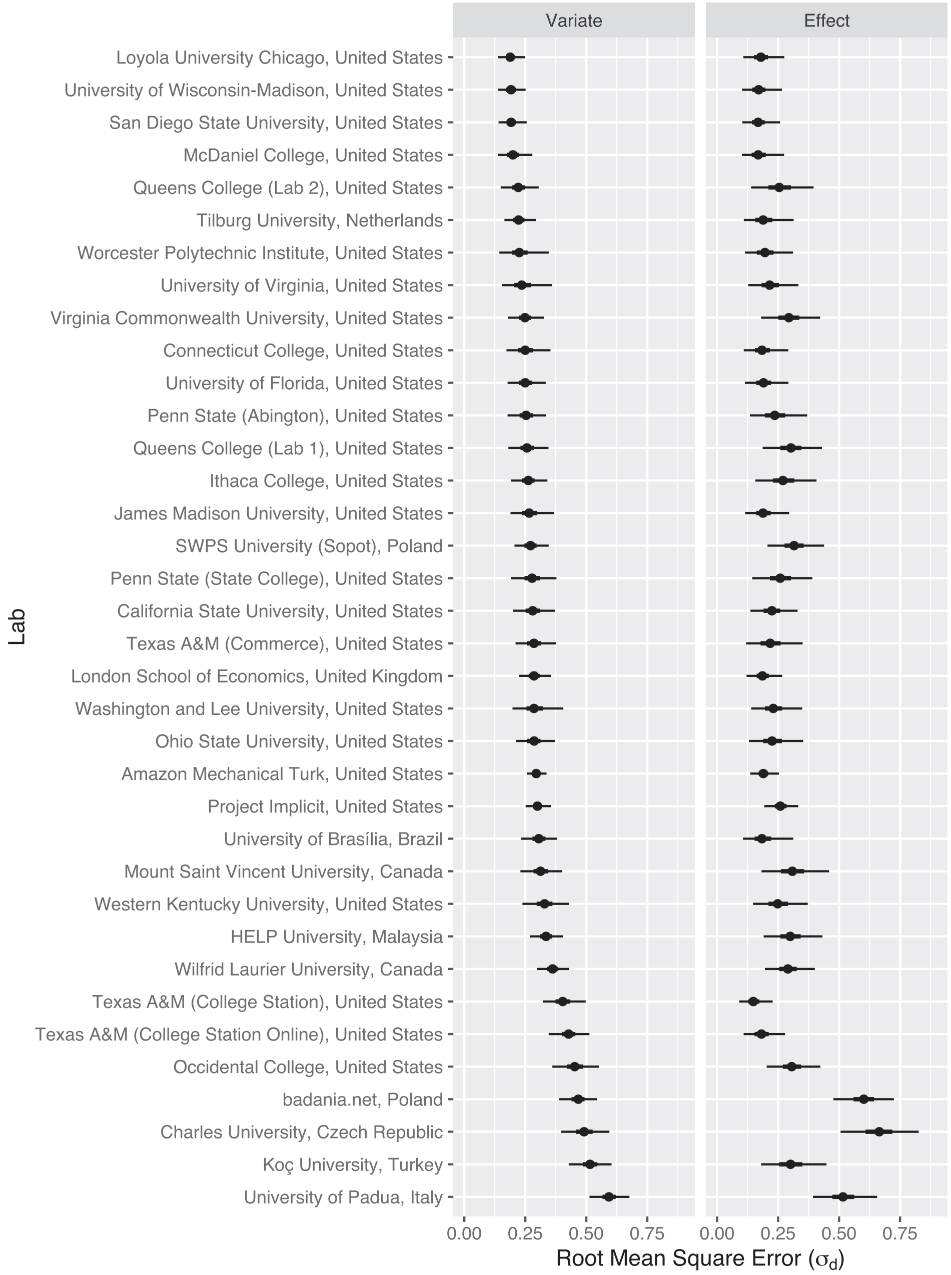

Figure 10. MLP lab atypicality estimates. Posterior median estimates are given by the points; 50% and 95% equal-tailed posterior interval estimates are given by the thick and thin lines, respectively. Labs are sorted by the posterior median estimate of the variate atypicality.

variates (i.e., the $\beta_{l,v}$) and the effects (i.e., contrasts of the $\beta_{l,v}$), which we present in Figure 10. The estimates suggest that no lab is atypical in terms variates but that badania.net, Charles University, and the University of Padua are somewhat atypical in terms of effects; these three labs are also among the four most atypical in terms of the estimates of variates.

4. Discussion

4.1. *Evaluations of Replicability*

In large-scale replication projects, replication has typically been evaluated based on a single study and dichotomously. In this article, we have used two characteristics shared by many large-scale replication projects—namely, their multilevel and multi-variate natures—to provide evaluations of replicability that are based on multiple studies, continuous, and multi-faceted.

A heretofore unmentioned difference between the two approaches to evaluating replicability is that the former is *external* in the sense that it compares the estimates from a large-scale replication project to those from some original study[4] while the latter is *internal* in the sense that it compares the estimates across the labs involved in a project. Consequently, the two need not be concordant. For example, from an external perspective, the Att-SNARC RRR "conclude[d] that we failed to

[4] In large-replication projects that are multilevel in nature such as RRRs and MLPs, the estimates from the large-scale replication project used for comparison to date have typically been single meta-analytic average estimates, and therefore the replication could be considered to be evaluated based on a single study.

replicate the effect reported by Fischer et al. (2003)" because "the effects we observed both within and across labs were minuscule and incompatible with those observed by Fischer et al. (2003)." However, from an internal perspective, the Att-SNARC RRR "successfully replicated" in the sense that estimates of the variates and effects varied very little across the labs involved in the project. Further, from an external perspective, the MLP concluded that they "successfully replicated 11 of 13" effects because the effects they observed attained "statistical significance" and were directionally consistent with those observed in the original studies of the phenomena. However, from an internal perspective, the MLP "failed to replicate" depending on the variate or effect in the sense that estimates of many variates and effects varied much across the labs involved in the project.

The external and internal approaches are of course not mutually exclusive, but the latter has heretofore been overlooked and has several advantages. First, estimates from original studies in many domains are typically quite imprecise; consequently, any comparison of the estimates from a large-scale replication project to those from an original study as per the external approach will also be quite imprecise even if the former estimates are precise. Second, unlike the external approach, the internal approach not only is based on multiple studies, is continuous, and is multi-faceted, but also follows directly from model parameters. Third, internal replicability would seem to be a necessary (though not sufficient) condition for external replicability because insofar as estimates vary across the labs involved in a project, they are likely to vary even more across labs that are not involved in the project.

If the estimates of the variation in variates and effects across the labs involved in a large-scale replication project are sufficiently low for certain variates and effects, that indicates those variates and effects are replicable. However, even if the estimates are not low, variates and effects can be considered replicable provided the variation is predictable. In particular, if there is some known covariate or set of covariates—ideally causal ones but also correlates—that can reliably predict that certain variates or effects from some labs will be larger while those from other labs will be smaller, then the variate or effect is replicable.

Estimating the covariation in variates and effects is thus important for evaluating replicability because it can suggest such covariates. In particular, a high degree of covariation in variates and effects suggests that a common covariate or set of covariates is causing or is associated with the variation in the variates or effects and facilitates the identification of such covariates when unknown, as illustrated by ocular dominance in the Att-SNARC RRR and political ideology in the MLP.

We note that such covariates are of two types, namely theoretically pertinent ones and method factors (i.e., anything pertaining to the implementation of the study that is not directly related to the theory under investigation; McShane et al. 2019). Regardless of type, these covariates should be accounted for in the study design and analysis so as to reduce variation and thus increase replicability. Theoretically pertinent covariates should also be incorporated into theory so as to enhance it. Further, whether a given covariate should be deemed theoretically pertinent or not will depend on the perspective of the researcher.

Finally, we have focused on replicability as evaluated by the variation in variates and effects at the lab level. Two aspects of this evaluation are noteworthy. First, evaluations of replicability to date have focused solely on effects and given no consideration to variates. However, it is necessary to consider both to have a truly multi-faceted evaluation of replicability. Further, one should expect effects considered solely on their own to provide an overly-optimistic evaluation of replicability; specifically, one should expect the variation and covariation in variates to be higher than that in effects due to the zero-sum nature of contrasts.

Second, we have focused the evaluation of replicability on the lab level because theories in psychological research are typically at the aggregate rather than individual level. In other domains where theories may be at the individual level, the variation in variates and effects at the subject level in addition to or in lieu of that at the lab level may be more germane to evaluating replicability.

4.2. Recommendations for Future Large-Scale Replication Projects

We have four recommendations for future large-scale replication projects. First, we recommend that large-scale replication projects analyze the data via an approach that provides all eight quantifications relevant for evaluating replicability, namely that of the variation and covariation in variates and effects at the lab and subject levels. Our modeling framework will prove useful for this purpose, and we discuss in the next subsection several possible model extensions that are motivated by designs typical in psychological research that our framework can accommodate.

Second, we recommend that large-scale replication projects employ repeated measures for all phenomena under investigation in the project (as was done for the single phenomenon in the Att-SNARC RRR and the Anchoring phenomenon in the MLP) rather than a single measure (as was done for all phenomena other than the Anchoring phenomenon in the MLP). This has several important benefits. Repeated measures of a phenomenon allow for a better accounting of lab and individual differences with respect to the phenomenon. This yields not only more substantive estimates of variation and covariation but also information directly relevant for evaluating replicability. For instance, as illustrated in the MLP, the repeated measures of the Anchoring phenomenon suggested lab and individual differences in knowledge (i.e., of the distance from San Francisco to New York City, the number of babies born per day in the United States, etc.); this in turn suggested that the Anchoring theory may or may not apply to various dependent measures associated with (i.e., operationalizations of) it depending on the lab or subject—something that seemed especially relevant for the dependent measure regarding the height of Mount Everest. Repeated measures also allow for an evaluation of measurement invariance which is a necessary precondition for an evaluation of replicability (see, e.g., Fabrigar and Wegener 2016).

Third, we recommend that large-scale replication projects focus on a single phenomenon as in the Att-SNARC RRR rather than many phenomena as in the MLP. This will increase covariation in variates and effects and thus the likelihood of identifying unknown covariates, whether theoretically pertinent ones or method factors.

At minimum, large-scale replication projects that investigate many phenomena should give careful consideration to theoretical relationships among the phenomena that they investigate when choosing the phenomena. Consideration to date has been given solely to pragmatic concerns. For example, in the MLP, the phenomena were chosen based on suitability for online presentation, length of study, simplicity of design, and diversity. Regarding diversity, our results suggest that the phenomena investigated by the MLP are not in fact so diverse; specifically, the dependent measures associated with many of these phenomena relate to political ideology. Thus, insofar as diversity of phenomena is of interest, diversity with respect to both the dependent measures and the experimental conditions associated with the phenomena is necessary.

Fourth, we recommend that large-scale replication projects vary the study materials within and across the labs involved in the projects rather than use the same materials (as was done in the Att-SNARC RRR and the MLP). This allows for an evaluation of the robustness of replicability, specifically by examining the degree to which the estimates of variation in variates and effects increases as the study materials vary (Baribault et al. 2018; McShane et al. 2019; DeKay et al. 2022).

4.3. Design-Based Model Generalization

To quantify the variation and covariation in variates and effects at the lab and subject levels, we introduced a multilevel multivariate modeling framework for analyzing all of the subject level data from large-scale replication projects jointly in a single analysis. Our framework employs a factor analytic structure for the variance-covariance matrices at the lab and subject levels that is specially tailored to the design of large-scale replication projects. Specifically, the factor analytic structure is constrained based on the design of these projects. This results in three distinct advantages: interpretability, adaptability, and parsimony.

Our design-based constraints stand in stark contrast to the unconstrained (or "exploratory") factor analytic structures typical in psychological research as well as the constrained (or "confirmatory") factor analytic structures also typical in psychological research in which the constraints are based on theoretical relationships among the variates assessed. For example, the estimates of the factor loadings in the Att-SNARC RRR presented in Figures 1 and 2 are not unconstrained (i.e., there are not 16 variates $\times$ 3 factors $=$ 48 freely varying loadings in each figure). Neither are they constrained based on theoretical relationships among the 16 variates. Instead, they are constrained based on the design of the Att-SNARC RRR, namely that it investigated a single phenomenon and featured four dependent measures and four experimental conditions, with all four dependent measures assessed for each experimental condition. These design-based constraints respect the fact that variates subsist in dependent measures which subsist in the phenomenon as a whole.

The model specification employed in the Att-SNARC RRR represents a "minimal" specification that is highly adaptable to variations in the design, as illustrated by the model extensions employed in the MLP. Indeed, our modeling framework is sufficiently general to accommodate an arbitrary number of phenomena, dependent measures, experimental conditions, levels, and covariates at any level. We therefore note several other possible model extensions that are motivated by designs typical in psychological research that our framework can accommodate and that will prove useful for the analysis of the data from past and future large-scale replication projects.

If our recommendation to employ repeated measures for all phenomena under investigation in a large-scale replication project is heeded, the basic three-factor structure employed in the Att-SNARC RRR can be extended to introduce a phenomenon factor as well as intercept, dependent measure, and variate factors for each phenomenon. In this case, the design-based constraints respect the fact that variates subsist in dependent measures which subsist in the particular phenomena under investigation which subsist in psychological phenomena as a whole.

In addition, if the experimental conditions are the same across the dependent measures (e.g., as is the case when there are repeated measures of the same phenomenon), the intercept factor can be extended so that the loadings can vary across the variates associated with each experimental condition, as illustrated at the lab level in the MLP. Further, if these experimental conditions follow a factorial design, factors corresponding to this design can instead be introduced.

Moreover, in observational research with no experimental conditions, the variate factor can simply be omitted. Alternatively, if subgroups of subjects such as demographic groups are of interest, those can play the role of the experimental conditions in defining variates.

To accommodate levels in addition to the lab level and the subject level (e.g., to treat labs as grouped by continent), a random effect for each additional level unit (e.g., continent) and variate can be introduced. Further, one can assume the vector of random effects for each additional level unit are independent and identically distributed according to the multivariate normal distribution with mean zero and variance-covariance matrix modeled in the same manner as at the lab level.

To accommodate covariates at any level, the various model parameters can be allowed to vary as a function of them. The choice of which parameter(s) vary and how they vary can be based on theory or results, as illustrated in the Att-SNARC RRR and the MLP.

Finally, to accommodate settings in which the subject level data is not available but lab level data is (specifically, an estimate of the mean of each variate from each lab, an estimate of the variance-covariance matrix of the variates from each lab, and the sample size from each lab), the model requires only minor modification.

Supplementary Materials

The Supplementary Materials contain data and code to reproduce all results (i.e., Figures 1–10) in the article.

ORCID

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

References

Baribault, B., Donkin, C., Little, D. R., Trueblood, J., Oravecz, Z., van Ravenzwaaij, D., White, C., De Boeck, P., and Vandekerckhove, J. (2018), "Meta-Studies for Robust Tests of Theory," *Proceedings of the National Academy of Sciences*, 115, 2607–2612. [1620]

Begley, C. G., and Ellis, L. M. (2012), "Raise Standards for Preclinical Cancer Research," *Nature*, 483, 531–533. [1605]

Berkey, C. S., Hoaglin, D. C., Antczak-Bouckoms, A., Mosteller, F., and Colditz, G. A. (1998), "Meta-Analysis of Multiple Outcomes by Regression with Random Effects," *Statistics in Medicine*, 17, 2537–2550, 1998. [1606]

Bourassa, D., McManus, I., and Bryden, M. (1996), "Handedness and Eye-Dominance: A Meta-Analysis of their Relationship," *Laterality*, 1, 5–34. [1609]

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26, 211–252. [1611]

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016), "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351, 1433–1436. [1605]

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikenstein, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018), "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, 2, 637–644. [1605]

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76, 1–32. [1607]

Colling, L. J., Szűcs, D., De Marco, D., Cipora, K., Ulrich, R., Nuerk, H.-C., Soltanlou, M., Bryce, D., Chen, S.-C., Schroeder, P. A., Henare, D. T., Chrystall, C. K., Corballis, P. M., Ansari, D., Goffin, C., Sokolowski, H. M., Hancock, P. J. B., Millen, A. E., Langton, S. R. H., Holmes, K. J., Saviano, M. S., Tummino, T. A., Lindemann, O., Zwaan, R. A., Lukavský, J., Becková, A., Vranka, M. A., Cutini, S., Mammarella, I. C., Mulatti, C., Bell, R., Buchner, A., Mieth, L., Röer, J. P., Klein, E., Huber, S., Moeller, K., Ocampo, B., Lupiáñez, J., Ortiz-Tudela, J., de la Fuente, J., Santiago, J., Ouellet, M., Hubbard, E. M., Toomarian, E. Y., Job, R., Treccani, B., McShane, B. B. (2020), "Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003)," *Advances in Methods and Practices in Psychological Science*, 3, 143–162. [1606]

DeKay, M. L., Rubinchik, N., Li, Z., and DeBoeck, P. (2022), "Accelerating Psychological Science with Metastudies: A Demonstration Using the Rrisky-Choice Framing Effect," *Perspectives on Psychological Science*, forthcoming. [1620]

Erosheva, E. A., and Curtis, S. M. (2017), "Dealing with Reflection Invariance in Bayesian Factor Analysis," *Psychometrika*, 82, 295–307. [1607]

Fabrigar, L. R., and Wegener, D. T. (2016), "Conceptualizing and Evaluating the Replication of Research Results," *Journal of Experimental Social Psychology*, 66, 68–80. [1619]

Fischer, M. H., and Brugger, P. (2011), "When Digits Help Digits: Spatial–Numerical Associations Point to Finger Counting as Prime Example of Embodied Cognition," *Frontiers in Psychology*, 2, 260. [1606]

Fischer, M. H., Castel, A. D., Dodd, M. D., and Pratt, J. (2003), "Perceiving Numbers Causes Spatial Shifts of Attention," *Nature Neuroscience*, 6, 555–556. [1606,1619]

Fodor, J. A. (1975), *The Language of Thought* (Vol. 5), Cambridge, MA: Harvard University Press. [1606]

Galton, F. (1880), "Visualized Numerals," *Nature*, 21, 252–256. [1606]

Gelman, A., Van Dyk, D. A., Huang, Z., and Boscardin, J. W. (2008), "Using Redundant Parameterizations to Fit Hierarchical Models," *Journal of Computational and Graphical Statistics*, 17, 95–122. [1607]

Gładziejewski, P., and Miłkowski, M. (2017), "Structural Representations: Causally Relevant and Different from Detectors," *Biology & Philosophy*, 32, 337–355. [1606]

Kalaian, H. A., and Raudenbush, S. W. (1996), "A Multivariate Mixed Linear Model for Meta-Analysis," *Psychological Methods*, 1, 227–235. [1606]

Kay, A. C., and Jost, J. T. (2003), "Complementary Justice: Effects of "Poor but Happy" and "Poor but Honest" Stereotype Exemplars on System Justification and Implicit Activation of the Justice Motive," *Journal of Personality and Social Psychology*, 85, 823–837. [1616]

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Barry Kappes, H., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van 't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A., and Nosek, B. A. (2014), "Investigating Variation in Replicability: A "Many Labs" Replication Project," *Social Psychology*, 45, 142–152. [1606]

McCulloch, R., and Rossi, P. E. (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 207–240. [1607]

McShane, B. B., and Böckenholt, U. (2018), "Multilevel Multivariate Meta-Analysis with Application to Choice Overload," *Psychometrika*, 83, 255–271. [1606]

McShane, B. B., Tackett, J. L., Böckenholt, U., and Gelman, A. (2019), "Large Scale Replication Projects in Contemporary Psychological Research," *The American Statistician*, 73, 99–105. [1605,1619,1620]

Muthén, B. O. (1994), "Multilevel Covariance Structure Analysis," *Sociological Methods & Research*, 22, 376–398. [1606]

Newell, A., and Simon, H. A. (1976), "Computer Science as Empirical Enquiry: Symbols and Search," *Communications of the ACM*, 19, 113–126. [1606]

Open Science Collaboration. (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. [1605]

Prinz, F., Schlange, T., and Asadullah, K. (2011), "Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets," *Nature Reviews Drug Discover*, 10, 712. doi:10.1038/nrd3439-c1. [1605]

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004), "Generalized Multilevel Structural Equation Modeling," *Psychometrika*, 69, 167–190. [1606]

Stan Development Team (2020), "Stan user's guide (version 2.24)." Available at *http://mc-stan.org/*. [1607]

Williams, D., and Colling, L. (2018), "From Symbols to Icons: The Return of Resemblance in the Cognitive Neuroscience Revolution," *Synthese*, 195, 1941–1967. [1606]

Wilson, M. (2002), "Psychonomic Bulletin & Review," *Six views of Embodied Cognition*, 9, 625–636. [1606]