Lack of group-to-individual generalizability is a threat to human subjects research

Aaron J. Fisher[a,1], John D. Medaglia[b,c], and Bertus F. Jeronimus[d]

[a]Department of Psychology, University of California, Berkeley, CA 94720; [b]Department of Psychology, Drexel University, Philadelphia, PA 19104; [c]Department of Neurology, University of Pennsylvania, Philadelphia, PA 19104; and [d]Department of Developmental Psychology, Faculty of Behavioural and Social Sciences, Groningen University, 9712 TS Groningen, The Netherlands

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved May 25, 2018 (received for review July 4, 2017)

Only for ergodic processes will inferences based on group-level data generalize to individual experience or behavior. Because human social and psychological processes typically have an individually variable and time-varying nature, they are unlikely to be ergodic. In this paper, six studies with a repeated-measure design were used for symmetric comparisons of interindividual and intraindividual variation. Our results delineate the potential scope and impact of nonergodic data in human subjects research. Analyses across six samples (with 87–94 participants and an equal number of assessments per participant) showed some degree of agreement in central tendency estimates (mean) between groups and individuals across constructs and data collection paradigms. However, the variance around the expected value was two to four times larger within individuals than within groups. This suggests that literatures in social and medical sciences may overestimate the accuracy of aggregated statistical estimates. This observation could have serious consequences for how we understand the consistency between group and individual correlations, and the generalizability of conclusions between domains. Researchers should explicitly test for equivalence of processes at the individual and group level across the social and medical sciences.

research methodology | replicability | idiographic science | generalizability | ecological fallacy

Inferences made in social and medical research typically result from statistical tests conducted on aggregated data. The implicit assumption is that group-derived estimates can be applied to understanding individual phenomenology, physiology, and behavior. However, statistical findings at the interindividual (group) level only generalize to the intraindividual (person) level if the processes in question are ergodic (1). Ergodic processes are equivalent for groups and individuals (homogeneity criterion) when their mean and variance remain consistent over time (stationarity criterion) (2). Because psychological and biological phenomena are organized within persons over time, generalizations that rely on group estimates are nonergodic if there are individual exceptions.

Group estimates can be derived from a cross-sectional measurement of individuals at one point in time, but intraindividual analyses require data collected over time—as the sample size becomes the number of repeated observations. Just as a random sample is required to be representative of a population to support generalizable claims about that population, the data sampled within an individual in time must be representative of that individual generally (i.e., stationary). When group generalizations obscure genuine individual differences, we may fail to describe natural processes and their natural kinds (3). If scientific consilience (4) and completeness (5) are our goals, this scenario should be avoided.

In this paper, we contend that (*i*) nonergodicity—specifically, the lack of generalizability from group to individual statistical estimates—is a threat to human subjects research, because we do not know the full scope of the problem and are not adequately studying it; and (*ii*) that scientists need to demonstrate the consistency between individual and group variability before generalizing results across levels of analysis. We will refer to this latter condition as the "group-to-individual generalizability" of a given statistical estimate. However, whether couched in prosaic terms, or within formal mathematical theorems, researchers have not systematically examined such generalizability in extant literatures, despite a number of calls to do so throughout the years (cf. refs. 6–11). Hitherto, the highest-impact publications in medical and social sciences have been largely based on data aggregated across large samples, with best-practice guidelines almost exclusively based on statistical inferences from group designs. The worst-case scenario—a global, uniform absence of group-to-individual generalizability due to nonergodicity in the social and medical sciences—would undermine the validity of our scientific canon in these domains. However, even moderate incongruities between group and individual estimates could result in imprecise or potentially invalid conclusions. We argue that this possibility should be formally tested, wherever possible, to be ruled out.

Ergodicity, the Ecological Fallacy, and Simpson's Paradox

The ergodic theorem is a general and formal mathematical expression that deals with the generalizability of statistical phenomena across levels and units of analysis. [While a more thorough explication of the ergodic theorem is outside of the scope of the present paper, readers are referred to Molenaar (1) for a comprehensive mathematical treatment of ergodicity in human subjects research.] Ergodic theory postulates that the

Significance

The current study quantified the degree to which group data are able to describe individual participants. We utilized intensive repeated-measures data—data that have been collected many times, across many individuals—to compare the distributions of bivariate correlations calculated within subjects vs. those calculated between subjects. Because the vast majority of social and medical science research aggregates across subjects, we aimed to assess how closely such aggregations reflect their constituent individuals. We provide evidence that conclusions drawn from aggregated data may be worryingly imprecise. Specifically, the variance in individuals is up to four times larger than in groups. These data call for a focus on idiography and open science that may substantially alter best-practice guidelines in the medical and behavioral sciences.

Author contributions: A.J.F. designed research; A.J.F. performed research; A.J.F. analyzed data; and A.J.F., J.D.M., and B.F.J. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: afisher@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1711978115/-/DCSupplemental.

Published online June 18, 2018.

necessary—although not always sufficient—conditions for ergodicity in human subjects data are that the structures of interindividual and intraindividual variation are asymptotically equivalent (1). The ergodic theorem can be understood as a general frame of reference to identify specific cases of statistical incongruity and inferential errors, including Simpson's paradox and the ecological fallacy. Simpson's paradox (12) is a statistical effect in which trends in subgroups differ from (or are even inverse to) the aggregate trend when the groups are combined. The ecological fallacy is a common and problematic statistical interpretation error, in which statistical inferences from groups are inappropriately generalized to individuals (13). An intuitive example is provided by Hamaker (14), who describes the correlation between typing speed and typos. At the group level, the correlation is negative, as experienced typists are both faster and more proficient. However, within individuals, the correlation is positive—the faster a given individual types, the greater the number of mistakes she or he will make relative to their own performance at slower speeds. Thus, the aggregation of the data produces an example of Simpson's paradox, and we would commit an ecological fallacy by concluding that the relationship observed at the group level represents any of the individuals in the group. Both Simpson's paradox and the ecological fallacy remind us that the individual level and group level are not necessarily related. The effects of nonergodicity in a given dataset should therefore be directly tested before any extrapolations are made.

Unfortunately, applied tests of ergodicity are uncommon in the social, behavioral, and medical sciences. While others have observed that processes within persons over time differ from processes sampled across persons (cf. refs. 1, 15, and 16), explicitly quantifying the scope and potential threat of this discrepancy in psychosocial and medical domains should be a routine focus of scientific inquiry. Whereas Pearl (17, 18) has demonstrated that there is no single diagnostic or correction for Simpson's paradox, we propose that there is a relatively easy way to directly test for nonergodicity and, thus, group-to-individual generalizability in statistical analyses. Quite simply, comparisons of the first and second moments (mean and variance) of intraindividual and interindividual distributions can inform us about the accuracy of generalizations between groups and individuals. To thoroughly examine group-to-individual generalizability across the social and medical sciences, prodigious collaborative efforts across all areas of human subjects research would be required. In the meantime, individual researchers can address the suitability of their data for generalizations from aggregated results to individual participants through apposite research designs and data collection paradigms. Specifically, scientists who endeavor to generalize findings between interindividual and intraindividual levels of analysis should collect many measurements within subjects over time—whether or not the research question is explicitly longitudinal. Moreover, sharing data and results could alleviate the burden to comprehensively test for ergodicity in future studies. Fortunately, as data resources become increasingly available through open access, we can begin to collectively confront this challenge. To determine the importance of this effort, we use six independent datasets of repeatedly sampled individuals to draw comparisons between intraindividual and interindividual variation.

What Exactly Should We Be Testing, and How? To generalize group findings to individuals, we should be interested in measures of central tendency and measures of covariation. The central tendency, typically represented as the mean or median, reflects the expected value of a distribution—the most likely value for a variable—and, thus, is a common proxy for the nature of the variable en masse. For a group-to-individual generalization to apply, the central tendency within individuals sampled over time should be the same as between individuals. Conveniently, common metrics are available for determining the similarity (vs. dissimilarity) of the respective central tendencies.

A common use of the sample mean is to determine whether one group is statistically different from another group as a function of the ratio between the mean difference and the spread around that difference (the SE). This is a signal-to-noise ratio: Signals that are roughly twice the size of the respective noise are considered to be statistically significant. In turn, the SE, which constitutes the denominator in this ratio, is a function of the sample size and the SD—the average deviation from the mean, across the sample. Thus, mean values should always be viewed in the context of SDs, and both should be examined when testing for nonergodicity.

As readers will recall, the SD is the square root of the variance, and the variance, in turn, is one-half of the joint variability—the "covariance" between two variables. The covariance represents the bedrock of quantitative analyses in social, behavioral, and medical sciences. The entire body of classical statistics based on the general linear model, including regression, analysis of variance, multilevel modeling, and structural equation modeling, all use the covariation between variables to support statistical inferences. For statistical estimates to generalize from groups to individuals, it is imperative that the relations among intraindividual variables are equivalent to the relations among interindividual variables.

Thus, because the central tendency is often employed to represent the overall data, tests that measure the validity of group-to-individual generalizations should examine the consistency of the mean across groups and individuals. Because the "signal" that the mean represents is only as strong as its relation to the variance, we should be equally concerned with equivalence in the variance (and SD). Finally, statistical inferences that are drawn from relations between variables should be tested for consistency in the bivariate and multivariate covariation among those variables.

What Is at Stake? The consequences of neglecting ergodic theory in social, behavioral, and medical fields may have substantial epistemic and practical consequences. In the absence of quantitative examination at the individual level, the consequences could range from zero if we are lucky to find one of the few ergodic processes in nature (19), to catastrophic if a process is quite nonergodic. In clinical research, diagnostic tests may be systematically biased and our classification systems may be at least partially invalid. In terms of theory development, we may have a misleading impression about the nature of psychological variables and their interactions. Regarding experiments, we may not design and employ research designs necessary to adequately diagnose these issues. Overall, we may need to reconsider how we communicate statistical principles to students and researchers and encourage them to design studies capable of making explicit tests of the consistency of intraindividual and interindividual variation across settings, paradigms, and constructs. Here, we begin to examine the stakes.

Comparison of Intraindividual and Interindividual Distributions

Across six previously published datasets, we quantified intraindividual and interindividual variation and covariation across multiple samples and constructs. We made roughly symmetric comparisons, calculating intraindividual and interindividual estimates across equivalent numbers of individual and cross-sectional datasets. Given the repeated-measures paradigms employed in each of the studies, multiple cross-sectional datasets were constructed from the between-subject aggregations of within-subject data on an observation-by-observation basis. That is, the concatenation of the set of first rows for each individual's repeated-measures data in a given sample yielded a cross-sectional dataset with n = number of participants. The concatenation of all second rows likewise yielded a cross-sectional dataset with n = number of participants, and so forth (see *Method* for complete details). This allowed us to go beyond single-point estimates to compare the distributions of comparable cross-sectional and intraindividual means and variances—both of the univariate descriptives for study variables, and of bivariate associations among those variables.

Method

The present study comprises a set of analyses of six previously published datasets at the group and individual levels. All study procedures were approved by the relevant institutional review boards and conducted in accordance with ethical standards.

Participants.

Sample 1. The first sample comprises 43 individuals with principal generalized anxiety disorder (GAD) and/or major depressive disorder (MDD), and 35 healthy control participants. Details regarding recruitment procedures, screening, and diagnostic interviewing are provided by Fisher et al. (20). The 78 participants were predominantly female ($n = 48$, 62%) and white ($n = 35$, 45%). Following study enrollment, participants' mobile phone numbers were entered into a secure web-based survey system. Four times a day for 30 d, participants received a text message containing a hyperlink to a web-based survey (each with a time stamp). Participants rated their experience of each item over the preceding hours using a 0–100 visual analog slider with the anchors "not at all" (=0) and "as much as possible" (=100). In addition to the extant *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5) (21) criteria for GAD and MDD, the surveys included an additional 11 items gauging positive affect (PA) (positive, energetic, enthusiastic, and content), negative affect (NA) (angry, afraid, and dwelled on the past), and behavioral avoidance (avoided people, avoided activities, sought reassurance, and procrastinated). This study was approved by the University of California Institutional Review Board (Committee for Protection of Human Subjects protocol no. 2014-03-6138; Committee for Protection of Human Subjects protocol title: Personalized Interventions for Anxiety and Depression). Written informed consent was obtained before participation.

Sample 2. Participants from sample 2 took part in a polysomnography study conducted at the National Center for Post-Traumatic Stress Disorder (PTSD) in Menlo Park, California. The present study only uses the data on heart rate (HR) and respiratory sinus arrhythmia (RSA). Complete details regarding recruitment procedures, screening, diagnostic interviewing, and physiological data collection and methodology (including electrocardiography signal cleaning and preparation) are published in the study by Fisher and Woodward (22). The present sample comprised 69 individuals, 16 healthy controls, 23 individuals with PTSD, 14 individuals with panic disorder, and 16 individuals with co-occurring PTSD and panic disorders. No group differences were found for HR or RSA (22). This study was authorized by the Stanford/Veterans Affairs (VA) Palo Alto Human Research Protection Program. All participants provided written informed consent.

Sample 3. Participants from sample 3 were taken from an ecological momentary assessment study of individuals with clinically diagnosed personality disorders (23, 24). Participants ($n = 101$) completed ratings of psychosocial and interpersonal experiences once per day for 100 d. Complete details regarding recruitment, diagnosis, and study procedures can be found in the studies by Wright et al. (23) and Wright and Simms (24). The present study utilized a subsample of 83 participants who completed a minimum of 83 surveys (*Analytic Approach*) on the scale-level variables PA and NA, using a subset of the Positive and Negative Affect Schedule (PANAS) (25) items assessed on a five-point scale (very slightly or not at all, a little, moderately, quite bit, and very much) "over the last 24 h." Daily PA was measured as the mean of the items: active, alert, attentive, determined, and inspired. Daily NA was measured as the mean of afraid, ashamed, hostile, nervous, and upset. To be consistent with the affect measure employed in sample 6 (below), it should be noted that the indicators for PA and NA in sample 3 are consistent with high-arousal PA and NA, specifically. This study was approved by the University at Buffalo, State University of New York Institutional Review Board. Written informed consent was obtained before participation.

Sample 4. Participants from sample 4 were taken from a randomized controlled trial (RCT) for clinical depression in which 64 individuals were randomized to receive imipramine or placebo (26). Before treatment, participants took part in an intensive repeated-measures paradigm in which they completed surveys 10 times per day for 6 d. These data have been described in detail elsewhere (cf. refs. 26 and 27). Participant mean age was 42.5 y (SD, 9.1), and the majority of the sample was female (74%). The present study uses the PA scale (mean of energetic, cheerful, satisfied, alert, calm, enthusiastic, strong, and happy) and NA scale (mean of hostile, depressed, tense, lonely, anxious, insecure, guilty, harried, and irritable). All study procedures were approved by the Medical Ethics Committee of Maastricht University Medical Centre, and all participants signed an informed consent form.

Sample 5. The fifth sample was drawn from a RCT for mindfulness-based cognitive therapy (MBCT), in which 64 individuals were randomized to MBCT and 66 were randomized to waitlist control (28). The present study utilized only those individuals randomized to MBCT. Consistent with sample 4, participants completed a daily survey paradigm before treatment. Each participant was given a digital wristwatch and assessment forms for completing daily self-reports. The watches were programmed to signal participant at random intervals within each of 10 90-min blocks between 7:30 AM and 10:30 PM. Assessments were completed for 6 consecutive days, with a maximum of 60 signals per study period. At each signal, participants were instructed to complete the self-assessment forms. Current mood and context were rated on seven-point Likert scales. Consistent with samples 3 and 4, the present study utilized scales for PA (mean of happy, satisfied, strong, enthusiastic, curious, animated, and inspired) and NA (mean of feeling down, anxious, lonely, suspicious, disappointed, insecure, and guilty). All study procedures were approved by the Medical Ethics Committee of Maastricht University Medical Centre, and all participants signed an informed consent form.

Sample 6. The final sample was taken from the ongoing naturalistic study "How Nuts are the Dutch" (HoeGekIsNL), in which 975 participants filled out diaries three times a day for 30 d. Complete study methodology and procedures can be found in van der Krieke et al. (29). In this subsample, the mean number of diary responses was 60.53 (SD, 18.10). The sample was 82% female ($n = 449$), and the mean age was 40.68 (SD, 13.53). The variables employed in the present study were high-arousal PA and NA and low-arousal PA and NA, as calculated from daily diary questions 5–16 (see ref. 29, p. 129). The present study included the 535 participants (55%) who completed at least 44 surveys. The HND study protocol was assessed by the Medical Ethical Committee of the University Medical Centre Groningen and judged to be exempted from the Medical Research Involving Human Subjects Act (in Dutch: WMO) because it concerned a nonrandomized open study targeted at anonymous volunteers in the general public (registration no. M13.147422).

Analytic Approach. We sought to quantify the degree to which group data can describe individual participants by comparing intraindividual and interindividual variation and covariation across multiple samples and constructs. We used intensive repeated-measures data to compare the univariate distributions of each variable, as well as the distributions of bivariate correlations calculated within subjects vs. those calculated between subjects. As outlined below, the same data (participants and variables) were utilized to make interindividual vs. intraindividual comparisons, so that we could directly assess the degree to which aggregations reflect their constituent individuals.

Temporal dependence. As has been discussed in detail elsewhere, statistical estimates of repeated-measures data can be biased due to temporal dependence (cf. refs. 30 and 31). That is, individuals will tend to correlate with themselves over time, leading to correlated errors in models that do not account for the underlying autocorrelation. Although one common way to handle the temporal dependence in repeated measures is to employ a multilevel model that partitions the variance of dependent variables into fixed and random effects (32), these models do not represent true intraindividual variation (33). Instead, they reflect individual deviations (in intercept and slope) from aggregated estimates. To isolate the true intraindividual variation and covariation, we used the Durbin–Watson test (34) to assess the degree of temporal dependence in each intraindividual variable.

Sources of lagged and cross-lagged temporal dependence were examined for each bivariate pair analyzed below. Regression models were constructed in which each intraindividual variable was regressed on its first-order autoregression (i.e., lag-1) and the first-order cross-lag for any bivariate pairs. For instance, because we examined separate bivariate correlations for depressed mood with anhedonia, and depressed mood with worry (sample 1), two regression models were conducted for the variable depressed mood, one with cross-lagged measurements of anhedonia, and one with cross-lagged measurements of *worry*. For any bivariate pair X and Y, the associated regression models are expressed algebraically as follows:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 X_{t-1} + e_t,$$

$$X_t = \beta_1 X_{t-1} + \beta_2 Y_{t-1} + e_t.$$

The Durbin–Watson test was employed to assess the degree of temporal dependence in the residuals of each regression model.

Inclusion of first-order lagged and cross-lagged variables effectively removed the temporal dependence from (i.e., whitened) the majority of individual data for each of the behavioral variables from samples 1, 3, 4, 5, and 6 (range, 86–99%; mean percentage, 97%). However, after conducting first-order regression models, temporal dependence remained in a handful of these behavioral data and the majority of the psychophysiological data from sample 2. For each intraindividual variable that yielded a Durbin–Watson test statistic of $P < 0.05$, additional regression models were conducted—each with an additional set of lagged and cross-lagged parameters—until the residuals for the regression model were found to be effectively whitened. Models that included lag-2 variables effectively whitened the remainder of all but one intraindividual behavioral variable, 57% of intraindividual HR data, and 72% of RSA data. Models with three or more lags were required for one intraindividual PA variable from sample 4, 39% of HR data, and 16% of RSA data from sample 2. *SI Appendix*, Table S1 presents the complete results of the temporal dependence tests for each variable.

Once a regression model yielded a nonsignificant Durbin–Watson test statistic, the residuals were retained for further analysis. We refer to these data as the AR residuals. Thus, for each bivariate correlation reported below, we report the means and SDs for intraindividual estimates derived from the raw data, as well as from the AR residual data.

Constructing interindividual cross-sections. As noted above, the present study sought to compare cross-sectional, aggregated data to individual, time-varying data to investigate the degree to which aggregations represent their constituent individuals. To make such comparisons, we required multiple intraindividual time series and interindividual cross-sections, which should be composed of the same individuals. The extant data from the six studies were composed of the former: individual time series ranging in length from 69 to 900 observations. These time series thus comprise repeated observations in time, collected on a minimum of 69 separate occasions. To make cross-sectional comparisons, each repeated observation was doubly employed: as a single observation within an individual time series and as a single participant in an aggregated cross-section. That is, if 100 individuals participate in a study and each individual provides 100 observations over time, we can represent the entire sample of data in a 100 (observations) × 100 (individuals) matrix. We can thus consider each row to be a sample of 100 separate individuals measured on one occasion, and each column to be the time series for a single individual, measured 100 times. Extending this to a multivariate space, if each individual's time series is organized such that observations in time are in rows and variables are in columns, the concatenation of the set of first rows will yield a cross-sectional dataset with between-subjects sample size of $n = 100$. The concatenation of all second rows will likewise yield a cross-sectional dataset with $n = 100$, and so forth.

For each sample, we endeavored to make comparisons that were symmetric in scale and maximally powered. Symmetry was pursued to limit the degree to which differences in statistical power affected the comparison of interindividual and intraindividual variation. Exceptions to symmetry are described below. For each sample, we identified the number of participants and rows per participant that would yield the maximum number of intraindividual and interindividual datasets.

Samples 1, 2, and 3 contained sample sizes of 78, 69, and 101 individuals, respectively. The data collection paradigms for each study yielded individual time series with an average of 130, 925, and 90 observations each. Because the length of the time series in samples 1 and 2 were longer than the number of participants for each study, we generated random number strings to select a random subset of 78 and 69 rows from each intraindividual dataset in samples 1 and 2, respectively. As described above, each randomly selected row was extracted from each of the intraindividual datasets and concatenated across participants to yield a cross-sectional, interindividual dataset. For sample 1, the comparisons of interindividual to intraindividual variation was conducted between 78 intraindividual datasets with an average time series length of 130 observations, and 78 interindividual datasets, each with a between-subjects sample size of $n = 78$. The comparison for sample 2 included 69 intraindividual datasets with an average time series length of 925 observations, and 69 interindividual datasets, each with a between-subjects sample size of $n = 69$.

Sample 3 contained a greater number of participants (101) than average observations per participant (mean, 89). We identified the maximum number of participants with an equivalent minimum number of observations per participant at 83, and took the first 83 observations of each intraindividual dataset to construct the cross-sectional comparison datasets. Thus, for sample 3, the comparison of interindividual to intraindividual variation was conducted between 83 intraindividual datasets with an average time series length of 94 observations, and 83 interindividual datasets, each with a between-subjects sample size of $n = 83$.

Samples 4 and 5 contained 63 and 64 participants, who each completed 30–60 observations. To generate symmetric comparisons for these samples, we allowed the interindividual datasets to have variable sample sizes. That is, for both samples, interindividual datasets 1–30 contained 63 and 64 participants, respectively. As the number of interindividual datasets increased, the overall sample size for each set decreased. For sample 4, the minimum interindividual sample size was $n = 58$, and for sample 5, the minimum interindividual sample size was $n = 52$.

Table 1. Means and SDs for variables as calculated from intraindividual distributions (left) and interindividual distributions (right)

	Intraindividual		Interindividual		
Sample/variable	Mean	SD	Mean	SD	IAV:IEV ratio
Sample 1					
Depressed mood	30.59	22.01	30.16	2.77	7.95
Anhedonia	28.94	20.00	28.01	2.34	8.55
Worry	38.21	24.60	37.83	2.48	9.92
Fear	23.05	20.05	21.70	2.34	8.57
Avoidance	29.41	19.49	29.12	2.77	7.04
Sample 2					
HR	65.03	8.94	67.80	1.82	4.91
RSA	7.32	4.08	6.78	0.52	7.85
Sample 3					
PA	2.67	0.74	2.68	0.08	9.25
NA	1.77	0.64	1.78	0.07	9.14
Sample 4					
PA	−0.29*	0.66	−0.53*	0.05	13.20
NA	0.27	0.91	0.47	0.09	10.11
Sample 5					
PA	3.74	0.93	3.67	0.21	4.43
NA	1.90	0.72	2.00	0.19	3.79
Sample 6					
PA	56.35	11.34	55.33	1.71	6.63
NA	27.42	13.13	27.04	2.04	6.44

Note: IAV:IEV ratio, ratio of intraindividual SD to interindividual SD. Significant difference between interindividual and intraindividual estimates at $^*P = 0.003$.

Finally, sample 6 contained 975 participants with time series that ranged from 0 to 93 observations. Symmetric comparisons between intraindividual and interindividual data would require discarding more than 800 participants. Consideration was given to (*i*) the minimum number of observations to yield robust estimates of the bivariate correlations for each individual, and (*ii*) the maximum number of participants with an equivalent minimum number of observations per participant. A sample size of 44 was selected as the minimum threshold for intraindividual correlations, yielding a sample of 535 intraindividual datasets, with an average of 68 observations per individual. Regarding interindividual comparisons, we identified 58 participants with a minimum of 85 observations. The first 85 consecutive rows of these 58 intraindividual datasets were concatenated to generate 85 interindividual datasets, each with a sample size of 58. Thus, for sample 6, symmetry was abandoned to analyze the full set of 535 selected time series. The comparison of interindividual to intraindividual variation was conducted between 535 intraindividual datasets with an average time series length of 68 observations, and 85 interindividual datasets, each with a between-subjects sample size of n = 58.

Results

Univariate Distributions. We first examined the univariate distributions of each of the 15 study variables at the interindividual and intraindividual levels. The means and SDs for each variable, at each level of analysis, are presented in Table 1. The mean estimates were statistically equivalent across 14 of the 15 variables (all p's > 0.05), with the relative interindividual and intraindividual means falling within 7% of each other (mean discrepancy, 3%). The mean estimates for intraindividual and interindividual PA in sample 4 exhibited a significant discrepancy of 45% ($P = 0.003$). However, despite the consistency across mean estimates, the SDs for intraindividual and interindividual estimates exhibited marked discrepancies, with the former at least 3.79 times larger than the latter across all 15 variables (min ratio, 3.79:1; max ratio, 13.20:1; mean ratio, 7.85:1).

Bivariate Correlations.

Sample 1. To compare the relative distributions of intraindividual covariation and interindividual covariation, we selected three bivariate correlations from sample 1: (*i*) depressed mood and anhedonia, (*ii*) depressed mood and worry, and (*iii*) avoidance of activities and fear. Depressed mood and anhedonia are the principal symptoms of MDD, and at least one of the two is required for a clinical diagnosis. Worry is the cardinal symptom of GAD, a disorder that has the highest rate of co-occurrence with MDD (35). Behavioral avoidance is a common feature of both mood and anxiety symptomatology, typically thought to be negatively reinforced via temporary reductions in fear and anxiety, thereby maintaining the symptoms over time (36).

Bivariate correlations between the three selected variable pairs were conducted across the 78 intraindividual datasets and 78 interindividual datasets. Fig. 1 presents the histograms for the intraindividual correlations (red) and interindividual correlations (blue), with results calculated from raw data (top) and AR residual data (bottom). The solid lines reflect the mean correlation, and the dashed lines reflect the median correlation. The mean intraindividual correlations (with SD in brackets) were r = 0.47 (0.23), r = 0.40 (0.22), and r = 0.23 (0.23) for raw data, and r = 0.41 (0.22), r = 0.33 (0.21), and r = 0.19 (0.21) for AR residuals. The mean interindividual correlations were r = 0.75 (0.07), r = 0.71 (0.08), and r = 0.46 (0.11). The disparities between

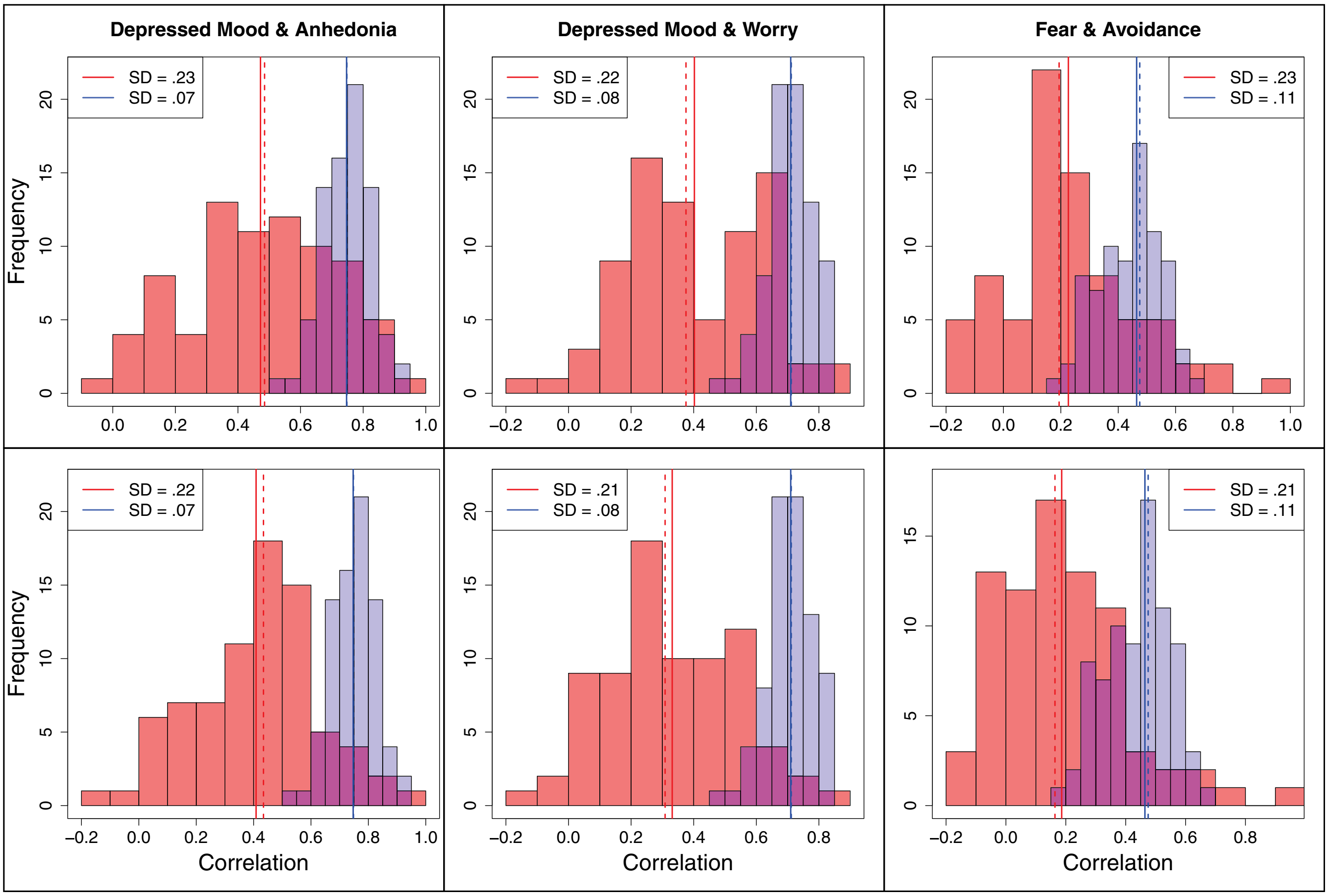

Fig. 1. Histograms for intraindividual (red) and interindividual (blue) correlations for four bivariate relationships in sample 1. *Top* depicts intraindividual correlations calculated from raw data. *Bottom* depicts intraindividual correlations calculated from data with temporal dependence removed.

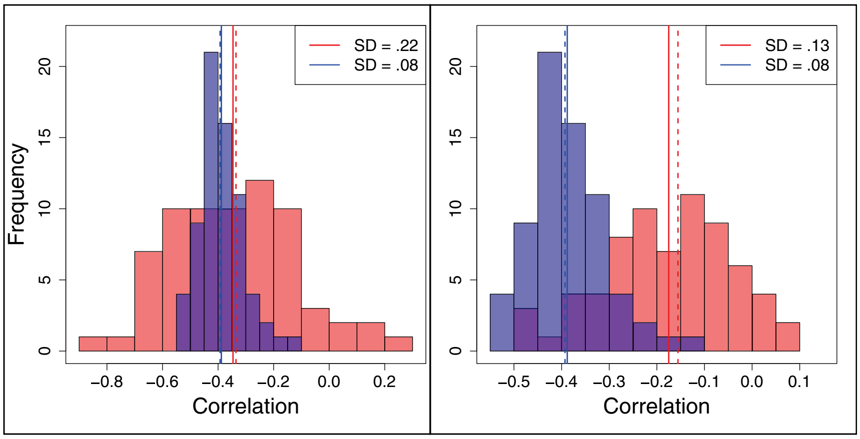

Fig. 2. Histograms for intraindividual (red) and interindividual (blue) correlations for RSA and HR. *Left* depicts intraindividual correlations calculated from raw data. *Right* depicts intraindividual correlations calculated from data with temporal dependence removed.

intraindividual and interindividual correlations were $r = 0.28$, $r = 0.31$, and $r = 0.22$ for raw data, and $r = 0.34$, $r = 0.38$, and $r = 0.27$ for AR residuals, with an average difference of $r = 0.27$ for raw data and $r = 0.33$ for AR residual data. Moreover, across the three comparisons, intraindividual SDs were 2.71 times larger than interindividual SDs for raw data and 2.46 times larger for AR residual data. These results reflect a substantially wider range of variability across intraindividual estimates, regardless of adjustment for temporal dependence.

Sample 2. For sample 2, a single comparison was calculated for the bivariate correlation between HR and RSA, two of the most common variables employed in psychophysiological research (37). Fig. 2 presents the histograms from the intraindividual (red) and interindividual (blue) correlations between HR and

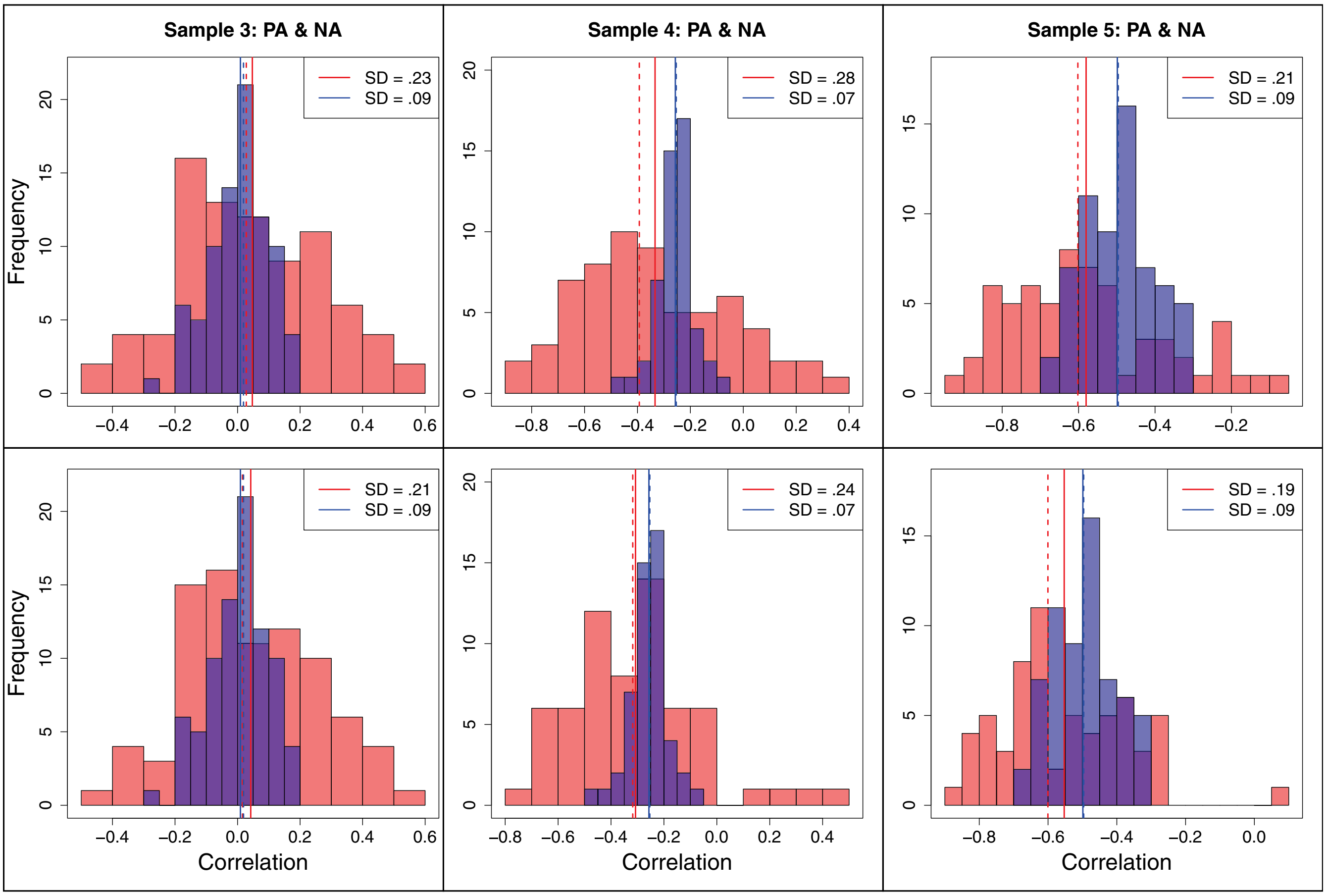

Fig. 3. Histograms for intraindividual (red) and interindividual (blue) correlations between positive affect (PA) and negative affect (NA) in samples 3, 4, and 5. *Top* depicts intraindividual correlations calculated from raw data. *Bottom* depicts intraindividual correlations calculated from data with temporal dependence removed.

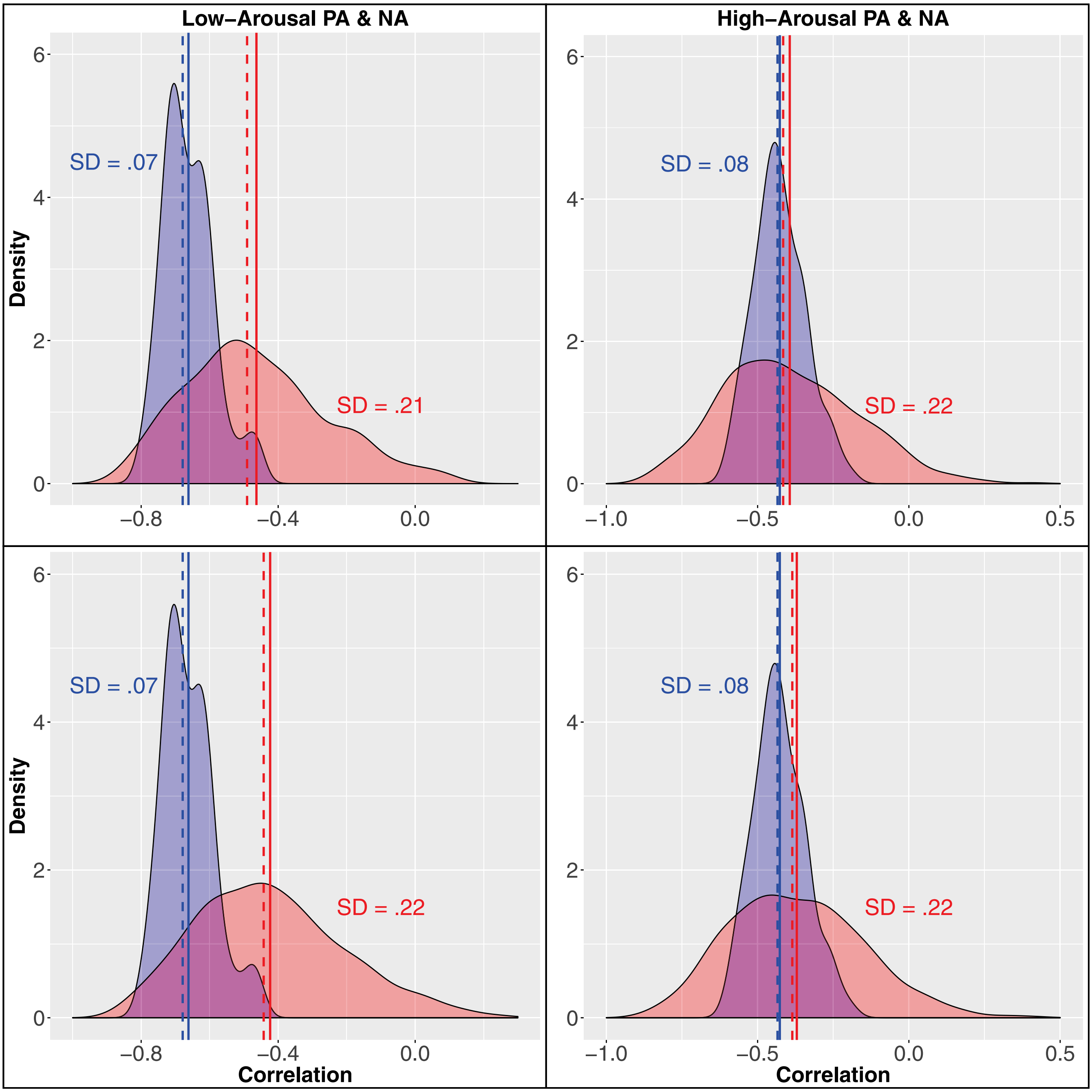

Fig. 4. Density plots for the distributions of correlations in sample 6 between low-arousal positive affect (PA) and negative affect (NA) (*Left*) and high-arousal PA and NA (*Right*). Red indicates distributions related to the 535 individual respondents. Blue indicates the distributions of the 85 cross-sectional comparison datasets. Intraindividual correlations were calculated from raw data (*Top*), and data with temporal dependence removed (*Bottom*).

RSA, with intraindividual estimates derived from raw data on the left and intraindividual estimates derived from AR residuals on the right. The mean intraindividual correlation (and SD) for HR and RSA in sample 2 was $r = -0.35$ (0.22) for raw data, and $r = -0.18$ (0.13) for AR residual data. The mean interindividual correlation was $r = -0.39$ (0.08). The disparity between means was $r = 0.04$ for raw data and $r = 0.21$ for AR residual data. Of note, whereas the intraindividual SD for raw data was 2.75 times larger than the interindividual SD, the SD for AR residual data was only 1.63 times larger, a reduction in magnitude of 59%.

Because the symmetric analysis eliminated more than 800 potential interindividual comparison cases, we reran the interindividual analysis with a sample size of $n = 900$. Results were consistent, with a mean interindividual correlation of $r = -0.40$ and SD of 0.09.

Samples 3, 4, and 5. Consistent with sample 2, a single comparison was calculated for samples 3, 4, and 5. Each sample provided data for PA and NA, although the operationalization of these variables differed between studies (*Method*). Fig. 3 presents the histograms for the intraindividual (red) and interindividual (blue) correlations between PA and NA for samples 3, 4, and 5 (left to right). The top panels depict comparisons calculated with raw intraindividual data, and the bottom panels depict comparisons calculated with AR residual intraindividual data. The mean intraindividual correlations for PA and NA in samples 3, 4, and 5 were $r = 0.05$ (0.23), $r = -0.33$ (0.28), and $r = -0.58$ (0.21), respectively, for raw data, and $r = 0.04$ (0.21), $r = -0.31$ (0.24), and $r = -0.55$ (0.19), respectively, for AR residual data. The mean interindividual correlations for sample 3, 4, and 5 were $r = 0.01$ (0.09), $r = -0.26$ (0.07), and $r = -0.50$ (0.09), respectively. The discrepancy between intraindividual and interindividual mean estimates were only $r = 0.04$, $r = 0.07$, and $r = 0.08$ for raw data and $r = 0.03$, $r = 0.05$, and $r = 0.05$ for AR residual data. However, consistent with sample 1, the intraindividual SDs for samples 3–5 were substantially larger than the interindividual SDs, with ratios of 2.56:1, 4:1, and 2.33:1 for raw data and 2.33:1, 3.42:1, and 2.11:1 for AR residual data. The mean SD ratios across the three studies were 2.96:1 and 2.62:1 for raw and residual data, respectively.

Sample 6. Finally, for sample 6, we calculated the intraindividual correlations for two operationalizations of PA and NA, low-arousal and high-arousal, across 535 participants. Fig. 4 presents the density plots for the intraindividual correlations of low-arousal PA and NA (left) and high-arousal PA and NA (right). The top panels depict comparisons made with raw data, and the bottom panels depict comparisons made with AR residual data. Density plots were employed due to the large discrepancy between intraindividual sample size ($n = 535$) and interindividual sample size ($n = 85$). Density plots reflect the shape of the underlying histograms, independent of the scale of measurement. The mean intraindividual correlation (and SD) for low-arousal PA and NA was $r = -0.46$ (0.21) for raw data and $r = -0.42$ (0.22) for AR residual data. The mean interindividual correlation for low-arousal PA and NA was $r = -0.66$ (0.07). The disparity between means was $r = 0.20$ for raw data and $r = 0.24$ for AR residual data. The mean intraindividual correlation (and SD) for high-arousal PA and NA was $r = -0.39$ (0.22) for raw data and $r = -0.37$ (0.22) for AR residual data. The mean interindividual correlation for high-arousal PA and NA was $r = -0.43$ (0.08). The disparity between means was $r = 0.04$ for raw data and $r = 0.06$ for AR residual data. The ratio of intraindividual SD to interindividual SD was 3:1 for raw data and 2.75:1 for AR residual data.

Discussion

In the present study, we made comparisons between intraindividual and interindividual variation across six independent samples from the United States and the Netherlands. Using intensive repeated-measures data, we compared the relative distributions of individual and group correlations across a number of constructs and data collection paradigms, using both raw data and data residualized against the first-order autoregression. Univariate distributions revealed relatively consistent mean estimates. However, the ratios of intraindividual to interindividual SDs ranged from 3.79 to 13.20, with the former 7.85 times larger than the latter on average. Regarding the expected values of the distributions of bivariate correlations, we found equivocal results—those for which group-to-individual generalizability did and did not apply. That is, we found evidence for the agreement in the central tendency of four of the nine comparisons, and disagreement in the remaining five comparisons. Specifically, there was fairly strong agreement in the expected values for the correlation between PA and NA, as operationalized by samples 3, 4, and 5. Related to this, the operationalization of high-arousal PA and NA in sample 6 exhibited fairly strong consistency across intraindividual and interindividual variation. The average discrepancy between intraindividual and interindividual estimates was $r = 0.06$ for raw data and 0.04 for AR residual data.

However, regarding the variance around the expected value, across all six samples and all nine comparisons the intraindividual covariation exhibited a spread that ranged from 2.09 to 4 times larger than that observed for interindividual estimates when calculated from raw data (mean ratio, 2.84:1) and a range of 1.63–3.43 times larger when calculated from AR residual data (mean ratio, 2.56:1). That is, the SD of correlations for individuals was, on average, nearly three times higher than for groups in the raw data and greater than two-and-a-half times larger in the residual data. Thus, while equivocal support was found for the group-to-individual generalizability of mean estimates, the current study found no support for ergodic agreement between the variances related to intraindividual and interindividual statistical estimates.

There are two important takeaways from these findings: (*i*) the fact that we observe substantial numerical disagreement between the individual and group estimates, and (*ii*) the consequences for generalizing group data to individuals. The former finding is self-explanatory, but worth emphasizing: Aggregated estimates did not consistently agree with individual estimates. While prior work has noted the implausibility of ergodicity in human subjects research, including the provision of mathematical proof (1), the present study provides an empirical demonstration of this effect across multiple settings and constructs. Regarding the generalizability of aggregate estimates, the present findings indicate that correlations between variables within individuals exhibit at least twice as much variation as those found within groups. The present findings suggest that, at best, group estimates can be considered expected values for normally distributed estimates—such as correlations—accompanied by substantially greater variability than the group SD. That is, even in the best-case scenario, we should not think of a correlation in group data as an estimate that generalizes to any given individual in the population. Stated bluntly, this implies that the temptation to use aggregate estimates to draw inferences at the basic unit of social and psychological organization—the person—is far less accurate or valid than it may appear in the literature. Indeed, even the best-case scenario is quite alarming: Only 68% of all individual correlational values fall within a range that would be predicted by group data to cover 99.7% of all possible correlations—a discrepancy of nearly 32%.

The worst-case scenario is clearly dire: It is plausible that inattention to nonergodicity and a lack of group-to-individual generalizability threaten the veracity of countless studies, conclusions, and best-practice recommendations. As our results in just a few studies illustrate, the nature of the threat varies in degree and kind across research domains and measurement modalities. For example, the discrepancies in sample 1 relate to the assessment, classification, and treatment of psychiatric syndromes. Gold-standard treatments for phobic avoidance are built on the assumption that the mitigation of avoidance behavior through exposure—the facilitation of repeated contact between the individual and the feared stimulus—will reduce the experience of fear over time. However, our results indicate that, on average, the strength of the association between fear and avoidance is weaker than group estimates would imply and that, for some individuals, there is no relationship between these phenomena whatsoever. It is thus imperative to test whether individuals who have no relationship represent a distinct natural kind that do not respond to treatment.

Similarly, the strong, positive covariation between depressed mood and worry has been so consistently replicated that some have argued that their respective clinical diagnoses—GAD and MDD—should belong in the same class of mental disorders (38). Consistent with this, the expected value for the group correlation between worry and depressed mood was $r = 0.71$. However, the expected value for the correlation between these constructs within individuals was 44% lower in raw data ($r = 0.40$), was 55% lower in AR residual data ($r = 0.32$), and exhibited an overall range from $r = -0.11$ to $r = 0.89$, suggesting that the putative covariance between these syndromes in the extant literature may underestimate the considerable variability in the relationship across individuals. Thus, taxonomic decisions based on the putative group correlation between these dimensions would likely lack ecological and clinical validity, and may undermine treatment planning and outcomes. These results also suggest that our taxonomies are unlikely to capture the diversity of natural kinds that may exist across humans.

Taken together, we believe the results of the present study point to a continuum of scenarios for identifying nonergodicity by examining the consistency between group and individual correlations and the degree to which conclusions from one domain can be generalized to the other. By framing such tests within the ergodic theorem, which stipulates strict and rigorous conditions for generalizations between levels of analysis, we can address basic issues in biopsychosocial research. The current data further reify the dangers of the ecological fallacy, empirically demonstrating that group-derived estimates should not be considered accurate proxies of individual processes. Regarding Simpson's paradox—that aggregations may misrepresent subgroups or individuals—the present study was consistent with the recommendation of Kievit et al. (39), who stressed that data should be explicitly tested for agreement in estimates across levels of analysis.

Importantly, visual inspection of the distributions of correlations for intraindividual and interindividual data revealed that the latter appeared to better approximate normality than the former. Thus, tests that assume normality may be more appropriate for between-subjects analyses than for within-subjects analyses, and the standard error of the mean may be a more appropriate measure for interindividual variability. Moreover, in addition to incongruities in interindividual and intraindividual statistical moments, nonergodicity in the natural sciences may also be driven by variation in temporal dependencies across individuals in a sample. While our current findings support the case to more carefully consider nonergodicity in general, they hold whether or not temporal dependencies are explicitly accounted for (*SI Appendix*). However, the presence of this variation in real-world data may provide further incentive to thoroughly investigate individual-level variability before making group-derived inferences.

One possible limitation of the present study is that, for some of the data sources, we focused on subsamples of the data to match the relative scale of cross-sectional and intraindividual dimensions. Although this was done to reduce bias when comparing group to individual data, it is important to note that small sample sizes may reduce the agreement between individual and group data because small data batches may not represent the larger population of observations (40). Explicitly using models that are applied to large samples, account for variation across time, and model the influences of covariates may help separate issues due to nonergodicity from issues due to unmodeled variation from other sources. Our results suggest that, for researchers collecting cross-sectional data, it may be worthwhile to collect subsamples of within-subjects data to test for potential nonergodicity. Furthermore, it may be important to apply grouping procedures if categorically different types of intraindividual variability are observed, facilitating scientific discovery and more appropriate generalization claims.

Importantly, we can use other approaches to consider within- and between-person variation to support inferences that are more representative of intraindividual variation where nonergodicity may be observed. Here, we focused on simple approaches that measure properties of variance and covariance, which can, in turn, support more complex data analysis techniques. Numerous, more sophisticated tools exist to examine variance component models (41, 42), multilevel models (43), contemporaneous and lagged structural equation models (44), and time series more generally (45). Our results suggest that, just as simple measures can exhibit significant differences in within- vs. between-subject variation, both the within- and between-subject variation should be explicitly modeled in any approach, and all models should be evaluated in terms of how well they represent individual-level processes before making generalizations from aggregate estimates.

On a final note, with increasing emphasis on the so called "replication crisis" (46), and an acknowledgment that incentive structures can lead to harmful scientific practices (47), the present paper adds a voice to the chorus of scholars calling for increased transparency and accountability. It is possible that the emerging emphasis on open data will help us detect and characterize instances of nonergodicity broadly in the life sciences. For example, open-data efforts have led to increasing researcher interactions and open conversations about replicability and validity in comparative bioinformatics (40), breast cancer prediction (48), and massive computational experiments (49). At the very least, open-data access would allow researchers to examine the extent to which previously published results may evince varying degrees of nonergodicity.

Our focus on a truly fundamental but rarely considered dilemma in human subjects research brings to the fore questions of scientific accuracy and generalizability. Even if incentives are adjusted, our ability to make valid inferences at the level of the individual will require substantial efforts in person-centered science. Encouragingly, with the increasing number of publicly available large-dimensional datasets, we can not only begin to quantify how pervasive the nonergodic threat may be, we can begin to broaden our sciences to quantify and articulate both intraindividual and interindividual differences, and in so doing build models of human physiology and behavior that improve the quality and accuracy of statistical inferences across levels and units of analysis.

ACKNOWLEDGMENTS. We thank all participants of the six studies for their participation and valuable contribution to this research project. Stephen Hinshaw and Rodolfo Mendoza-Denton are thanked for comments on earlier drafts of this manuscript. We are grateful to Steven Woodward, Aidan Wright, Nicole Geschwind, Evelien Snippe, Nancy Nicolson, Marieke Wichers, and Daniela Barge-Schaapveld for sharing data and providing assistance with data preparation. J.D.M. acknowledges support from the Office of the Director and National Institute of Mental Health at the National Institutes of Health (1-DP5- OD021352).

1. Molenaar PCM (2004) A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement* 2:201–218.
2. Molenaar PCM, Campbell CG (2009) The new person-specific paradigm in psychology. *Curr Dir Psychol Sci* 18:112–117.
3. Quine WV (1969) *Natural Kinds. Essays in Honor of Carl G. Hempel* (Springer, New York), pp 5–23.
4. Wilson EO (1998) Consilience among the great branches of learning. *Daedalus* 127: 131–149.
5. Carrier M (2012) *The Completeness of Scientific Theories: On the Derivation of Empirical Indicators Within a Theoretical Framework: The Case of Physical Geometry* (Springer Science and Business Media, Dordrecht, The Netherlands).
6. Hamaker EL, Dolan CV, Molenaar PCM (2005) Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate Behav Res* 40:207–233.
7. Salvatore S, Valsiner J (2010) Between the general and the unique. *Theory Psychol* 20: 817–833.
8. Medaglia JD, Ramanathan DM, Venkatesan UM, Hillary FG (2011) The challenge of non-ergodicity in network neuroscience. *Network* 22:148–153.
9. Molenaar PCM (2008) On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Dev Psychobiol* 50:60–69.
10. Lamiell JT (1981) Toward an idiothetic psychology of personality. *Am Psychol* 36: 276–289.
11. Castro-Schilo L, Ferrer E (2013) Comparison of nomothetic versus idiographic-oriented methods for making predictions about distal outcomes from time series data. *Multivariate Behav Res* 48:175–207.
12. Simpson EH (1951) The interpretation of interaction in contingency tables. *J R Stat Soc B* 13:238–241.
13. Robinson WS (1950) Ecological correlations and the behavior of individuals. *Am Sociol Rev* 15:351–357.
14. Hamaker E (2012) Why researchers should think "within-person": A paradigmatic rationale. *Handbook of Research Methods for Studying Daily Life* (The Guilford Press, New York), pp 43–61.
15. Beltz AM, Wright AGC, Sprague BN, Molenaar PCM (2016) Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment* 23:447–458.

16. Fisher AJ (2015) Toward a dynamic model of psychological assessment: Implications for personalized care. *J Consult Clin Psychol* 83:825–836.
17. Pearl J (2009) *Causality: Models, Reasoning, and Inference* (Cambridge Univ Press, Cambridge, UK).
18. Pearl J (1999) Simpson's paradox: An anatomy (UCLA Cognitive Systems Laboratory, Los Angeles), Technical Report R-264.
19. Boltzmann L (1877) On the relation between the second law of the mechanical theory of heat and the probability calculus with respect to the theorems on thermal equilibrium. *Kais Akad Wiss Wien Math Natumiss Classe* 76:373–435.
20. Fisher AJ, Reeves JW, Lawyer G, Medaglia JD, Rubel JA (2017) Exploring the idiographic dynamics of mood and anxiety via network analysis. *J Abnorm Psychol* 126: 1044–1056.
21. American Psychiatric Association (2013) *The Diagnostic and Statistical Manual of Mental Disorders: DSM-5* (American Psychiatric Association Publishing, Washington, DC).
22. Fisher AJ, Woodward SH (2014) Cardiac stability at differing levels of temporal analysis in panic disorder, post-traumatic stress disorder, and healthy controls. *Psychophysiology* 51:80–87.
23. Wright AG, Hopwood CJ, Simms LJ (2015) Daily interpersonal and affective dynamics in personality disorder. *J Pers Disord* 29:503–525.
24. Wright AG, Simms LJ (2016) Stability and fluctuation of personality disorder features in daily life. *J Abnorm Psychol* 125:641–656.
25. Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: The PANAS scales. *J Pers Soc Psychol* 54: 1063–1070.
26. Barge-Schaapveld DQCM, Nicolson NA (2002) Effects of antidepressant treatment on the quality of daily life: An experience sampling study. *J Clin Psychiatry* 63: 477–485.
27. Snippe E, et al. (2017) The impact of treatments for depression on the dynamic network structure of mental states: Two randomized controlled trials. *Sci Rep* 7:46523.
28. Geschwind N, Peeters F, Drukker M, van Os J, Wichers M (2011) Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *J Consult Clin Psychol* 79:618–628.
29. van der Krieke L, et al. (2016) HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *Int J Methods Psychiatr Res* 25:123–144.
30. Matthews JN, Altman DG, Campbell MJ, Royston P (1990) Analysis of serial measurements in medical research. *BMJ* 300:230–235.
31. Beck N, Katz JN, Tucker R (1998) Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *Am J Pol Sci* 42:1260–1288.
32. DiPrete TA, Forristal JD (1994) Multilevel models: Methods and substance. *Annu Rev Sociol* 20:331–357.
33. Molenaar PCM (2005) Rejoinder to Rogosa's commentary on "A manifesto on psychology as idiographic science." *Measurement* 3:116–119.
34. Durbin J, Watson GS (1950) Testing for serial correlation in least squares regression. I. *Biometrika* 37:409–428.
35. Kessler RC, et al. (2008) Co-morbid major depression and generalized anxiety disorders in the National Comorbidity Survey follow-up. *Psychol Med* 38:365–374.
36. Craske MG, Waters AM (2005) Panic disorder, phobias, and generalized anxiety disorder. *Annu Rev Clin Psychol* 1:197–225.
37. Cacioppo JT, Tassinary LG, Berntson G (2007) *Handbook of Psychophysiology* (Cambridge Univ Press, Cambridge, UK).
38. Watson D (2005) Rethinking the mood and anxiety disorders: A quantitative hierarchical model for DSM-V. *J Abnorm Psychol* 114:522–536.
39. Kievit RA, Frankenhuis WE, Waldorp LJ, Borsboom D (2013) Simpson's paradox in psychological science: A practical guide. *Front Psychol* 4:513.
40. Tyekucheva S, Parmigiani G (2017) Bioinformatic analysis of epidemiological and pathological data. *Pathology and Epidemiology of Cancer*, eds Loda M, Mucci LA, Mittelstadt ML, Van Hemelrijck M, Cotter MB (Springer International Publishing, Cham, Switzerland), pp 91–104.
41. Anderson DA, Aitkin M (1985) Variance component models with binary response: Interviewer variability. *J R Stat Soc B* 47:203–210.
42. Kang HM, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42:348–354.
43. Snijders TA (2011) Multilevel analysis. *International Encyclopedia of Statistical Science*, ed Lovric M (Springer, Berlin), pp 879–882.
44. Kim J, Zhu W, Chang L, Bentler PM, Ernst T (2007) Unified structural equation modeling approach for the analysis of multisubject, multivariate functional MRI data. *Hum Brain Mapp* 28:85–93.
45. Wei WW (2006) Time series analysis. *The Oxford Handbook of Quantitative Methods in Psychology* (Oxford Univ Press, Oxford), Vol 2.
46. Maxwell SE, Lau MY, Howard GS (2015) Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *Am Psychol* 70:487–498.
47. Harris R (2017) *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions* (Hachette Book Group, New York).
48. Bernau C, et al. (2014) Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* 30:i105–i112.
49. Monajemi H, Donoho DL, Stodden V (2016) Making massive computational experiments painless. *2016 IEEE International Conference on Big Data (Big Data)* (IEEE, New York), pp 2368–2373.