Social Psychological Bulletin
Psychologia Społeczna

Diary Studies in Social and Personality Psychology: An Introduction With Some Recommendations and Suggestions

John Nezlek[ab]

[a] *SWPS University of Social Sciences and Humanities, Poznań, Poland.* [b] *College of William and Mary, Williamsburg, VA, USA.*

Social Psychological Bulletin, 2020, Vol. 15(2), Article e2679, https://doi.org/10.32872/spb.2679

Received: 2020-01-13 • **Accepted:** 2020-03-24 • **Published (VoR):** 2020-07-31

Handling Editor: Ewa Gruszczyńska, SWPS University of Social Sciences and Humanities, Warsaw, Poland

Corresponding Author: John Nezlek, SWPS University of Social Sciences and Humanities in Poznań, Kutrzeby 10, 61-719 Poznań, Poland. E-mail: jnezlek@swps.edu.pl

Abstract

The present paper provides an overview of diary style research. This includes descriptions of different methods and the types of research questions for which they are appropriate. Data analytic methods are described and some recommendations are provided. Recommendations regarding the preparation of manuscripts describing the results of diary studies are also provided.

Keywords

diary studies, intensive repeated measures, multilevel modeling

Overview

This article does not concern a study per se; rather, it is intended as an overview of the methods that can be used to conduct what is often referred to as a “diary” or “diary style” study. Given the limitations of a journal article, not all topics are covered in detail. My goal in writing this article was to provide an introduction with the expectation that interested readers who are unfamiliar with diary methods will use this article as a starting point for their further education.

You may notice that this article does not contain a lot of references and citations. This was due to various factors. Many of the points I make could be supported by numerous citations and selecting one or even a small number of them would be essentially arbitrary and therefore relatively uninformative. On the other hand, some of the points I make reflect my experience conducting diary research since 1973 and no citations are available.

Nevertheless, researchers interested in learning more about conducting diary studies (or intensive repeated measures designs) might find the following useful:

1. Mehl and Conner (2012) – An edited volume that contains 36 chapters describing various aspects of diary style studies.
2. Nezlek (2012) – A description of how to conduct diary style studies of different types and how to analyze the data generated in such studies.
3. Bolger and Laurenceau (2013) – A description of how to conduct diary style studies and how to analyze the data generated in such studies, primarily focusing on interval or signal contingent methods, terms that are described below.

These are all books to which there may be limited access. I have written a series of articles about diary studies and how to analyze the data collected in diary studies. These are articles in journals and so they might be more accessible: Nezlek (2001, 2003, 2007, 2008), although they primarily concern data analyses.

What Is a Diary Study?

The term diary study has no clear definition. It is based on the word diary, which is usually defined as something such as a book in which one keeps a daily record of events and experiences. For present purposes, we will consider a diary study to be a study in which participants provide data on a regular basis over an extended period of time. Regular basis may refer to once or a few times each day, once or a few times each week, and so forth. Extended periods of time may consist of a few days, a few weeks, or longer. The frequency with which data are collected and the time over which data are collected need to be often and long enough to provide a sample of people’s lives that is sufficient to provide a basis for making inferences about the topic of the study. Other terms that have been used to describe such data collection protocols are “experience sampling” and “intensive repeated measures.”

By their nature, diary studies concern naturally occurring phenomena, “Life as it is lived” as described by Bolger, Davis, and Rafaeli (2003). Diary studies can examine phenomena that are difficult or impossible to study in the laboratory. For example, it would seem to be difficult to study the development of close relationships, particularly intimate relationships, in a fixed, controlled laboratory setting. Taking the phenomenon out of its real life context destroys it. Similarly, understanding how characteristics such as depression manifest themselves in daily life cannot be studied in the laboratory.

In some instances, it is the naturally occurring variability in the environment that is the focus of a study. For example, although one can study the mechanisms of ostracism and the processes that underlie reactions to it, it is only through studying naturally occurring ostracism that one can understand the importance of ostracism in people's lives. What if ostracism occurred so rarely that few people experienced it? By the way, this is not the case (Nezlek, Wesselmann, Wheeler, & Williams, 2012).

I am not proposing that controlled laboratory studies are worthless. Rather, I believe that laboratory studies have important limitations and that some of the limitations can be overcome (at least partially) by studies of naturally occurring phenomena. The two types of research can and should complement each other.

Different Types of Methods

Although various typologies to describe diary studies exist, I think Wheeler and Reis (1991) provide a straightforward summary of the different methods. They distinguished event contingent, interval contingent, and signal contingent methods. In an event contingent study, a record is generated every time a certain type of event occurs. A good example of this is a social interaction diary study (e.g., Wheeler & Nezlek, 1977). In such a study, every time a person has a social interaction he or she is asked to describe it. In an interval contingent study, a record is generated following the passage of a certain amount of time, e.g., at the end of each day, or at 8:00, 14:00, 20:00 each day. In a signal contingent study, a record is generated whenever a participant receives some type of signal such as a beep from a phone or watch (Csikszentmihalyi & Larson, 1987). For this reason, such studies are sometimes referred to as "beeper studies."

None of these methods is better than the others. Each has advantages. The selection of which method depends upon the hypotheses or questions of interest. What is the target of inference? What do you want to say?

Design Considerations: What, How Frequently, and for How Long?

The most important consideration when designing any study is the nature of the phenomenon in which you are interested. For diary style studies, in addition to theoretical questions, you need to consider the nature of the phenomenon. Unlike experimental studies in which circumstances are created, diary studies rely on naturally occurring phenomena. If the behavior or state that is the focus of your study does not occur during your study, you will not have anything to study.

Broadly speaking, are you interested in something that is easily recognizable and occurs on a regular basis for most people such as social interaction? Or are you interested in something that is more subtly defined and may not occur that often for many people such as transcendental spiritual experiences? The more frequent and common phenomena are, the better suited they are for studying using a diary of some kind. Infrequent phenomena that are not experienced by most/many people can be studied using a diary, but you will have a lot of "extra" data that are not relevant to your interests. In such cases a single occasion survey may be appropriate, or perhaps a less intensive data collection protocol such as once a month.

Assuming you are interested in something that can be defined relatively unambiguously and that occurs with some regularity leads to questions about your specific interest in this entity. Are you interested in how people experience or perceive something, how such perceptions covary with other state-level measures, and/or relationships between dispositional characteristics and perceptions and the covariation among state-level measures? You should design your study to answer such questions. The more clearly you formulate these questions the easier it will be to design your study. Certainly, you can start with vague questions such as: "Do people feel worse on days when they experience interpersonal stress than on days when they don't," but before you conduct your study you will need to define precisely what "feel worse" and "interpersonal stress" are.

Once you have defined the constructs of interest, it is simply a matter of selecting/designing measures of these constructs. The clearer and more precisely you have defined your constructs, the easier it will be to measure them. If you are having a lot of trouble figuring out how to measure something you may want to re-evaluate the definition of what you are measuring. Designing measures may not be automatic, but it should not be arduous.

It will probably be easier for you to decide how to measure dispositional characteristics such as traits than constructs at the within-person level. Measures of trait level constructs abound. Measuring constructs that you think will vary within-persons will probably be more challenging. In my work, I frequently use state-level analogs of trait-level measures. I tend not to use trait level measures as they have been designed for two reasons. Most important, many trait level measures tend to be too long to administer on a daily basis or multiple times each day. Also, the wording of trait level measures is typically not appropriate for diary administration (e.g., people are asked how they typically feel or think, or how they feel or think on average). Finally, most trait level measures have numerous items from which researchers can choose items for use in diary studies.

When creating state level analogs of trait level measures, I examine a trait-level measure and select a few items (typically three or four) that can be reworded for administration at the state-level. When available, I consult factor analyses and use items with higher loadings, assuming they are not highly redundant. Shorter measures may not measure a construct as broadly as a longer measure, but this does not mean that shorter measures are not valid. The validity of diary measures created in this way can be examined using relationships between trait level measures and means of diary measures (e.g., Nezlek & Plesko, 2001, p. 205).

When presenting items to participants I emphasize the timeframe to which the item applies. For example, in a study that uses end of day reports, items will start with a stem such as “Today I felt…”. In a beeper study, the stem might be “Now, I feel…”. Such stems help respondents focus their thoughts on the construct that you want to measure.

When designing measures for administration on a within-person basis do not feel bound by the trait-level measure on which you are basing your measures. For example, the Rosenberg Self-Esteem Scale (RSE), one of the most widely used scales in psychology, has 10 items. You probably do not need to use 10 items to measure self-esteem (e.g., Robins, Hendin, & Trzesniewski, 2001). Moreover, if participants are completing the measure every day (or a few times each day) and you ask all 10 items you will probably bore or annoy them. I should note that in a series of studies I have measured daily self-esteem successfully using four items. In this context, I define success in terms of internal consistency, stability across time, and relationships to scores of trait level measures including the trait RSE. See Nezlek (2017) for a discussion of estimating reliability within the context of multilevel analyses such as those used to analyze diary-style data.

I know that some of you are thinking: “But the Rosenberg Self-Esteem Scale has 10 items. If you are not asking all 10 items you are not measuring self-esteem!!” I understand such thinking, but such a belief confuses a construct with the measure of a construct. Although scores on the RSE are measures of self-esteem, scores on the RSE are not self-esteem.

Regardless of how you decide what items to include, you need to keep in mind the total number of items you are asking participants to answer. At some point (I wish that I could say exactly when), participants will become overloaded, and they will stop making the distinctions that you want them to make. I have seen beeper studies in which participants were asked to make 30 or more judgments such as emotions just a few hours apart. I know that participants provided answers (the data were presented); I am not certain that they distinguished the items as carefully as the researchers hoped or assumed they did.

Keep in mind that people have fixed or limited cognitive resources, and each answer requires some of these resources. If you ask participants to provide too many responses, they will not have enough resources to respond to each item thoughtfully. They may provide answers, but as the number of questions increases these answers are likely to increasingly reflect the influence of some dominant underlying dimension such as the hedonic dimension. For example, as discussed in Nezlek (2012, p. 31), in studies using social interaction diaries, as the number of ratings of individual interactions increased, the correlation between two basic ratings (satisfaction and intimacy) increased. My recommendation is to use fewer, better items. Less is more.

Sample Sizes

With good reason, researchers are expected to provide an estimate of the power of their designs. Unfortunately, it is not easy to estimate the power for the multilevel modeling analyses (MLM) that are standard for analyzing the data produced in diary style studies (described in the next section). As explained in Nezlek and Mroziński (2020), determining power for MLM analyses requires making assumptions about numerous parameters other than the size of the effect one wants to find. These parameters include error covariances, the reliabilities of coefficients, and the distributions of variances of both outcomes and predictors. This difficulty translates into difficulties in deciding how many participants are needed in a study and how many occasions of measurement are needed.

Although Nezlek and Mroziński do not provide specific advice regarding sample sizes, based on my experience (and the simulations conducted by Nezlek and Mroziński), I offer the following advice, which assumes a design in which days are nested within persons. Note that these estimates are for the sample sizes of the final sample that will be used for analysis. Researchers should anticipate that some percent (perhaps 10%) of participants will be excluded because they do not comply with the research protocol, and similarly, some percent of day-level observations will be excluded.

If hypotheses concern only relationships between a mean of a daily measure and person level measure (such as a trait), 50 participants and 7 days should be adequate. If hypotheses concern only within-person relationships between a single predictor and an outcome and do not concern cross-level interactions involving slopes, 100 participants and 10 days should be adequate. Finally, if analysts are interested in cross-level interactions (i.e., modeling individual differences in Level-1 slopes) researchers will need to include at least 125 participants who provide at least 14 days of data.

I stress that these recommendations are guidelines that can be useful for planning purposes. They are not formal recommendations. Formal power analyses require having a priori estimates of numerous parameters, which can be based on previous studies. If this is not possible, I recommend calculating achieved power after the fact. I should add that many well-intentioned editors and reviewers blindly request (or demand) a priori estimates of power without a full understanding of what is needed to provide such estimates.

There are also the issues of recruiting participants and maximizing compliance with a research protocol. For researchers who rely on students who participate in research as part of a course requirement, recruiting is not particularly difficult. The researcher primarily needs to be certain that the compensation (e.g., credits) is equitable. Recruiting participants from the general public can be done online or through social networks. Regardless, I recommend that researchers be as transparent as possible regarding the study. For example, my colleagues and I have told potential participants that a study is about daily experience (or social interaction), and that many participants in the past have found the study to be enjoyable and informative. This last point is in fact true. Finally, I emphasize the fact that the study will take only a few minutes (less than 10) each day, a limit to which I have adhered.

Researchers need to design study protocols so that compliance with a protocol does not interfere with participants' lives too much; otherwise, participation will destroy what a study is examining. The most important consideration is how much time answering the questions will take. Pre-testing (not using research assistants or colleagues) can help determine this. Remember, although participants may be interested in a study, they are probably not as interested as you are, and they are most certainly not interested in the subtleties in which you are interested. Interference can also be considered in terms of the number (and likely/possible circumstances) of assessment occasions each day.

Compliance with a protocol can be understood in two ways. Do participants provide answers or do they answer the questions? For example, although participants who are asked to provide 40 ratings multiple times each day may provide these ratings, I would not be confident that they provided all of these answers thoughtfully. As mentioned previously, as the number of total items increased in my social interaction diary studies, the correlation between two measures increased. I cannot offer some all-purpose recommendation other than to be sensitive to the possibility that lengthy protocols will undermine the validity of all responses.

Data Analysis

Multilevel modeling (MLM) is the current standard for best practice to analyze diary style data. Although a full discussion of using MLM is beyond the scope of this article, I present a brief description here. Interested readers can consult Nezlek (2001, 2003, 2007) for brief articles about how to use MLM to analyze diary style data, and more detailed accounts can be found in Nezlek (2011, 2012) and in Bolger and Laurenceau (2013). The primary rationale for using MLM is that it takes into account the fact that in a hierarchically nested data structure units of observation are not independent. The independence of observations (technically, the independence of error) is a foundational assumption of OLS regression, which means that OLS regression cannot be used to analyze a multilevel data set.

Logic of MLM

When describing MLM I rely on Bryk and Raudenbush's (1992) now classic explanatory framework in which separate models are described for each level of analysis. For a diary study in which daily diaries are nested within persons this would be represented with separate models for the within- and between-person variables of interest. A basic model is below.

Within-person	$y_{ij} = \beta_{0j} + r_{ij}$.
Between-person	$\beta_{0j} = \gamma_{00} + u_{0j}$.

In this model, there are *i* days nested within *j* persons. The mean for each of *j* persons is β_0, and the overall mean is γ_{00}. The variance of r_{ij} is the Level-1 (or within-person) variance, and the variance of u_{0j} is the Level-2 (or between-person) variance. Note that the intercept from Level 1 (β_{0j}) becomes an outcome at Level 2. In essence, a set of Level-1 coefficients (in this case only a mean) is estimated for each Level-2 unit (person in this example), and then these coefficients are analyzed at Level 2. It is important to note that in reality, all these parameters are being estimated simultaneously. Bryk and Raudenbush's contribution was to represent the models at different levels of analysis separately to make it clearer what is being done.

Conceptualizing the Model

Assuming that you are using some type of MLM, the first consideration is the number of levels in the analyses. What is nested within what? Typically, decisions about the number of levels in a design are straightforward and are usually dictated or suggested by data. In a diary study, this is usually occasions (days, interactions, etc.) nested within persons.

It is useful to keep in mind that each level of analysis represents a sample, and if there are not enough units of observation at a level of analysis, it will not be possible to model the variance at the level of analysis. For example, assume students from three schools are measured once a day for two weeks. Conceptually, this could be considered as a three-level model (days within students and students within schools). Nevertheless, three schools do not provide enough information to estimate the random variability associated with sampling schools. Three schools do not constitute a sample of schools. In such a case, school can become a person level variable in a two-level model.

The same considerations apply to within-person sampling. For example, assume participants maintain a daily diary twice, perhaps before and after an intervention of some kind. One might be tempted to nest days within times (pre vs. post) and times within persons. Although appealing in some ways, such a model would not be the best approach. Two time periods do not constitute a sample of time periods. In such a case, time of assessment (pre vs. post in this example) would be represented at the day level as a fixed effect.

Often, researchers conduct studies in which multiple observations are collected each day for each person. The classic example of this is the “beeper study” described previously. Typically, the data from such studies are treated as two-level models in which occasions of measurement are nested within persons. The fact that observations are nested within days is ignored. I do not think this is good practice, and if you collect such data, I encourage you to consider analyzing them with three-level models (observations nested within days and days nested within persons). You may have trouble estimating some of the random effects, but your estimates will have taken into account the possibility that what appear to be occasion-level relationships are in fact day-level relationships.

Broadly speaking, I urge you to ignore advice about using ICCs (intra-class correlations) as a method of determining whether nesting is appropriate or not. An ICC describes the distribution of variances, and some argue that if there is no variance at a specific level of analysis that level of analysis can be ignored. Although apparently sensible, this criterion ignores the possibility that relationships between two variables may vary across units of analysis when means do not.

For example, in the data set contained in Table 1, the ICCs for both variables are 0. All groups have the same mean for both variables. Based on the ICCs, the nested structure of these data should be ignored. If the 30 observations are treated as one group, the correlation between x and y is 0. Clearly this is not an accurate description of the relationships in the sample. The x-y relationship in Groups 1, 2, and 3 is perfectly negative, whereas it is perfectly positive in Groups 4, 5, and 6.

Table 1

Varying Relationships Within-Groups When ICCs Are 0

Group 1		**Group 2**		**Group 3**	
x	**y**	**x**	**y**	**x**	**y**
1	5	1	5	1	5
2	4	2	4	2	4
3	3	3	3	3	3
4	2	4	2	4	2
5	1	5	1	5	1
Group 4		**Group 5**		**Group 6**	
x	**y**	**x**	**y**	**x**	**y**
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5	5	5	5

Centering

Within the multilevel context, when predictors are selected, decisions about how they will be centered need to be made. Centering refers to the reference value from which the deviations of a predictor are taken. In OLS regression, this is usually the mean for a variable. The situation is somewhat more complicated in MLM analyses. I discuss this issue in terms of two-level models, e.g., days nested within persons. More details about centering can be found in the references previously provided and in Enders and Tofighi (2007).

At Level 2, there are two options: grand-mean centering and uncentered (also referred to as zero-centered). When a Level-2 predictor is entered grand-mean centered, deviations are taken from the grand mean, and the intercept represents the expected value for an observation that has a value that is at the grand mean of the predictor. When a Level-2 predictor is entered uncentered, deviations are taken from 0, and the intercept represents the expected value for an observation that has a value of 0 on the predictor.

At Level 1, there are three options: grand-mean centering, uncentered, and group-mean centering. Similar to Level 2, when a Level-1 predictor is entered grand-mean centered, deviations are taken from the grand mean, and the intercept represents the expected value for an observation that has a value that is at the grand mean of the predictor. Also similar to Level 2, when a Level-1 predictor is entered uncentered, deviations are taken from 0, and the intercept represents the expected value for an observation that has a value of 0 on the predictor. The third option that is available for Level-1 predictors is group-mean centering. When a Level-1 predictor is entered group-mean centered, deviations are taken from the group mean, and the intercept represents the expected value for an observation that has a value that is at the group mean of the predictor.

Note that “group” in this instance refers to the Level-2 unit of analysis. In a diary study in which days are nested within persons, groups would be persons. When a predictor is group-mean centered the intercept represents the mean for the outcome measure. Group-mean centering is the multilevel equivalent of conducting a regression analysis for each group (person), and then analyzing the coefficients generated by these analyses as outcomes in person-level analyses.

For Level-2 predictors, I recommend grand-mean centering continuous measures and zero-centering categorical measures. This makes interpreting the coefficients and generating predicted values (see below) relatively straightforward. For Level-1 predictors, I recommend group-mean centering continuous measures and zero-centering categorical measures. Note that grand-mean centering Level-1 predictors introduces Level-2 variance into the Level-1 model because the reference point for a grand-mean centered predictor is the grand mean, which represents the mean of all observations.

Modeling Error

Modeling error is another aspect of MLM that can be puzzling for analysts whose primary experience is with OLS regression. In contrast to OLS regression in which there is only one error term, in MLM, each Level-1 predictor can have its own error term. Moreover, the covariances among these individual error terms are also estimated. Collectively, the individual error terms and their covariances are referred to as the “error structure.” When thinking about modeling error it is important to keep in mind that how error is modeled can affect estimates of fixed effects, which are typically the focus of hypotheses. So, even though error structures are rarely the focus of hypotheses, you need to ensure that error is modeled properly to ensure that tests of your fixed effects are accurate.

In MLM, Level-1 coefficients can be modeled in one of three ways. (1) Randomly varying – a fixed effect and a random effect (error term) are estimated. (2) Non-varying – a fixed effect is estimated but no random effect is estimated. (3) Non-randomly varying – a fixed effect is estimated, no random effect is estimated, but there is a Level-2 predictor of the Level-1 coefficient. It is important to note that the absence of a random effect (#2) does not mean that an effect does not vary. It means that the random effect cannot be estimated reliably, or in other words, random and true variance cannot be distinguished.

In most software packages, individual error terms for each coefficient are tested for significance. I recommend dropping error terms that are not significant – I see no reason to estimate parameters that cannot be estimated reliably. Noting this, I recommend using a more generous significance level than .05 (at least .10, perhaps .15) when making decisions about including random error terms. This second recommendation reflects the fact that conceptually, most coefficients are random (they represent a coefficient from a sample of coefficients), and this randomness should be modeled if at all possible. If you are concerned about the influence of fixing an effect (i.e., not estimating a random error), run the model with and without the random error term to see how the results differ.

Generally speaking, I recommend not spending too much time describing error structures when writing up the results of analyses. As I said previously, hypotheses rarely concern error structures per se. There are some differences in the strength of inference between randomly varying coefficients and coefficients that do not vary randomly (e.g., the calculation of confidence intervals is more problematic if there is no random error term), but for most intents and purposes, this difference is not important. Finally, how error terms are specified varies across software packages, and analysts need to be careful that their models are estimating the error structures they want to estimate.

Sequence of Analyses

Regardless of the specific structure, I recommend starting your analyses with “totally unconditional” models of each of the measures. Totally unconditional refers to the fact that there are no predictors at any level of level analysis. Such a model was presented at the beginning of the section “Logic of MLM.” Such models provide the basic descriptive statistics of a MLM: the mean and the variance estimates at each level of analysis. Although such models typically do not test hypotheses per se, they do provide valuable information about the distribution of variances, information that can be used to guide and evaluate further analyses.

Assuming a two-level model, I recommend constructing the Level-1 model and then examining differences in the Level-1 coefficients at Level 2. For example, if days are nested within persons, the Level-1 model would describe day-level (or within-person) relationships, and the Level-2 model would describe relationships between person-level measures such as personality traits and the phenomena represented by the Level-1 coefficients. When constructing models, it is important to recognize that hypotheses regarding coefficients are tested against a null of 0.

For example, assume you want to model daily self-esteem as a function of daily stress. The model would look like this:

Within-person	$y_{ij} = \beta_{0j} + \beta_{1j} * \text{Stress} + r_{ij}.$
Between-person	$\beta_{0j} = \gamma_{00} + u_{0j}.$
	$\beta_{1j} = \gamma_{10} + u_{1j}.$

Note that the relationship (slope) between self-esteem (y) and stress (β_{1j}) now becomes an outcome at Level 2. The null hypothesis is that the mean slope (γ_{10}) between self-esteem and stress is 0. This model can be extended to include a predictor of the slope, creating what is sometimes called a “slopes as outcomes” or cross-level interaction.

In the following model, the slope between daily self-esteem and stress is modeled as a function of extraversion. The model is presented below, and whether the esteem-stress slope varies as a function of extraversion is tested by the significance of the γ_{11} coefficient.

Within-person	$y_{ij} = \beta_{0j} + \beta_{1j} * \text{Stress} + r_{ij}.$
Between-person	$\beta_{0j} = \gamma_{00} + \gamma_{01} * \text{Extraversion} + u_{0j}.$
	$\beta_{1j} = \gamma_{10} + \gamma_{11} * \text{Extraversion} + u_{1j}.$

Constructing a Level-1 model consists of deciding which predictors you want to include, and this includes *how these predictors are centered and how error is modeled.* Although the rule is not absolute, common practice is to (initially) enter the same predictors at Level 2 for all Level-1 coefficients (intercept and slopes). This was done in the example above. The reason for this is that if a Level-2 predictor is not entered, the assumption is that it is not significant, and such assumptions need to be tested. Moreover, all coefficients are estimated based on covariance matrices, and if an element of the matrix is missing, this changes the estimates of the other elements.

Regardless, particularly at Level 1, be conservative in terms of adding predictors. I recommend using what are called “forward-stepping” procedures. When using forward-stepping procedures, predictors are added to models one at a time and checked for significance. Of course, groups of predictors can also be added in sequence. This is the opposite of what are called “backward-stepping” stepping procedures that start with the most complex models and delete terms. Backward-stepping procedures are commonly used in single level regression analyses.

The reason for this recommendation is based on the number of parameters that are estimated in a MLM analysis. MLM estimates more parameters than are estimated in a comparable OLS regression, and the number of parameters increases non-linearly as a function of the number of predictors. For example, the simple model, $Y_{ij} = \beta_{0j} + r_{ij}$, estimates three parameters, the mean, and the two variance estimates. Adding a predictor, $Y_{ij} = \beta_{0j} + \beta_{1j} (x) + r_{ij}$, estimates six parameters, fixed and random effects for intercept and the predictor, the correlation between the two random terms, and the Level-1 variance. Adding a second predictor, $Y_{ij} = \beta_{0j} + \beta_{1j} (x_1) + \beta_{2j} (x_2) + r_{ij}$, estimates 10 parameters, fixed and random effects for the intercept and the two predictors, the correlations between the three random terms, and the Level-1 variance. As you can see, adding a predictor requires adding more than one parameter, and as the total number of predictors increases, the number of parameters each predictor requires also increases. In contrast, in OLS regression, adding a predictor requires the estimation of only one more parameter.

It is important to consider the number of parameters a model estimates because you need to be careful about trying to estimate more parameters than your data can estimate. The more data you have, the more parameters you can estimate. Nonetheless, to my knowledge, there are no quick and easy guidelines for this. As discussed, I recommend starting simply and keeping models “lean and tight.” Opt for fewer parameters that are estimated well rather than many parameters that are estimated poorly or not so well.

Once the Level-1 model is finalized (including the error structure), then predictors can be added at Level 2. Although adding predictors at Level 2 does not involve the same type of increases in the number of parameters estimated by adding predictors at Level 1, I still encourage analysts to use forward-stepping algorithms. I believe that the possible risk of inflating Type-I error posed by forward stepping is outweighed by the improvement in the quality of the parameters that are estimated.

Software Options

Particularly for beginning or less experienced analysts I strongly recommend using the program HLM (Raudenbush, Bryk, Cheong, & Congdon, 2019). HLM was designed to conduct MLM, and the interface allows users to select modeling options such as centering and error terms easily and unambiguously. A free trial version of the program, good for 14 days, is available at: https://ssicentral.com/index.php/product/hlm

I know that some of you, perhaps many, are thinking: “But what about R? If I don’t use R people will think I am naïve.” Let us put aside social pressure and discuss reality. R is a very powerful platform, and for experienced users it can provide numerous, sophisticated options that are not available using other platforms. Unfortunately, it also provides numerous opportunities for analysts to misspecify models, and R modules do not have many (if any) safeguards built in. I will not belabor this point. MLM can be conducted using numerous software packages (SPSS, SAS, Mplus, Stata, R, and others), and if the *same* model is specified, the results will be the same. You can use whatever program you want to conduct MLM; just be certain that the program is running the model you want to run. If you use HLM this is easy to determine.

Preparing Manuscripts

Preparing manuscripts describing the results of diary-based studies requires a bit more attention to some details than preparing a manuscript describing an experimental study or a single level regression analysis. These differences are not dramatic, but they can be important for readers (and reviewers). A thorough description of what was done and what was found that is not cluttered with extraneous detail will increase readers' appreciation of the contribution of a piece of research. Much of what I describe here is standard good practice, but I thought some of it was worth repeating. For a discussion of some of these issues within the context of social and personality psychology, see Nezlek (2001). Readers interested in EMA studies that concern health psychology/behavioral medicine can consult Stone and Shiffman (2002), although be advised that they do not discuss MLM.

Method

1. A brief description of participants, including how they were recruited and if they were compensated. The specific description will vary as a function of the study at hand, but at the least some basic demographic variables should be provided, age, sex, employment, education, and so forth. There is no reason to spend too much time on this. A few brief sentences should be enough.
2. A description of what data were deleted from the analyses and why. Diary studies present different demands than single occasion studies, and dropout or exclusion (bad data) rates are typically higher in diary studies than in single occasion studies. I have deleted as many as 10% of participants from the primary analyses of a diary study. Nonetheless, deleting participants is not a crime or a sin. In contrast, deleting participants and not disclosing the fact that they were deleted or not providing a clear rationale for their deletion is not ethical. Similarly, individual observations (e.g., days in a daily diary study) can also be deleted for various reasons, typically because an entry has not been made at an appropriate time (e.g., the middle of the afternoon for a study that asks for end of day reports). Researchers need to establish criteria for inclusion/exclusion at both the within- and between-person levels *before they conduct their analyses* to avoid the temptation to select the data that "work." In my daily diary research, I have used different criteria to define what a valid entry is (e.g., collegians have a different daily rhythm than non-students), and for a two-week study, I typically cut at seven days of valid data. The critical issue is that you ensure

that to the extent possible the data you retain for analysis are not systematically biased. One specific recommendation is that what are called “next day” reports (e.g., describing on Wednesday morning what happened on Tuesday) generally seem to be reasonably unbiased (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004). Finally, keep it simple. It makes no sense to spend two paragraphs describing the specific reasons why 10 different participants were excluded from the analyses.

3. A clear description of the data that were collected. This includes the specific questions that were asked and the response scales that were used. You need to provide enough detail so that someone can repeat your study. Sometimes this can be done with a figure, and sometimes these details can be included in online supplemental materials. Regardless of how, they need to be available.

Results

1. Start the results with a description of the basic analytic framework. For example, for a daily diary study, days could be treated as nested within persons. For a social interaction study, social interactions might be nested within persons. The first model should be a totally unconditional model, i.e., no predictors at a level of analysis. Such basic models can then be used to generate the basic descriptive statistics of a MLM: the mean, and the variance estimates at each level of analysis. Note that in MLM there is more than one variance estimate. I urge you to be cautious when reporting correlations at the within-person level. First, keep in mind that correlations between means that are aggregated across occasions of measurement do not describe within-person relationships. Second, correlations based on samples in which all the Level-1 observations in a study (e.g., diary entries) are treated as a single sample are not accurate because they confound between- and within-person variances. It is possible to estimate within-person correlations using some specialized software packages (e.g., Mplus), but describing how to do this is beyond the scope of this article.
2. Describe how predictors were entered and how error was modeled. Following the presentation of the descriptive statistics, the sequence of the models that are presented will vary as a function of the focus of the study. Regardless, it is *essential* to describe centering and error terms. Centering changes what slopes and intercepts mean, and error terms are part of model specification (and misspecification). Finally, unless there are compelling reasons to do otherwise (e.g., non-significant error terms), all Level-1 coefficients should be modeled as randomly varying.
3. Avoid number clutter. When describing the results of analyses, it usually suffices to report the coefficients (what the hypotheses invariably concern) and the *t* and *p* values accompanying the coefficients. There is no need to report standard errors because *t* = coefficient/*SE*. Once you have two of these, the third is determined. There is also no need to report model fits and comparisons of model fits. The focus in MLM

(vs. *SEM*) is not on how well a model reproduces a covariance matrix; rather, the focus is on significance tests of coefficients. Some researchers test the significance of a coefficient by comparing the fit of a model with that coefficient to a model without it. Although well-intended, such comparisons include the error structures of the two models (individual error terms and their covariances) not just the fixed effects (the coefficients). Hypotheses typically do not concern error structures and so tests of hypotheses should not include error structures.

4. Explain results with predicted values. Often, the results of MLM analyses can be difficult to understand based upon coefficients alone. For example, the meaning of a cross-level interaction (slopes as outcomes analysis) may not be readily apparent from the coefficients themselves. Moreover, MLM estimates unstandardized coefficients (despite well-intended but sometimes misguided suggestions about how to standardize them), and so the meaning of a coefficient needs to be evaluated within the context of the variances of the measures that are in a model. I have found it useful to estimate predicted values +/- 1 *SD* from the mean, a standard for regression analyses. For example, one could write “On days that were high in positive affect, self-esteem was higher than it was on days when positive affect was lower (5.5 vs. 4.8).” NB: The within-person *SD* of a variable needs to be estimated from a totally unconditional model of that variable. See Nezlek (2011, 2012) for descriptions of how to do this.

Conclusions, Thoughts About the Future, and so Forth

The use of diary style methods has increased markedly over the past few decades. What was once innovative is now commonplace. Nevertheless, rapid growth in any domain is often accompanied by “growing pains.” For example, my sense is that many researchers do not have a sufficient understanding of the multilevel analyses that are commonly used to analyze diary-style data. Moreover, as analytic options become more complex (e.g., multilevel mediation, multilevel *SEM*, greater use of Bayesian estimates), this problem will become more, not less, serious. For future generations, such problems can be addressed by formal instruction in postgraduate programs. Individuals who have finished their formal education can attend workshops, read books, and so forth.

In this regard, it is important to distinguish the ability to conduct an analysis from knowing if an analysis is correct and what the results mean. Consulting established experts, such as Raudenbush and Bryk, Kreft and deLeeuw, Hox, and Goldstein (to name a few), can help researchers understand the rationale for MLM. Without understanding why, researchers run the risk of running models that are technically accurate but are not accurate in terms of answering the questions a study was designed to answer.

I also see challenges (and opportunities) in terms of deciding what types of methods are appropriate for what types of questions/topics. “One size does not fit all.” Some topics such as mood variation might be better studied using a frequent measurement strategy (e.g., multiple times a day), whereas other topics such as self-evaluation might be better studied on a daily basis. There is no right and wrong for such decisions. Researchers need to define carefully the constructs in which they are interested and develop measures of these constructs.

I also encourage researchers to consider event-contingent methods, i.e., examining characteristics of and reactions to specific types of events/occurrences. Contemporary research has relied heavily on various types of interval-contingent methods (e.g., end of day, random measures during a day), but for events/occurrences that are easily recalled and frequent enough to provide a basis for inference, event-contingent protocols may be appropriate. Moreover, interval- and event-contingent protocols can be combined in the same study. Regardless of the specific protocol, researchers need to be careful to avoid asking for so many responses that the validity of individual responses is compromised. Participants may provide responses, but this does not necessarily mean they have provided answers.

Diary methods are tools that can be used to examine a wide variety of questions. The only limit is researchers’ imagination. I hope that this article has provided some information that can help researchers transform their ideas into reality.

Funding: Preparation of this paper was supported by grant 2018/31/B/HS6/02822 awarded to John Nezlek from the Polish National Science Centre (Narodowe Centrum Nauki).

Competing Interests: The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments: The author has no support to report.

References

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*, 579-616. https://doi.org/10.1146/annurev.psych.54.101601.145030

Bolger, N., & Laurenceau, J.-P. (2013). Intensive longitudinal methods: An introduction to diary and experience sampling research. In T. D. Little (Series Ed.), *Methodology in the Social Sciences.* New York, NY, USA: Guilford Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Advanced Qualitative Techniques in the Social Sciences: Vol. 1. Hierarchical linear models: Applications and data analysis methods.* New York, NY, USA: SAGE.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease, 175*, 526-536. https://doi.org/10.1097/00005053-198709000-00004

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121-138. https://doi.org/10.1037/1082-989X.12.2.121

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The Day Reconstruction Method. *Science, 306*, 1776-1780. https://doi.org/10.1126/science.1103572

Mehl, M. R., & Conner, T. S. (Eds.). (2012). *Handbook of research methods for studying daily life.* New York, NY, USA: Guilford Press.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event and interval contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin, 27*, 771-785. https://doi.org/10.1177/0146167201277001

Nezlek, J. B. (2003). Using multilevel random coefficient modeling to analyze social interaction diary data. *Journal of Social and Personal Relationships, 20*, 437-469. https://doi.org/10.1177/02654075030204002

Nezlek, J. B. (2007). A multilevel framework for understanding relationships among traits, states, situations, and behaviors. *European Journal of Personality, 21*, 789-810. https://doi.org/10.1002/per.640

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2*, 842-860. https://doi.org/10.1111/j.1751-9004.2007.00059.x

Nezlek, J. B. (2011). Multilevel modeling for social and personality psychology. In J. B. Nezlek (Series Ed.), *The SAGE Library in Social and Personality Psychology Methods.* London, United Kingdom: SAGE.

Nezlek, J. B. (2012). Diary methods for social and personality psychology. In J. B. Nezlek (Series Ed.), *The SAGE library in social and personality psychology methods.* London, United Kingdom: SAGE.

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality, 69*, 149-155. https://doi.org/10.1016/j.jrp.2016.06.020

Nezlek, J. B., & Mroziński, B. (2020). Applications of multilevel modeling in psychological science: Intensive repeated measures designs [Applications du modèle multiniveau dans les sciences psychologiques: Les plans à mesures répétées intensives]. [in English]. *L'Année Psychologique, 120*, 39-72. https://doi.org/10.3917/anpsy1.201.0039

Nezlek, J. B., & Plesko, R. M. (2001). Day-to-day relationships among self-concept clarity, self-esteem, daily events, and mood. *Personality and Social Psychology Bulletin, 27*, 201-211. https://doi.org/10.1177/0146167201272006

Nezlek, J. B., Wesselmann, E. D., Wheeler, L., & Williams, K. D. (2012). Ostracism in everyday life. *Group Dynamics, 16*, 91-104. https://doi.org/10.1037/a0028029

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2019). *HLM 8 for Windows* [Computer software]. Skokie, IL, USA: Scientific Software International, Inc.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27*, 151-161. https://doi.org/10.1177/0146167201272002

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine, 24*, 236-243. https://doi.org/10.1207/S15324796ABM2403_09

Wheeler, L., & Nezlek, J. (1977). Sex differences in social participation. *Journal of Personality and Social Psychology, 35*, 742-754. https://doi.org/10.1037/0022-3514.35.10.742

Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality, 59*, 339-354. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x