Published in final edited form as:

Res Hum Dev. 2017 ; 14(3): 253–270. doi:10.1080/15427609.2017.1340052.

The Questionable Ecological Validity of Ecological Momentary Assessment: Considerations for Design and Analysis

Nilam Ram^{a,b}, Miriam Brinberg^a, Aaron L. Pincus^a, and David E. Conroy^a

^aPennsylvania State University

^bGerman Institute for Economic Research (DIW), Berlin

Abstract

Experience sampling (ESM), diary, ecological momentary assessment (EMA), ambulatory monitoring, and related methods are part of a research tradition aimed at capturing the ongoing stream of individuals' behavior in real-world situations. By design, these approaches prioritize *ecological validity*. In this paper, we examine how the purported ecological validity these study designs provide may be compromised during data analysis. After briefly outlining the benefits of EMA-type designs, we highlight some of the design issues that threaten ecological validity, illustrate how the typical multilevel analysis of EMA-type data can compromise generalizability to “real-life”, and consider how unobtrusive monitoring and person-specific analysis may provide for more precise descriptions of individuals' actual human ecology.

Keywords

intraindividual variability; sequence analysis; entropy

Throughout the last century, researchers have conscientiously sought methods to capture the ongoing stream of individuals' behavior or thought (see Hormuth, 1986; Wilhelm, Perez, & Pawlik, 2012, for historical accounts). Developmentalists, in particular, have contributed much to innovation and refinement of research perspectives and study designs aimed at identifying and parsing the ecology of human development (e.g., Bronfenbrenner, 1977). Given that developmental processes are not constrained to and do not occur only within the laboratory, developmental science has adopted techniques that allow for observation of behavior in natural settings (e.g., home, classroom). Csikszentmihalyi and colleagues (e.g., Csikszentmihalyi & Larson, 1984), for instance, used then recently available beeper technology to sample and study adolescents' psychological reactions to everyday activities and experiences. As technology has advanced, so too has researchers' ability to engage participants in their natural habitats. Ubiquitous computing, handheld and laptop devices,

Address correspondence regarding this manuscript to: Nilam Ram, Department of Human Development & Family Studies, Pennsylvania State University, 407 Biobehavioral Health Building, University Park, PA 16802, USA or nur5@psu.edu.

Author Notes

Nilam Ram, Department of Human Development & Family Studies, Pennsylvania State University and German Institute for Economic Research (DIW), Berlin; Miriam Brinberg, Department of Human Development & Family Studies, Pennsylvania State University; Aaron L. Pincus, Department of Psychology, Pennsylvania State University; David E Conroy, Department of Kinesiology, Pennsylvania State University.

portable global positioning systems, and smartphones that communicate with multiple wearable sensors are providing for collection of real-time data in real-life contexts (Harari et al., 2016; Reis, 2012). With modern mobile computing, it is now easier than ever to launch studies wherein individuals of all ages provide comprehensive reports about hundreds of experiences as they go about their daily lives (e.g., Ram et al., 2014). A pervasive problem, though, is that usually very little is known about the specific settings in which daily-life measurements are obtained. Taking seriously the idea that the human ecology is a complex biopsychosociocultural milieu filled with multiple influences and feedback processes (Bronfenbrenner, 1977), we are somewhat concerned that “ecological momentary assessment” studies may not actually deliver as much of the specific types of validity as they promise.

External validity is generally considered as the extent to which research findings generalize across people, places, and time. Ecological validity is more specifically defined as *the extent to which research findings generalize to settings typical of everyday life* (Stone & Schiffman, 2011). As such, generalization to specific contexts depends heavily on both adequate *sampling* of stimuli, persons, situations, and time (i.e., study design and data collection) and defensible *statistical inference* (i.e., data analysis).

Experience sampling (ESM), diary, ecological momentary assessment (EMA), ambulatory monitoring, and related methods (Bolger et al., 2003; Csikszentmihalyi & Larson, 1984; Mehl & Conner, 2012; Shiffman et al., 2008) can be considered as part of a longstanding research tradition aimed at capturing “the ways a person usually behaves, the regularities in perceptions, feelings and actions” (Fiske, 1971, p. 179). A primary objective of these methods is to sample the ongoing stream of behavior or thought in its natural sequence and occurrence, and thereby, provide an accurate and valid representation of individuals’ function that generalizes to real-world situations. By design, through repeated assessment of individuals’ behavior and experience in their natural environments, these approaches prioritize *ecological validity*.

In this paper, we highlight how the purported ecological validity these study designs provide may be compromised during data collection and/or analysis. Our general thesis is that assumptions embedded in the analytical models typically used to derive inferences from experience sampling data undercut the very elements of ecological validity the approach seeks to address. In the sections that follow we will (1) briefly outline the benefits of EMA-type designs, (2) highlight some design issues that threaten internal and external validity, (3) illustrate how the typical multilevel analysis of EMA-type data compromises ecological validity, and (4) consider how unobtrusive monitoring and person-specific analysis may enhance descriptions of the human ecology. Up front, our intent is not to discourage use of EMA-type study designs. We are heavily invested. Rather, our goal is to better understand how to collect and analyze these data streams – in ways that accurately extract and represent regularities in the human ecology.

Study Design for Ecological Validity

Experience sampling and related study designs are often distinguished by three features (see also Bolger et al., 2003; Shiffman et al., 2008): (a) data are collected in real-world environments, as individuals go about their daily lives, (b) assessments focus on individuals' current or recent state, and (c) individuals complete multiple assessments over time in accordance with a specific time sampling scheme (e.g., interval sampling, momentary sampling, random interval, fixed interval, event-contingent; see Bolger & Laurenceau, 2013; Ram & Reeves, in press). Studies are explicitly designed in ways that provide for generalization to individuals' real lives, reduction of error or bias associated with retrospective reporting, and study of intraindividual change. The intensive longitudinal nature of the data collection allows for examination of individual differences in intraindividual variability, intraindividual covariation, person-situation interactions, change processes, causality, etc. (see Baltes & Nesselroade, 1979; Hamaker, 2012; Stone et al., 2007). Many researchers have turned to EMA-type designs because of the advantages they afford over cross-sectional and experimental designs for understanding how psychological and behavioral processes manifest within the context of naturally occurring changes in individuals' environments – *ecological validity*. A basic presumption of the entire method is that, because EMA-type data are collected – *in situ* – in individuals' natural environments, the findings derived from them should be more generalizable to real-world, real-life experiences and processes (Shiffman et al., 2008).

Common Threats to Ecological Validity

Of course, EMA-type designs are prone to many of the same limitations raised in other types of studies (e.g., survey- and laboratory-based designs), including the validity / reliability / sensitivity of measurement instruments, (lack of) compliance with the protocol, and reactivity to assessment procedures. Labeling a study as an EMA study does not absolve us from careful consideration of the sampling issues central to generalization. For example, in a typical EMA design, individuals are pinged every few hours with a questionnaire and asked to report on their current feelings and behaviors. The subjective nature of the reports brings a variety of potential saliency, accessibility, recency, and social desirability biases that compromise validity and reliability of the repeated measures (Schwartz, 2007; Hufford, 2007). Choice of response scale (e.g., 5 point Likert vs. 1000-point slider) compromises both reliability and validity of measurement by filtering sensitivity to both within-person change and “noise”. For example, when using a clinical depression scale to measure depressive symptoms in healthy populations, most participants will score 0 or 1 on a 20-point scale. Similarly, sensors tracking specific types of behavior (e.g., physical activity) may not accurately reflect different individuals' movement or sleep/wake patterns as intended. Devices designed for a specific purpose or population may (not) provide for reliable or valid measurement of the particulars of each individuals' real-world settings. As such, findings may only generalize to a very specific subset of human ecology.

Participants' lack of *compliance* with a carefully designed EMA-type study protocol is also of persistent concern. In particular, when individuals do not provide enough reports, data samples do not provide adequate coverage of individuals' natural environments. Reports of

compliance in EMA-type studies range from 11% to over 80 or 90% (Stone et al., 2002; Hufford & Shields, 2002), with differences noted across both the specifics of the measurement protocol (e.g., intensity of reporting) and the practicalities surrounding participant incentives, training, and data monitoring. Suggestions for keeping compliance high generally point towards the declining expense of and increasing comfort with personal electronic devices and the possibility of using real-time data transfer from mobile devices to support monitoring of report completion and provision of “motivational” support (e.g., Ram et al., 2014; but see also Newcomb et al., 2016). Participants’ *reactivity* to EMA-type data collection procedures can also compromise ecological validity. When the assessment protocol influences or prompts change in individuals’ behavior or feelings, data samples are repeatedly obtained in situations that no longer represent individuals’ natural ecology. They are obtained in a new reality. Results from studies examining participants’ reactivity to EMA-type protocols are mixed (see Barta, Tennen, & Litt, 2012 for a review). Some studies note how their protocols may have prompted behavioral change through reflection, social desirability, fatigue, or feedback processes (e.g., changing eating behavior after seeing one’s daily step count), while other studies find no evidence of reactivity (Reynolds, Robles, & Repetti, 2016). Our general impression is that potential for reactivity differs across content areas, protocols, individuals, and time scales. For example, hourly reporting about social interactions is unlikely to provoke change in interaction partners, while daily reporting of specific foods consumed is more likely to provoke dietary change. Researchers using EMA-type designs must explicitly consider if and how they and their measurement devices become “participant observers” in individuals’ natural ecologies (Corbin & Strauss, 2008). Pilot studies may be extremely useful for identifying potential threats to ecological validity.

Prospects for Unobtrusive Monitoring

Ironically, many of these potential threats to ecological validity may be addressed by reducing the “participant” part of “participant observer”. Compliance issues and reactivity effects can be eliminated by reducing interaction with the participant. In traditional EMA designs, participants’ active engagement with researcher-designed measurement instruments requires substantial time and effort (e.g., questionnaires every few hours). The digitization of information / communication that has accompanied technological advances, however, means more of individuals’ daily life experiences can be observed through passive monitoring of digital traces. Participants’ passive provision of digital traces greatly reduces participant burden – and thereby practically eliminates validity problems related to compliance and reactivity. All participants need to do is live – while researchers (and/or hospitals, corporations, governments, etc.) extract information about their human ecology from the scatter of digital crumbs left behind.

Design of unobtrusive EMA-type studies is well underway. Researchers are using paradigms wherein mobile devices (e.g., laptops, smartphones, cameras) are used to sample a wide variety of information streams, including audio (electronically activated recorder; Mehl, 2012), communication logs (calls; Saeb et al 2014), video (in-home cameras; Kim & Teti, 2014), geographic location (smartphone GPS), illumination (smartphone light sensors; Wang et al. 2014); proximity to others (RFID badges; Salathe et al, 2010), internet browsing (browser logging), and laptop use (screenshot logging; Mark et al., 2005; Yeykelis,

Cummings, & Reeves, 2014). Following the same arguments made for EMA-type studies (e.g., Shiffman et al., 2008), findings derived from these digital sweeps are considered generalizable to individuals' real-world, real-life experiences and processes – i.e., ecologically valid. Less clear in some cases, though, is how the data map back to existing behavioral and psychological typologies. Paralleling the challenges faced when EMA-type designs were first being introduced (see Hektner, Schmidt, & Csikszentmihalyi, 2007), the validity burdens shift back to the researcher – and the challenging task of deriving valid, reliable, and sensitive measurement inferences from rich and noisy digital streams.

Time Sampling as a Threat to Ecological Validity

A unique, or at least prominent, issue for both active and passive EMA-type studies is embedded in choice of time sampling scheme. Is the sampling interval appropriate for the phenomenon being studied? Should individuals be pinged with a questionnaire every 30 minutes or every 2 hours? Should laptop screenshots be taken every 5 seconds or every 60 seconds?

Time sampling schemes are often difficult to determine because theory or knowledge of the actual time scale on which behaviors occur in the real world is underdeveloped or unknown. Generally, however, sampling schemes fall into two categories (see Ram & Reeves, in press). With an *interval sampling* scheme, a researcher or participant records the number of times a specific behavior was observed during a predetermined interval of time (e.g., a 24-hour period). An advantage of interval sampling is the rich information obtained about the frequency and duration of behaviors that have a clearly defined beginning and end. A disadvantage is that interval sampling requires researchers and/or collection systems' undivided and continuous attention (or some kind of retrospection). It can also present local storage challenges as the dataset increases in size to log frequent observations. In contrast, with a *momentary sampling* scheme, the presence or level of a specific behavior is observed intermittently at predetermined moments in time (e.g., each hour on the hour). An advantage of the momentary sampling method is the low effort required. The researcher or participant is only engaged at specific points in time, and thus does not need to attend to or continuously monitor behavior. The disadvantage is that behaviors occurring in the intervals between assessments are completely missed. Research on sampling scheme choice has rightfully concentrated on parsing reliability of interval or momentary sampling schemes in situations where ground-truth continuous data are already available (e.g., video recordings of children's classroom behavior; Radley, O'Handley, & Labrot, 2015). As might be expected, differences are noted across the specific sampling schemes considered, the particular behavior of interest, and the type of individuals being studied (Kearns, et al., 1990; Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007; Powell, Martindale, & Kulp, 1975; Powell, Martindale, Kulp, Martindale, & Bauman, 1977). In sum, knowledge about what time sampling scheme to use in which situation is fuzzy, even though the length of the assessment intervals or the intervals between momentary assessments has important implications for the quality of inference.

Generalization to real-world experience – ecological validity – is compromised when the data samples provide an incomplete picture of the temporal sequences of events or

experiences. For example, study of diurnal cycles is simply not possible with a once-per-day momentary sampling scheme. Abstracting from the formal mathematics from which sampling intervals can be derived (i.e., Nyquist-Shannon sampling theorem; Nyquist, 2002), interval sampling designs should use intervals that are much longer than the length of the behavior of interest, so that multiple (e.g., > 3) occurrences could occur within any given interval. In many daily diary studies, for instance, participants report on the previous 24-hour interval – a length that is sufficient for multiple stressor events to have occurred. In momentary sampling designs, the interval between assessments should be short relative to the rate of change of the phenomena of interest; that is, the behavior should span multiple assessments. For example, in many experience sampling studies, participants report on stressor experiences every 2 hours – with the implication that stressors span multiple assessments (i.e., last many hours).

Here also, in addition to piloting protocols that sample at different cadences (e.g., 2-hours 7-hours), unobtrusive monitoring can enhance researchers' ability to obtain robust and comprehensive sampling of stimuli, situations, and time. For example, the StudentLife study at Dartmouth made use of sensors and microphones embedded in mobile phones that provide for continuous and unobtrusive monitoring of stress in real-life situations through collection and analysis of activity, sleep, location, and conversations (Wang et al., 2014). Speech patterns in particular provide for detection of stress, without any need for participants to provide self-reports (Lu et al., 2012). As the field moves toward unobtrusive EMA-type studies that track behavior continuously, we will obtain a more complete and informative picture of the temporal sequences of real-world behavior – and thereby increase potential for ecological validity. Along the way, researchers will need to demonstrate the reliability and validity of the unobtrusive measurements – and likely embark a whole series of studies demonstrating congruence between participants' subjective perceptions and measures derived from the unobtrusive monitoring (e.g., Hovsepian et al, 2015). The new digital data streams may eventually make it possible to tackle the research questions discussed in this paper using unobtrusive measures.

Multilevel Modeling Assumptions as a Threat to Ecological Validity

As noted earlier, ecological validity is better when research findings generalize to settings typical of everyday life. Notably generalization “to the settings and people common in today's society” (Wegener & Blankenship, 2007) depends on both adequate sampling and *statistical inference* (i.e., data analysis). One of the most predominant approaches for analysis of EMA-type data is the multilevel model (e.g., Bolger & Laurenceau, 2013). For example, and portending our empirical example, EMA-type (e.g., daily diary) data may be used to study individual differences in “stressor reactivity” – a construct defined as the within-person association between stressor experience (stressor-day vs. no-stressor-day) and affective well-being (e.g., daily negative affect, NA). We, and many others (e.g., Almeida, 2005; Koffer et al., 2016) use multilevel models of the form,

$$Negative\ Affect_{it} = \beta_{0i} + \beta_{1i}StressorDay_{it} + e_{it} \quad 1$$

where the repeated measures of negative affect for person i at occasions t are modeled as a function of β_{0i} , a person-specific *baseline NA* (intercept) coefficient; β_{1i} , a person-specific *stressor-reactivity* coefficient that indicates the extent a person's NA systematically differs between no-stressor (=0) and stressor (=1) days; and e_{it} , residual error that is assumed normally distributed. The person-specific baseline (intercept) and stressor reactivity coefficients are, in turn, modeled as

$$\beta_{0i} = \gamma_{00} + \dots + u_{0i} \quad 2$$

$$\beta_{1i} = \gamma_{10} + \dots + u_{1i} \quad 3$$

where γ_{00} and γ_{10} describe the “prototypical person” in the sample and u_{0i} and u_{1i} describe between-person differences that are assumed multivariate normally distributed with variances $\sigma_{u_0}^2$, $\sigma_{u_1}^2$ and correlation $r_{u_0u_1}$. In our example scenario, the specific parameter of greatest interest is γ_{10} , which indicates the prototypical person's “stressor reactivity”. Additional within-person predictors can be added at either the within-person or between-person levels, but being peripheral to our current focus, are not written out explicitly here.

The multilevel model, in this case one focused on the relation between a single predictor variable (*dailystressor_{it}*) and the outcome variable (*negativeaffect_{it}*), implicitly maps the data to what Campbell and Stanley (1963) referred to as *The Equivalent Time Samples Design* (Design 8, pp. 43–46). In brief, pushing beyond the usual form of experimental designs, Campbell and Stanley outlined single group quasi-experimental designs wherein two equivalent samples of occasions, one in which the experimental variable was present (stressor days; X_1) and in other of which it is absent (non-stressor days; X_0), are used to examine the effects of experimental variable (daily stressors). Following their notation, the equivalent time samples design (for occasions nested within a person) is structured as follows,

$$X_1 O \ X_0 O \ X_1 O \ X_1 O \ X_0 O \ X_0 O \ \dots,$$

where the paired Xs and Os indicate that both the “eXperimental variable” and “Outcome variable” are assessed at every repeated assessment. The random temporal sequencing of levels of X is intended, to ensure that a simple alternation of conditions does not get confounded with other periodicity effects (e.g., daily, weekly, or monthly cycles).

As noted explicitly in the description provided above, analyses of EMA-type data invoked in a multilevel modeling framework are specifically structured to examine the differences in the within-person means across levels of the “experimental variable” ($X = \textit{dailystressor}_{it}$) using a multi-person ensemble of within-person time-series “experiments” where stressors were introduced and withdrawn multiple times. In contrast to a balanced experimental design where each individual is equivalently exposed to the daily stressor treatment, the quasi-

experimental design uses randomization to equate the time periods in which the treatment was present to those in which it was absent. In both cases, the analytical framework assumes the daily stressors are manipulated or distributed such that, *for each individual*, some days are no-stressor days *and* some days are stressor days.

Usually, the intent of EMA-type designs is to obtain a *random sample* of experience that might be considered and treated analytically as representative of individuals' natural human ecology. From an ecological validity perspective, and in line with Brunswik's (1955) original points about *representative design*, this means the "experimental variables" should be randomly sampled in ways that directly match how the variables are distributed in real life. For the study of daily stress reactivity, this implies we should consider both individuals' total exposure to daily stressors and the temporal distribution (i.e., sequence) of their exposure. Our concern is that qualitative differences in (temporal) distribution of daily stressors preclude generalization of findings derived from a typical multilevel analysis of EMA-type data to individuals' real lives. Specifically, heterogeneity in the distribution of individuals' natural exposure to daily stressors suggests that individuals should not be lumped together in an aggregate analysis. When individuals with different exposure patterns are analyzed together, we undermine the very ecological validity that EMA-type designs intend to prioritize.

Well before the advent of multilevel modeling, Campbell and Stanley (1963) highlighted the basic concern we are struggling with now. After noting that generalization from intensive repeated measures designs is limited to frequently tested populations (reactivity to procedures concern noted earlier), they elaborate on the main issue we raise here (which they refer to more generally as *multiple-X interference*),

"The effect of X_1 , in the simplest situation where it is being compared with X_0 , can be generalized only to conditions of repetitious and spaced presentations of X_1 . No sound basis is provided for generalization to situations in which X_1 is continually present, or to the condition where in which it is introduced once and once only. In addition, the X_0 condition or the absence of X is not typical of periods without X in general, but is only representative of absences of X interspersed among presences. ... Thus, while the design is usually internally valid, its external validity may be seriously limited for some types of content" (pp. 44).

The implication is that we need to carefully examine whether the presumed random pattern of absences and presences of the experimental variable generalizes across occasions (within-person) *and* generalizes across persons. By simple illustration, generalization of findings derived from weeks or persons wherein stressors are distributed across the seven days of the week [Mo, Tu, We, Th, Fr, Sa, Su] according to Pattern A (weekday-only stressors) = [1,1,1,1,1,0,0] should not be generalized to weeks or persons with Pattern B (weekend-only stressors) = [0,0,0,0,1,1]. Particularly problematic, of course, are cases where there is no "turbulence" at all. Pattern C (always no-stressor) = [0,0,0,0,0,0,0] certainly does not generalize to Pattern D (always stressor) = [1,1,1,1,1,1,1], and neither actually provides the possibility to obtain an estimate of stress-reactivity. We can question whether the stress-reactivity construct is even relevant in such cases. More generally, it seems that when the

structure of *The Equivalent Time Samples Design* differs systematically across groups of occasions (e.g., “measurement bursts”) or persons – i.e., non-equivalence (or non-randomness) in the temporal distribution/sequence of $dailystressor_{it}$ – generalizations (i.e., claims of ecological validity) may not be justified. Beyond the usual threats to ecological validity (e.g., compliance, reactivity), we are concerned that the standard multilevel modeling approach used in analysis of EMA-type data presumes an ecological validity not present in the naturalistic study design. Specifically, the pooling together of heterogeneous occasions or persons together into a single model likely imbues the analysis with an assumption about ecological validity – generalization across natural environments – that is unwarranted.

An Illustration of an EMA-type Design’s Ecological Validity Problem

To illustrate the problem and the implications for the standard multilevel analysis, we use data collected via smartphone-based questionnaires from 150 individuals participating in an EMA-type study with three 3-week measurement bursts (see Ram et al., 2014, for further details). We make use of data from the first burst of measurement, with additional study of contrast in ecological validity between the first week of data (invoking a typical EMA-type design being used in the literature, e.g., National Study of Daily Experiences) and the entire three-week burst of data.

In Figure 1, each row in the main panel depicts individuals’ naturally occurring experience of daily stressors (see Koffer et al., 2016 for assessment details) across the first week of assessment. Temporal distribution of each individuals’ exposure to different levels of the “experimental variable” ($dailystressor_{it}$) is shown as an ordered sequence of light orange = non-stressor day ($dailystressor_{it} = 0$) and dark orange = stressor day ($dailystressor_{it} = 1$) cells (white = missing). Similarities and differences among individuals’ stressor ecology are evident, both day-to-day (horizontally within-person) and person-to-person (vertically between-persons).

Looking at the individual sequences in Figure 1, we see there are a substantial number of cases that are extremely problematic for study of stressor reactivity, specifically, the sequences that lack variance in stressor experience. At the very bottom are $n = 2$ individuals who only experienced no-stressor days, Pattern C = [0,0,0,...,0,0,0]. Just above them are $n = 91$ individuals who only experienced stressor days, Pattern D = [1,1,1,...,1,1,1]. And finally, the top $n = 57$ rows are individuals who only experienced some non-stressor and stressor days. These three sets of individuals clearly have qualitatively different life experiences in the time period that was sampled. Thus, as per Campbell and Stanley (1963), there is “no sound basis” for generalization from one group’s “real life” to the other’s.

Further, the *never-stressor* and *always-stressor* patterns (which constitute 62% of the sample) have none of the variance needed for quantification of daily stressor-reactivity (β_{1i} from Equation 1). Person-specific regression of the corresponding $negativeaffect_{it}$ scores on $dailystressor_{it}$ are untenable due to non-variance of the predictor, and inclusion in a multilevel analysis of stressor-reactivity seems questionable. To illustrate, parameter estimates (β_{0i} on x -axis, β_{1i} on y -axis) derived from 150 person-specific regressions (each

based on ~7 observations) are plotted in the upper panel of Figure 2 as blue points. Failure of fit for individuals due to non-variance of the predictor means that estimates are only available for $n = 57$ persons. In parallel, Bayes empirical estimates of β_{0i} and β_{1i} derived from the multilevel model of Equations 1 to 3 (based on 1042 repeated observations nested within 150 persons) are shown as black points, with the “prototypical person” implied by the fixed effects model parameters ($\gamma_{00} = 12.83$ and $\gamma_{10} = 9.85$) shown as a large red point. Light gray lines connecting estimates from the two analyses illustrate the compression invoked by the multilevel model, with unconnected points indicating where the *never-stressor* and *always-stressor* individuals got placed within the assumed multivariate normal distribution of between-person differences. The extremely high correlation ($r = .94$) between baseline negative affect (intercepts, β_{0i}) and stressor-reactivity (slopes, β_{1i}) scores derived from the multilevel model indicate that things are going haywire in the fitting process. Clearly, pooling of qualitatively different stressor ecologies, and generalization to a “prototypical” human ecology is problematic. This is, of course, not a new discovery. As highlighted by the deep literature on mixture modeling (e.g., McLachlan & Peel, 2000) and idiographic analysis (e.g., Molenaar, 2004; Nesselroade & Molenaar, 2016), aggregation is often unwarranted. Here, in the context of a discussion about ecological validity and EMA-type designs and analysis, we again highlight the basic problem with pooling individuals into a single analysis. After looking “under the hood” in our own data, we are concerned that other studies may be facing similar problems. Even in the often analyzed and very often cited 8-day National Study of Daily Experiences, for example, 10.7% ($n = 216$ of 2022) of persons in the sample are problematic *never-stressor* and *always-stressor* cases.

Quantitative and Qualitative Examination of a Study’s Ecological Validity Potential

With intent to better understand the opportunities for and limitations on ecological validity that exist within a given set of EMA-type data, we illustrate how a quantitative (turbulence) and a qualitative (clustering) examination (i.e., pre-processing) of the homogeneity/heterogeneity of individuals’ human ecologies may indicate which groups of individuals and/or occasions might be pooled for separate group-specific (or person-specific) analysis and inference.

Turbulence.

As noted earlier (quote from Campbell & Stanley, 1963), the effect of the “experimental variable” can be generalized only to similarly repetitious and spaced presentations of specific situations (i.e., temporal distribution of X_1). In our illustration, this means we need to carefully examine whether the presumed random pattern of absences and presences of daily stressors generalizes across persons (and/or groups of occasions = bursts of measurement). There are many ways to quantify the amount and distribution of individuals’ *dailystressor_{it}* time-series (e.g., intraindividual mean, standard deviation, and entropy; see Ram & Gerstorf, 2009). Here, we make use of a “turbulence” metric that explicitly considers both the range of values and the temporal sub-sequencing of those values (Elzinga & Liefbroer, 2007). *Turbulence* of a specific sequence, Z_i , is specifically useful here because it accounts for both the number of distinct subsequences, $\phi(Z_i)$, and the variance of the

durations spent in distinct states, $s_t^2(Z_i)$ – i.e., the repetition and duration/spacing of the experimental variable. Formally,

$$turbulence(Z_i) = \log_2 \left(\phi \left(Z_i \right) \frac{s_{t,max}^2(Z_i) + 1}{s_t^2(Z_i) + 1} \right) \quad 4$$

with $s_{t,max}^2(Z_i) = (T_i - 1)(1 - \bar{\tau}(Z_i))^2$, where T_i is the total length of the sequence and $\bar{\tau}(Z_i)$ is the average the state durations in that sequence (see also Gabadinho et al., 2011). By analogy, turbulence captures the intuitive relative ordering one might have when standing on a beach looking at the ocean waves. The sequence of waves may be characterized by relative calm with small waves (low turbulence), large regularly-timed waves (low-to-moderate turbulence?), or an irregular mix of small and large waves (high turbulence).

For clarity of interpretation, the sequences in the middle panel of Figure 1 are ordered by their $turbulence_i$ scores, low-turbulence sequences at the bottom and high-turbulence sequences at the top. Across the 150 persons, turbulence scores, shown in the middle panel of Figure 1 range from 1.0 (the minimum possible value) to 5.04 ($M = 1.89$, $SD = 1.29$), but have a distinctly non-normal distribution. We see the over-preponderance of $turbulence_i = 1.0$ scores for the *never-stressor* and *always-stressor* persons' sequences in the histogram on the right side of Figure 1. This quantitative metric suggests these $n = 93$ “non-turbulent” individuals have distinct stressor ecologies that can be separated from the more turbulent and relatively normally distributed stressor ecologies in the top portion of the figure. Whether pooling of, or generalization across, all individuals with $turbulence_i > 1.0$ is warranted is not yet known (as far as we can tell, there are not yet any studies/simulations), but subsequent analyses can be informed by the now quantified differences in the observed temporal distributions, and/or turbulence can be used as a moderator in the analysis (e.g., as predictor in Equations 2 and 3). Generalization to “real-life” (ecological validity) would then be explicitly *conditioned* on the turbulence of individuals' natural ecology.

Clustering.

The turbulence metric provides for a quantitative examination of the homogeneity/heterogeneity in the presumed random pattern of absences and presences of daily stressors. Alternatively these similarities/differences can also be examined using a qualitative approach. The objective then is to identify subgroups of individuals with qualitatively similar/different stressor ecologies – so we can make informed decisions about the opportunities for pooled analysis and the limitations on ecological validity. Aware that longer within-person time-series and more sensitive measurement and conceptualization of daily stressors may provide for more robust sampling of individuals' natural ecologies, we examine how a count variable, the *number of daily stressors* individuals' checked from a 14-category list, is distributed across the first 21-days of assessment. The expanded temporal sequence of each individuals' exposure to different levels of the ordinal “experimental variable” ($\#ofdailystressors_{it}$) is shown in the main panel of Figure 3, with darker colors indicating a greater number of daily stressors. Again, similarities and differences among

individuals' stressor ecologies are evident, both week-to-week (horizontally within-person) and person-to-person (vertically between-persons).

Seeking to cluster together individuals with similarly repetitious and spaced presentations of the “experimental variable” (i.e., total and temporal distribution of # of daily stressors we make use of a method specifically designed for identification of ordered patterns – *sequence analysis* – which emerged for analysis of DNA sequences (Sankoff & Kruskal, 1983), was adapted for study of employment sequences (Abbott, 1995), and most recently to dyadic EMA-type data (Brinberg, Ram, Hülür, Brick, & Gerstorf, in press). The general principle is to determine how much editing is needed to turn one individuals' EMA design (i.e., stressor ecology) into another's design. When very few edits are needed, we may consider the two individuals' natural exposure to situational variance similar. However, when many edits are needed, the two individuals' environments may be qualitatively different. Formally, dynamic programming and optimal matching (e.g., Needleman & Wunsch, 1970) are used to efficiently obtain the minimum *edit distance* between each pair of individual sequences. The resulting distance matrix is then used to identify qualitatively distinct clusters using standard hierarchical cluster analysis methods. In practice, we took the 150 sequences shown in Figure 3, ran them through the TraMineR package in R (Gabadinho, et al., 2011; Studer & Ritschard, 2015; scripts eventually available at <https://quantdev.ssri.psu.edu>) to obtain the 150×150 distance matrix, and subsequently identify 5 clusters of individuals with similar temporal distribution of daily stressors (Ward, 1963; single linkage method). Within-cluster similarity and between-cluster differences are highlighted in Figure 3. Most dramatically, the natural environments in which individuals from Cluster 4 live are quite different from the natural environments in which individuals from Cluster 3 and Cluster 5 live. From an ecological validity perspective, findings from one group should not be generalized to the other. Paralleling the identification of qualitatively different clusters, the distribution of turbulence scores shown in the right panel of Figure 3 has some multimodality – further indicating the sample may have been drawn from multiple populations. In sum, pre-analysis of the experimental variable suggests that ecological validity would be compromised if all individuals were pooled into the same analysis. To underscore this point, we highlight that both cluster membership and turbulence are related to age ($F_{4,145} = 7.35, p < .001; r = -.36, p < .001$); meaning that the younger and older individuals in the sample may have participated in what could be viewed as substantially different EMA studies, with systematically different exposures to the “experimental” variable.

To examine the ecological validity implications a bit further, we derived person-specific parameter estimates from 150 person-specific regressions of the corresponding *negativeaffect_{it}* scores on *#ofdailystressor_{it}* (e.g., Equation 1) which are shown as blue points in the lower panel in Figure 2 (β_{0i} on *x*-axis, β_{1i} on *y*-axis). This time, using an EMA-type design with longer time-series and an ordinal experimental variable, only $n = 3$ cases are untenable because of non-variance of the predictor. While their inclusion in a multilevel analysis of stressor-reactivity is still questionable, it will at least not cause the severe analytical problems we saw in the 7-day, binary experimental design. In parallel to the person-specific analyses of the 21-day ordinal data, Bayes empirical estimates derived from the multilevel model of Equations 1 to 3 (based on 3107 repeated observations nested within 150 persons) are shown as black points, with the “prototypical person” implied by the fixed

effects model parameters ($\gamma_{00}=16.17$ and $\gamma_{10}=2.95$) shown as a large red point. Light gray lines connecting estimates from the two analyses illustrate the compression invoked by the multilevel model, and highlight similarity in the potential inferences made from the collection of person-specific analyses and the pooled analyses (correlation of the two β_{0i} estimates and two β_{1i} estimates were .90 and .70, respectively). In other words, no red flags emerge in the fitting process. We got a viable result that would normally be assumed to generalize across the sample of persons and situations or to a “prototypical human ecology” – which in this case might be considered as the weighted average, median, or modal ecology derived from the 5 clusters shown in Figure 3. However, this approach makes us uneasy. Our pre-analysis suggested clearly that, when using a single multilevel model, we are pooling qualitatively different stressor ecologies. We may, by opting for the person-specific analyses ($N=150$ separate regressions), avoid the potential unwarranted generalization to a “real life” that an individual will never actually experience. An intermediate approach would be to *condition* the generalization on cluster membership, in the same way that we might condition on age. Indeed, when cluster membership is included as an additional predictor in the multilevel model, we find that stressor reactivity is moderated by cluster membership ($\chi^2=4.74, p=.03$); further indication that generalization of findings should be limited to cluster-specific (and what may perhaps be age-specific or group-invariant) human ecologies.

Prospects for Person-specific Modeling

Campbell and Stanley (1963) noted that most experiments employing *The Equivalent Time Samples Design* use relatively few repetitions of each experimental condition, and highlighted how observation across just a few occasions contrasted with the “emerging extension of sampling theory represented by Brunswik” – *ecological validity*. They note explicitly how larger, representative, and equivalent random samplings of time provide for better statistical precision – in that when X_1 is only presented once, differences between X_0 and X_1 are less prone to bias from extraneous events. Portending the nested structure of multilevel models, they viewed it “essential that at least two occasions be nested within each treatment” (p. 45). As seen in Figure 1, there may, in a short 7-day design, be many cases where the essential manipulation (which can be quantified in this case as turbulence $> \sim 2.4$) is never observed. This problem is alleviated when moving to a 21-day design (although not all individuals are exposed to all levels of *#ofdailystressors_{it}* at least twice). Indeed, as the observation period lengthens all individuals may eventually be exposed to the entire range of ecologies. An irony emerges.

The multilevel model can be useful when the number of repeated measures is small. Information from the time-series of similar individuals is used to infer relations among variables. That is, between-person information is used to derive more precise within-person estimates. However, as pointed out in our 7-day example, when the individuals are dissimilar, the model goes haywire. Availability of more occasions (and better measurement) corrects, or at least alleviates the problem. As the number of occasions increases, so too does the opportunity to observe differences and similarities. For example, the $n=91$ *always-stressor* patterns in Figure 1 are now distributed across differentiated clusters in Figure 3. The unique aspects of individuals’ human ecologies manifest, and the need to borrow information is reduced – as illustrated by the fact that it was not possible to conduct person-

specific analyses for $n = 93$ (62%) individuals with 7-day data, but only $n = 3$ (2%) individuals with 21-day (ordinal) data. The person-specific regression estimates provide nearly the same picture as the multilevel model, without pooling across clusters. As the number of occasions increases – within-person time-series of 100 or 200 – the information-borrowing aspect of the multilevel model is no longer needed. With long time-series (i.e., adequate sampling), the person-specific models are viable and capture the within-person relations accurately. Pooling is no longer necessary. Each individual can be treated as a unique individual with a unique human ecology. Person-specific findings can then remain person-specific – ecologically valid and without need for formulation of a “prototypical” person/ecology or questionable generalization to individuals living a different life.

Acknowledgements

This work was supported by the National Institute on Health (R01 HD076994, RC1- AG035645, R24 HD041025, UL TR000127), the Penn State Social Science Research Institute.

References

- Abbott A (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21, 93–113.
- Almeida DM (2005). Resilience and vulnerability to daily stressors assessed via diary methods. *Current Directions in Psychological Science*, 14, 64–68.
- Baltes PB, & Nesselroade JR (1979). History and rationale of longitudinal research In Nesselroade JR, & Baltes PB (Eds.), *Longitudinal Research in the Study of Behavior and Development* (pp. 1–39). New York, NY: Academic Press.
- Barta WD, Tennen H, & Litt MD (2012). Measurement reactivity in diary research Mehl MR, & Conner TS (Eds.), *Handbook of research methods for studying daily life* (pp. 108–123). New York, NY: Guilford Press.
- Bolger N, Davis A, & Rafaeli E (2003). Diary methods: capturing life as it is lived. *Annual Review of Psychology*, 54, 579–616. doi: 10.1146/annurev.psych.54.101601.145030
- Bolger N, & Laurenceau JP (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford Press.
- Brinberg M, Ram N, Hülür G, Brick TR, & Gerstorf D (in press). Analyzing dyadic data using grid-sequence analysis: Inter-dyad differences in intra-dyad dynamics. *Journal of Gerontology: Psychological Sciences*.
- Bronfenbrenner U (1977). Toward an experimental ecology of human development. *American Psychologist*, 32, 513–531.
- Brunswik E (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217.
- Campbell DT, & Stanley JC (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: RandMcNally.
- Corbin J, & Strauss A (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Thousand Oaks, CA: Sage.
- Csikszentmihalyi M, & Larson R (1984). *Being adolescent: Conflict and growth in the teenage years*. New York, NY: Basic Books.
- Elzinga CH, & Liefbroer AC (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population*, 23, 225–250.
- Fiske D (1971). *Measuring the concept of personality*. Chicago. IL: Aldine.
- Gabardinho A, Ritschard G, Muller NS, & Studer M (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1–37.

- Hamaker EL (2012). Why researchers should think “within-person”: A paradigmatic rationale Mehl MR, & Conner TS (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford Press.
- Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, & Gosling SD (2016). Using smartphones to collect behavioral data in psychological Science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, 11, 838–854. [PubMed: 27899727]
- Hektner JM, Schmidt JA & Csikszentmihalyi M (2007). Epistemological foundations for the measurement of experience In Hektner JM, Schmidt JA, & Csikszentmihalyi M (Eds.) *Experience sampling method: Measuring the quality of everyday life* (pp. 3–15). New York, NY: Sage.
- Hormuth SE (1986). The sampling of experiences in situ. *Journal of Personality*, 54, 262–293.
- Hovsepian K, al’Absi M, Ertin E, Kamarck T, Nakajima M, & Kumar S (2015, 9). cStress: Towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 493–504). ACM.
- Hufford MR, & Shields AL (2002). Electronic diaries: an examination of applications and what works in the field. *Applied Clinical Trials*, 11, 46–56.
- Kearns K, Edwards R, & Tingstrom DH (1990). Accuracy of long momentary time-sampling intervals: Implications for classroom data collection. *Journal of Psychoeducational Assessment*, 8, 74–85. 10.1177/073428299000800109
- Kim B-R, & Teti DM (2014). Maternal emotional availability during infant bedtime: An ecological framework. *Journal of Family Psychology*, 28, 1–11. [PubMed: 24377731]
- Koffler RE, Ram N, Conroy DE, Pincus AL, & Almeida D (2016). Stressor diversity: Introduction and empirical integration into the daily stress model. *Psychology and Aging*, 31, 301–320. [PubMed: 27294713]
- Lu H, Frauendorfer D, Rabbi M, Mast MS, Chittaranjan GT, Campbell AT, ... & Choudhury T (2012). Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 351–360). ACM.
- Mark G, Gonzalez VM, & Harris J (2005, 4). No task left behind?: Examining the nature of fragmented work. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 321–330). Portland, OR: ACM.
- McLachlan GJ, & Peel D (2000). *Finite mixture models*. New York, NY: Wiley.
- Meany-Daboul MG, Roscoe EM, Bourret JC, & Ahearn WH (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis*, 40, 501–514. [PubMed: 17970263]
- Mehl MR (2012). Naturalistic observation sampling: The electronically activated recorder (EAR) In Mehl MR, & Conner TS (Eds.), *Handbook of research methods for studying daily life* (pp. 176–192). New York, NY: Guilford Press.
- Mehl MR, & Conner TS (Eds.). (2011). *Handbook of research methods for studying daily life*. New York, NY: Guilford Press.
- Molenaar PC (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2, 201–218.
- Needleman SB, & Wunsch CD (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48, 443–453. [PubMed: 5420325]
- Nesselroade JR, & Molenaar PC (2016). Some behavioral science measurement concerns and proposals. *Multivariate Behavioral Research*, 51, 396–412. [PubMed: 27248831]
- Newcomb ME, Swann G, Estabrook R, Corden M, Begale M, Ashbeck A, ... & Mustanski B (2016). Patterns and predictors of compliance in a prospective diary study of substance use and sexual behavior in a sample of young men who have sex with men. *Assessment*, Advanced online publication.
- Nyquist H (2002). Certain topics in telegraphic transmission theory. *Proceedings of the IEEE*, 86, 280–305. (Reprinted from *Transactions of the AIEE*, pp. 617–644, 1928)
- Powell J, Martindale A, & Kulp S (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, 8, 463–469. [PubMed: 16795509]

- Powell J, Martindale B, Kulp S, Martindale A, & Bauman R (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis*, 10, 325–332. [PubMed: 16795556]
- Radley KC, O’Handley RD, & Labrot ZC (2015). A comparison of momentary time sampling and partial-interval recording for assessment of effects of social skills training. *Psychology in the Schools*, 52, 363–378.
- Ram N, Conroy D, Pincus AL, Lorek A, Rebar AH, Roche MJ, & Gerstorf D (2014). Examining the interplay of processes across multiple time-scales: Illustration with the Intraindividual Study of Affect, Health, and Interpersonal Behavior (iSAHIB). *Research in Human Development*, 11, 142–160. doi: 10.1080/15427609.2014.906739 [PubMed: 26989350]
- Ram N & Gerstorf D (2009). Time structured and net intraindividual variability: Tools for examining the development of dynamic characteristics and processes. *Psychology and Aging*, 24, 778–791. [PubMed: 20025395]
- Ram N, & Reeves B (in press). Time sampling In Bornstein M (Ed.), *The SAGE encyclopedia of lifespan human development* (pp. xxx-xxx). New York, NY: Sage.
- Reis HT (2012). Why researchers should think “real-world”: A conceptual rationale In Mehl MR, & Conner TS (Eds.), *Handbook of research methods for studying daily life* (pp. 3–21). New York, NY: Guilford Press.
- Reynolds BM, Robles TF, & Repetti RL (2016). Measurement reactivity and fatigue effects in daily diary research with families. *Developmental Psychology*, 52, 442. [PubMed: 26689757]
- Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, & Mohr DC (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of Medical Internet research*, 17, e175. [PubMed: 26180009]
- Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, & Jones JH (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107, 22020–22025.
- Sankoff D, & Kruskal JB (Eds.). (1983). *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Reading, MA: Addison Wesley.
- Schmuckler MA (2001) What is ecological validity? A dimensional analysis. *Infancy*, 2, 419–436.
- Schwartz N (2007). Retrospective and concurrent self-reports: The rationale for real-time data capture In Stone A, Shiffman S, Atienza AA, & Nebeling L (Eds.), *The science of real-time data capture: Self-reports in health research* (pp. 11–26). New York, NY: Oxford University Press.
- Shadish WR, Cook TD, & Campbell DT (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin.
- Shiffman S, Stone AA, & Hufford MR (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1–32.
- Stone AA, & Shiffman SS (2011). Ecological validity for patient reported outcomes In Steptoe A, Freedland K, Jennings JR, Llabre MM, Manuck SB & Susman EJ (Eds.), *Handbook of behavioral medicine: Methods and applications* (pp. 99–112). New York, NY: Springer.
- Stone AA, Shiffman S, Schwartz JE, Broderick JE, & Hufford MR (2002). Patient non-compliance with paper diaries. *BMJ*, 324(7347), 1193–1194. [PubMed: 12016186]
- Studer M, & Ritschard G (2015). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society, Series A*.
- Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, ... & Campbell AT (2014). StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 3–14). ACM.
- Ward JH, Jr (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.
- Wegener DT, & Blankenship KL (2007). Ecological validity In Baumeister RF & Vohs KD (Eds.) *Encyclopedia of social psychology* (pp. 275–276). Thousand Oaks, CA: Sage.
- Wilhelm P, Perrez M, & Pawlik K (2012). Conducting research in daily life: A historical review In Mehl MR & Conner TS (Eds.), *Handbook of research methods for studying daily life* (pp. 62–86). New York, NY: Guilford Press.

Yeykelis L, Cummings JJ, & Reeves B (2014). Multitasking on a single device: Arousal and the frequency, anticipation, and prediction of switching between media content on a computer. *Journal of Communication*, 64, 167–192.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Synopsis

EMA-type designs are providing for discovery of much new knowledge about what, when, and how individuals act, think, and feel in their real lives. Tasked with questioning the ecological validity of such designs, we highlighted situations where ecological validity may be compromised during data collection (e.g., compliance with and reactivity to study procedures) and illustrated how assumptions (e.g., homogeneity) embedded in the analytical models typically used to derive inferences from experience sampling data undercut the potential for generalization across persons. We are excited to explore how new technologies and person-specific analytic methods may facilitate collection and description of very dense and long time-series that support valid assessment and discovery of individuals' unique human ecologies.

Figure 1.

Temporal sequences of daily stressor experience across 7-days. *Left panel:* Each row depicts individuals' naturally occurring experience of non-stressor (orange) and stressor (black) days (white = missing). Similarities and differences among never-stressor, always-stressor, and mixed stressor ecologies are highlighted. *Middle panel:* Quantitative differences with respect to *turbulence* (see Equation 4) are shown in the colored vertical bar – paired with the corresponding row in the left panel. *Right panel:* Sample-level distribution of turbulence scores, highlighting that the sample is indeed a mixture of individuals with non-variant (turbulence = 1.0) and varying stressor ecologies (turbulence > 1.0).

Figure 2.

Regression parameter estimates from 7-day binary design (Upper panel) and 21-day ordinal design (Lower panel). Estimates derived from person-specific models (blue points) connected via gray lines to the corresponding Bayesian empirical estimates derived from a multilevel model (black points), with the fixed effect parameters for the prototypical person (red point)

Figure 3.

Temporal sequences of daily stressor experience across 21-days. *Left panel:* Each row depicts number of stressors (darker color indicates more stressors) experienced on each day (white = missing), with sequences organized by similarity into 5 clusters separated by blue lines. *Right panel:* Sample-level distribution of turbulence scores, highlighting that the sample may be a mixture of individuals with different *patterns* of exposure to the experimental variable.