

## Planned Missing-data Designs in Experience-sampling Research: Monte Carlo Simulations of Efficient Designs for Assessing Within-person Constructs

By: [Paul J. Silvia](#), [Thomas R. Kwapil](#), Molly A. Walsh, Inez Myin-Germeys

Silvia, P.J., Kwapil, T.R., Walsh, M.A., & Myin-Germeys, I. (2014). Planned missing data designs in experience sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods*, 46(1), 41-54. doi: 10.3758/s13428-013-0353-y

The final publication is available at Springer via <http://dx.doi.org/10.3758/s13428-013-0353-y>

\*\*\*© Springer. Reprinted with permission. No further reproduction is authorized without written permission from Springer. This version of the document is not the version of record. Figures and/or pictures may be missing from this format of the document. \*\*\*

### Abstract:

Experience-sampling research involves trade-offs between the number of questions asked per signal, the number of signals per day, and the number of days. By combining planned missing-data designs and multilevel latent variable modeling, we show how to reduce the items per signal without reducing the number of items. After illustrating different designs using real data, we present two Monte Carlo studies that explored the performance of planned missing-data designs across different within-person and between-person sample sizes and across different patterns of response rates. The missing-data designs yielded unbiased parameter estimates but slightly higher standard errors. With realistic sample sizes, even designs with extensive missingness performed well, so these methods are promising additions to an experience-sampler's toolbox.

**Keywords:** Missing data | Experience sampling methods | Efficient designs | Maximum likelihood | Ecological momentary assessment

### Article:

The idea behind experience sampling is simple: to understand what is happening in people's everyday lives, "beep" them with a device, and ask them to answer questions about their current experiences (Bolger & Laurenceau, 2013; Conner, Tennen, Fleeson & Barrett, 2009).

Experience-sampling methods have been increasingly used in social, clinical, and health psychology to examine topics as wide-ranging as substance use (Buckner, Zvolensky, Smits, Norton, Crosby, Wonderlich & Schmidt, 2011), romantic relationships (Graham, 2008), psychosis (Oorschot, Kwapil, Delespaul & Myin-Germeys, 2009), self-injury (Armey, Crowther & Miller, 2011), and personality dynamics (Fleeson & Gallagher, 2009). But as in many things, the devil is in the details. For fixed-interval and random-interval designs, in which people are beeped at fixed or random times throughout the day, researchers must decide how many items

should be asked at each beep, how many beeps should be given each day, and how many days the study should last. These three design parameters—number of items per beep, number of beeps daily, and number of total days—tug against each other, and thus require compromises to avoid participant burden, fatigue, and burnout. For example, asking many items per beep usually requires using fewer beeps per day; conversely, using many beeps per day usually requires asking fewer questions per beep (Hektner, Schmidt & Csikszentmihalyi, 2007). If the demands on participants are too high, both the quality and quantity of responses will suffer.

In this article, we describe some useful planned missing-data designs (Enders, 2010; Graham, Taylor, Olchowski & Cumsille, 2006). Although such designs have been popular in cross-sectional and longitudinal work, researchers have not yet considered their potential for experience-sampling studies. Planned missing-data designs allow researchers to omit some items at each beep. This reduces the time and burden needed to complete each assessment, and it enables researchers to ask more questions in total without expanding the number of questions asked per beep. We suspect that such designs could thus be useful for experience-sampling research because space in the daily-life questionnaire is tight. Using real data from a study of hypomania and positive affect in daily life (Kwapil, Barrantes-Vidal, Armistead, Hope, Brown, Silvia & Myin-Germeys, 2011), first we illustrate some common planned missing-data designs and describe their typical performance. Then we present two Monte Carlo simulations that evaluate the performance of planned missing data across a range of within-person (Level 1) and between-person (Level 2) sample sizes and patterns of response rates.

### **The basics of planning missing data**

During the dark ages of missing-data analysis, missing observations were usually handled with some form of deletion or simple imputation (Graham, 2009; Little & Rubin, 2002). With the advent of modern methods—particularly multiple imputation and maximum likelihood—researchers have many sophisticated options for analyzing data sets with extensive missingness, if certain assumptions are made (Enders, 2010). The rise of methods for analyzing missing data has brought new attention to planned missing-data designs (Graham et al., 2006), which have a long history but were relatively impractical until the development of modern analytic methods. Sometimes called *efficiency designs*, planned missing-data designs entail deliberately omitting some items for some participants via random assignment. In cross-sectional research, for example, the items might be carved into three sets, and each participant would be randomly assigned two of the sets. This method would reduce the number of items that each participant answered (from 100 % to 67 %) without reducing the overall number of items asked. In longitudinal research, for another example, researchers could randomly assign participants to a subset of waves instead of collecting data from everyone at each wave. Such methods save time and money, and they can “accelerate” a longitudinal project, such as in cohort sequential designs (Duncan, Duncan & Stryker, 2006).

Planned missing-data designs allow researchers to trade statistical power for time, items, and resources. Unlike unintended types of missing data, planned missing data are missing completely at random (MCAR): The observed scores are a simple random sample of the set of complete scores, so the likelihood of missingness is unrelated to the variable itself or to other variables in the data set (Little & Rubin, 2002). As a result, maximum likelihood methods for missing data can be applied that assume that the data are MCAR or missing at random (MAR)—in the latter case, that the likelihood of missingness is unrelated to the variable itself but is related to other variables in the data set (Little & Rubin, 2002). The body of simulation research on maximum likelihood methods, viewed broadly, has shown that standard errors are somewhat higher—consistent with the higher uncertainty that stems from basing an estimate on fewer data—but the coefficients themselves are unbiased estimates of the population values (see Davey & Savla, 2010; Enders, 2010; McKnight, McKnight, Sidani & Figueredo, 2007). The higher standard errors are the price that researchers pay for the ability to ask more items or to reduce the time needed to administer the items. This can be a good investment, if the savings in time and money enable asking more items, sampling more people, or yielding better data, such as by reducing fatigue or increasing compliance.

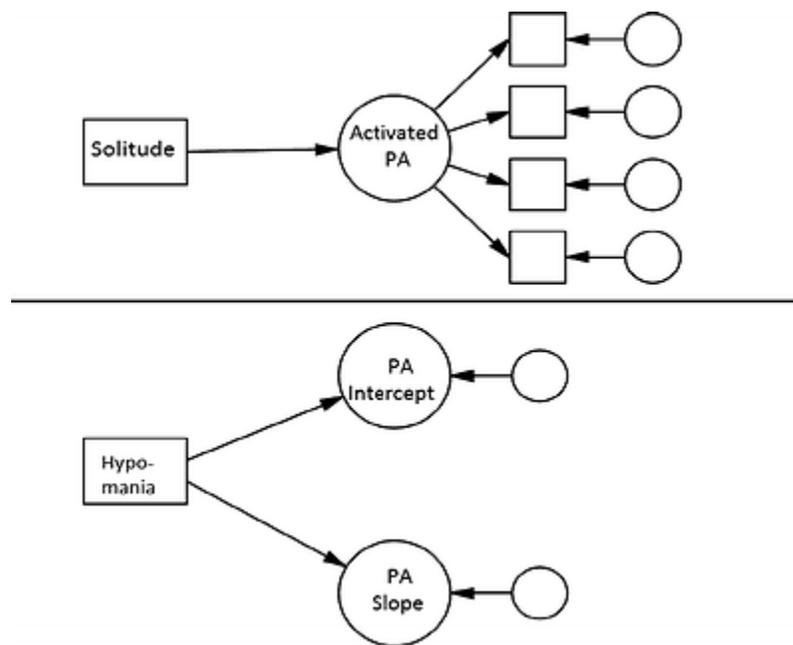
### **An empirical demonstration of planned missing-data designs**

To illustrate different missing-data designs in action, we used data collected as part of a study of hypomania's role in daily emotion, cognition, and social behavior (Kwapil et al., 2011). In this study, 321 people took part in a 7-day experience-sampling study. Using personal digital assistants (PDAs), we beeped the participants eight times per day for 7 days. One beep occurred randomly within each 90-min block between noon and midnight. Compliance was good: People completed an average of 41 daily-life questionnaires ( $SD = 10$ ). In the original study, each person received all items at each beep.

The key to implementing planned missing-data designs for beep-level, within-person constructs is to model the Level 1 constructs as latent variables. Multilevel models with latent variables are combinations of familiar structural equation models and hierarchical linear models (Mehta & Neale, 2005; Muthén & Asparouhov, 2011; Skrondal & Rabe-Hesketh, 2004); Heck and Thomas (2009) have provided a thorough didactic introduction for researchers interested in learning to run multilevel structural equation models. Instead of averaging across items—especially for cases in which researchers would average across three or more items—researchers can specify the items as indicators of a Level 1 latent outcome. Apart from enabling the use of planned missing designs, modeling latent Level 1 variables has the other advantages of latent variable modeling, such as the ability to model measurement error.

Figure 1 depicts the multilevel model used for our examples. The outcome variable is *activated positive affect* (PA), which is a latent variable defined by the observed items *enthusiastic*, *excited*, *energetic*, and *happy*. The Level 1 predictor *solitude* is a binary variable that reflects whether people are alone or with other people. Much research has shown

that PA is higher when people are with others (Burgin, Brown, Royal, Silvia, Barrantes-Vidal & Kwapil, 2012; Silvia & Kwapil, 2011; Watson, 2000). Solitude was group-mean centered (i.e., within-person centered) using the observed solitude scores, and the slopes and intercepts were modeled as being random. At Level 2, we estimated a main effect of hypomania, measured with the Hypomanic Personality Scale (Eckblad & Chapman, 1986), on PA (shown as the path from Hypomania to PA Intercept), as well as a cross-level interaction between solitude and hypomania (shown as a path from Hypomania to PA Slope). The Hypomanic Personality Scale measures a continuum of trait-like variability in bipolar spectrum psychopathology (see Kwapil et al., 2011). Hypomania was grand-mean centered using the observed scores. As is typical for experience-sampling research, the data were fully complete at Level 2 but partially complete at Level 1 because not everyone responded to every beep (Silvia, Kwapil, Eddington & Brown, 2013). The models were estimated using Mplus 6.12 using maximum likelihood estimation with robust standard errors.



**Fig. 1** A multilevel model with a latent outcome

***Matrix designs for three within-person indicators***

A typical experience-sampling study of mood would have many affect items to capture a range of affective states. If a construct includes three items, researchers can use a family of matrix designs. In our study, we included many Level 1 items that assessed PA. For this example, we will consider three of them: *enthusiastic*, *excited*, and *energetic*.

A simple matrix design for three indicators would randomly administer two of the three items at each beep. Table 1 depicts the design. In our example, people would be asked *enthusiastic* and *excited* a third of the time, *enthusiastic* and *energetic* a third of the time,

and *excited* and *energetic* a third of the time—no one would receive all three items at any beep. Table 2 depicts the covariance coverage for this sampling design. The diagonal of the table displays the proportion of complete data for an item. Each item is 67 % complete because it is asked two thirds of the time. The off-diagonal elements represent the proportions of complete data for a covariance between two items. All of the off-diagonal cells have a value of 33 % because each item is paired with each other item a third of the time. This last point is a critical feature of the matrix design, and of planned missing designs generally: Each item must be paired with every other item. If an off-diagonal cell has a value of zero—that is, two items never co-occur—the model will fail unless advanced methods beyond the scope of this article are applied (see, e.g., Willse, Goodman, Allen & Klaric, 2008).

**Table 1** Depiction of a matrix sampling design for three indicators of activated positive affect

	<b>Enthusiastic</b>	<b>Excited</b>	<b>Energetic</b>
Item Group 1	X	X	
Item Group 2		X	X
Item Group 3	X		X

X indicates that an item was administered

**Table 2** Covariance coverage matrix for the three-indicator matrix design

	<b>Enthusiastic</b>	<b>Excited</b>	<b>Energetic</b>
Enthusiastic	.67		
Excited	.33	.67	
Energetic	.33	.33	.67

The diagonal elements indicate the proportions of complete data for an item; the off-diagonal elements indicate the proportions of complete data for a pair of items

What are the consequences of using such a design? We estimated the multilevel model described earlier: a latent *activated PA* variable was formed by the three items, and it served as the Level 1 outcome predicted by solitude (Level 1), hypomania (Level 2), and their interaction. The predictor variables in these models had complete scores—in these and other planned missing-data models, the missingness is varied only for the outcomes. We estimated two models. The first used the complete-cases data file, in which everyone was asked all three items at all occasions. The second used a matrix sampling design, in which we randomly deleted observed Level 1 cases on the basis of the matrix sampling design in Table 1, which yielded a data set with the covariance coverage displayed in Table 2. The intraclass correlations (ICCs) for the

individual items—the proportions of the total variance at Level 2—ranged from .243 to .267. To estimate an ICC for the latent variable, researchers must ensure that it is invariant across the two levels, thereby placing the Level 1 and Level 2 variances on the same scale (Mehta & Neale, 2005). Equating the latent variable across Levels 1 and 2 is a good idea, regardless of missing data, and it is accomplished by constraining the factor loadings to be equal across levels (see Heck & Thomas, 2009, chap. 5; Mehta & Neale, 2005, p. 273). The ICC for the latent activated PA variable was .353.

Table 3 displays the patterns of effects. Comparing the parameters and standard errors reveals very little difference: Not much is lost when only two of three items are asked. Consistent with formal simulation research, the primary difference was in the standard errors. The estimates of regression paths and variances were at most modestly affected, but the standard errors were somewhat higher for elements of the model that hinged on the three PA items. Not surprisingly, the standard errors were most inflated for the factor loadings for the three items that served as indicators of the latent activated PA variable.

**Table 3** Parameter estimates for the complete-cases model and the matrix sampling model for three indicators

	<b>Complete</b>	<b>Matrix</b>
L1 solitude effect	.263 (.032)	.263 (.033)
L2 hypomania effect	.039 (.006)	.038 (.006)
Cross-level interaction	-.003 (.004)	-.002 (.004)
L1 CFA: Enthusiastic loading	1 (0)	1 (0)
L1 CFA: Excited loading	1.077 (0.019)	1.073 (0.032)
L1 CFA: Energetic loading	0.961 (0.020)	0.991 (0.031)
L1 latent PA residual variance	1.326 (0.057)	1.299 (0.067)
L2 random slope residual variance	0.127 (0.035)	0.117 (0.034)
L2 random intercept residual variance	0.627 (0.062)	0.618 (0.064)

The loading for *enthusiastic* was fixed to 1 in each model. The numbers in parentheses are standard errors

### ***Matrix designs for four indicators***

All else being equal, standard errors will increase as missingness becomes more extensive. To illustrate this, we can consider a model with four within-person indicators. Table 4 shows a

balanced matrix design for four indicators. It adds another item, *happy*, to the within-person activated PA latent variable. In this design, each beep asks two of the four items, and all pairs of items are asked. The covariance coverage is thus much lower than in the three-item matrix design. As Table 5 shows, 50 % of the data are missing for each item, and each covariance is estimated on the basis of about 16 % of the beeps.

**Table 4** Depiction of a matrix sampling design for four indicators of activated positive affect

	<b>Happy</b>	<b>Enthusiastic</b>	<b>Excited</b>	<b>Energetic</b>
Item Group 1	X	X		
Item Group 2	X		X	
Item Group 3	X			X
Item Group 4		X	X	
Item Group 5		X		X
Item Group 6			X	X

X indicates that an item was administered

**Table 5** Covariance coverage matrix for the four-indicator matrix design

	<b>Happy</b>	<b>Enthusiastic</b>	<b>Excited</b>	<b>Energetic</b>
Happy	.50			
Enthusiastic	.16	.50		
Excited	.16	.16	.50	
Energetic	.16	.16	.16	.50

The diagonal elements indicate the proportions of complete data for an item; the off-diagonal elements indicate the proportions of complete data for a pair of items

As before, we estimated two multilevel models: one using the complete-cases data file, and another in which Level 1 were randomly deleted data to create a matrix sampling design for the four items. The ICCs for the indicators ranged from .227 to .267, and the ICC for the latent activated PA variable was .345. Table 6 displays the effects and their standard errors. As before, and as expected, the factor loadings and regression weights are essentially the same, but the standard errors are larger, particularly for elements of the model involving the individual items.

**Table 6** Parameter estimates for the complete-cases model, the matrix sampling model, and the anchor test model

	<b>Complete</b>	<b>Matrix</b>	<b>Anchor Test</b>
L1 solitude effect	.221 (.025)	.213 (.025)	.229 (.025)
L2 hypomania effect	.029 (.004)	.028 (.004)	.029 (.004)
Cross-level interaction	-.002 (.003)	-.002 (.003)	-.002 (.003)
L1 CFA: Happy loading	1 (0)	1 (0)	1 (0)
L1 CFA: Enthusiastic loading	1.285 (0.031)	1.359 (0.052)	1.281 (0.036)
L1 CFA: Excited loading	1.393 (0.029)	1.472 (0.052)	1.397 (0.031)
L1 CFA: Energetic loading	1.216 (0.029)	1.297 (0.049)	1.215 (0.032)
L1 latent PA residual variance	0.808 (0.039)	0.734 (0.048)	0.806 (0.040)
L2 random slope residual variance	0.074 (0.020)	0.060 (0.018)	0.071 (0.020)
L2 random intercept residual variance	0.373 (0.036)	0.341 (0.037)	0.373 (0.036)

The loading for *happy* was fixed to 1 in each model. The numbers in parentheses are standard errors

### ***Anchor test designs for within-person indicators***

In the matrix sampling designs that we have discussed, no single item is always asked. Sometimes, however, a researcher may always want to ask a particular item. An item might be a “gold standard” item, an item that reviewers expect to see, or a particularly reliable item. For such occasions, researchers can use *anchor test designs*. For example, in an experience-sampling study of positive affect, one might want to always ask *happy*, given its centrality to the construct, but to ask the other items only occasionally.

An example of an anchor test design is shown in Table 7, where people are asked three of the four items at each beep. The item *happy* is always asked, and two of the three remaining items are occasionally asked. (This particular design is akin to Graham et al.’s, 2006, three-form design with an X-set asked of everyone.) Table 8 shows the covariance coverage for the four items. Because *happy* is always asked, it occurs 67 % of the time with the other three items. The occasional items, in contrast, are missing 33 % of the time and occur 33 % of the time with each other. This matrix illustrates the sense in which the item *happy* “anchors” the set of items: Data for the item most central to the latent construct—both the item itself and its relations with other items—are less sparse than data for the items more peripheral to the latent construct. Like the

matrix design, the anchor test design requires each item to co-occur with every other item—no off-diagonal cell is zero—but the degree of co-occurrence varies between pairs.

**Table 7** Depiction of an anchor test design for four indicators of activated positive affect

	<b>Happy</b>	<b>Enthusiastic</b>	<b>Excited</b>	<b>Energetic</b>
Item Group 1	X	X	X	
Item Group 2	X		X	X
Item Group 3	X	X		X

X indicates that an item was administered

**Table 8** Covariance coverage matrix for the anchor test design

	<b>Happy</b>	<b>Enthusiastic</b>	<b>Excited</b>	<b>Energetic</b>
Happy	1			
Enthusiastic	.67	.67		
Excited	.67	.33	.67	
Energetic	.67	.33	.33	.67

The diagonal elements indicate the proportions of complete data for an item; the off-diagonal elements indicate the proportions of complete data for a pair of items

Table 6 depicts the results in the final column and enables a comparison with both the complete-cases design and the matrix sampling design. In its coefficients and standard errors, the anchor test design is much closer to the complete-cases design than is the matrix sampling design. For the coefficients, neither design is especially discrepant, but the estimates for the anchor test design are particularly close to the complete-cases design. For the standard errors, both the anchor test and matrix sampling designs have higher standard errors for the parts of the model involving the four indicators, but the standard errors for the anchor test design are appreciably less inflated than those for the matrix sampling design. The better performance for the anchor test design is intuitive: Both designs have four indicators, but the anchor test design has less overall missingness (as is shown in their covariance coverage matrices).

### **A Monte Carlo study of level 1 and Level 2 sample sizes**

The empirical demonstrations that we have presented have their strengths: They capture the coarse character of real experience-sampling data, and they reflect the kinds of effects that researchers can expect when using these methods in the field. Nevertheless, formal Monte Carlo

simulations, although more artificial, can illuminate the behavior of the planned missing-data designs across a range of factors and conditions (Mooney, 1997). One subtle but important strength of Monte Carlo designs for missing-data problems is the ability to evaluate effects relative to the true population parameters. In our empirical simulations, the population values were unknown, so the planned missing designs were evaluated against the complete-cases sample. Missing-data methods, however, seek to reproduce the population values, not what the observed data would look like if there were no missing values (Enders, 2010). Monte Carlo simulations specify the population values, so we can gain a more incisive look at the performance of designs with different patterns of missing within-person data.

Our first simulation examined three designs: the complete cases, an anchor test design, and a matrix design, all for four indicators. We picked these because the anchor test design is relatively cautious (one item is always asked), and the matrix design is relatively risky (the covariance coverage is only 16 %). To obtain population values for the Monte Carlo simulations, we used the observed values from the complete-cases analysis of our hypomania study used in the empirical simulations (Kwapil et al., 2011). Using values from actual data increases the realism, and hence the practicality, of Monte Carlo simulations: The values of the parameters and ICCs reflect values that researchers are likely to encounter (Paxton, Curran, Bollen, Kirby & Chen, 2001). The model that we estimated was the same as in the empirical simulation: Hypomania (Level 2), solitude (Level 1), and their cross-level interaction predicted activated PA, a latent outcome with four indicators.

Our simulation focused on the roles of the Level 1 and Level 2 sample sizes in the performance of the various designs. These elements are important for two reasons. First, one would expect sample size to play a large role: With smaller sample sizes, especially at Level 1 (the number of beeps), one should encounter more convergence failures and higher standard errors. Second, Level 1 and Level 2 sample sizes are under a researcher's control, whereas many other aspects of a model (e.g., the ICC) are not. For our simulation we used a 3 (type of design: complete cases, anchor test, matrix)  $\times$  4 (Level 2 sample size: 50, 100, 150, 200)  $\times$  4 (Level 1 sample size: 15, 20, 45, 60) design. Simulation research commonly includes values that are unrealistic to illuminate a model's performance under dismal situations. In our analysis, a Level 2 sample size of 50 would strike most researchers as relatively small and a poor candidate for planned missing-data designs; the remaining levels (100, 150, and 200) are more realistic and consistent with typical samples in experience-sampling work. Likewise, a Level 1 sample size of 15 is obviously too small, 30 is somewhat low for a week-long experience-sampling project, and 45 and 60 represent realistic values for a typical 7- to 14-day sampling frame.

To simplify the reporting of the results, we focused on three parameters: the Level 1 main effect of solitude on PA, the Level 2 main effect of hypomania on PA, and the factor loading for one of the Level 1 PA confirmatory factor analysis (CFA) items (the item *excited*, chosen randomly). We sought three kinds of outcomes. First, how many times did the model fail to terminate normally? Second, how biased—that is, how discrepant from the population values—were the

estimates of the regression weights and factor loadings? And third, how did the missing-data designs affect standard errors across the Level 1 and Level 2 sample sizes? The simulated data were generated and analyzed using Mplus 6.12 using maximum likelihood. Each condition had 1,000 simulated samples.

**Estimation problems**

How often did the multilevel model fail to converge? The numbers of times that the model failed to terminate normally are shown in Table 9 and Fig. 2. Not surprisingly, termination failures were most common at the lowest sample sizes (50 people responding to 15 beeps), particularly when a missing-data design was used. Even in the worst case, however, the model failed to terminate normally only 2.10 % of the time. As Fig. 2 shows, termination failures were essentially zero for all designs once reasonable sample sizes (at least 100 people and 30 beeps) were achieved.

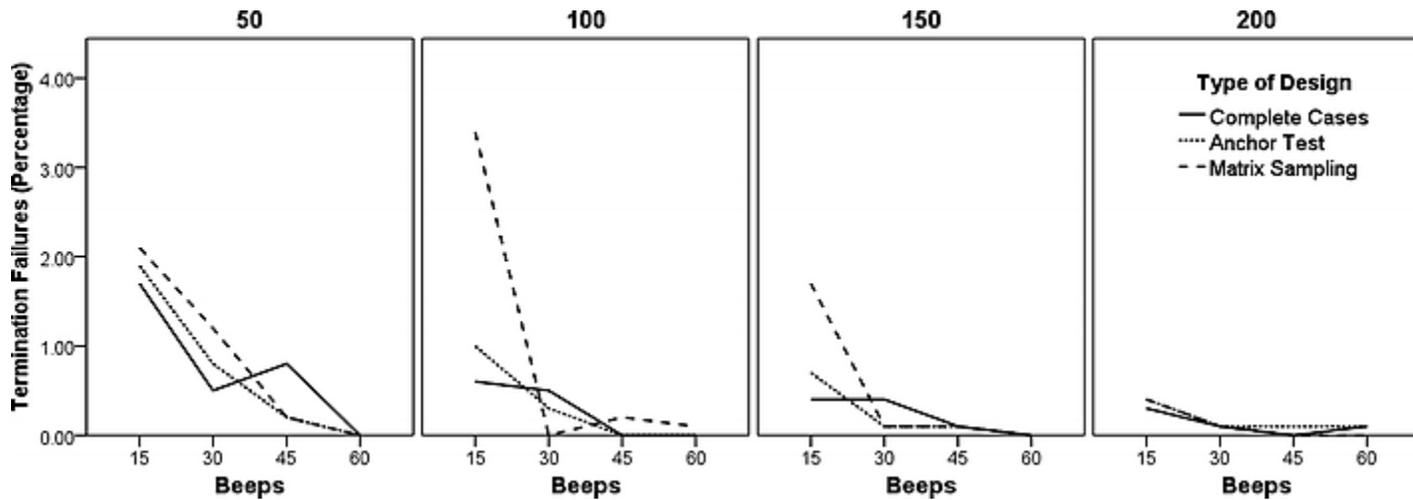
**Table 9** Effects of the planned missing-data design and the Level 1 and Level 2 sample sizes on termination failures, parameter biases, and standard errors

		Complete Cases				Anchor Test				Matrix			
		50	100	150	200	50	100	150	200	50	100	150	200
Termination failures	15	1.70	0.60	0.40	0.30	1.90	1.00	0.70	0.40	2.10	3.40	1.70	0.40
	30	0.50	0.50	0.40	0.10	0.80	0.30	0.10	0.10	1.20	0.00	0.10	0.10
	45	0.80	0.00	0.10	0.00	0.20	0.00	0.10	0.10	0.20	0.20	0.10	0.00
	60	0.00	0.00	0.30	0.10	0.00	0.00	0.00	0.10	0.00	0.10	0.00	0.00
L1 factor loading bias	15	0.187	-0.065	0.050	0.043	0.739	0.043	0.000	-0.057	1.392	0.057	0.136	0.251
	30	0.151	0.179	0.000	0.014	0.194	0.151	0.122	-0.043	0.660	0.215	0.029	0.143
	45	0.258	0.122	0.187	-0.036	0.158	-0.057	0.093	-0.029	0.344	0.201	0.244	0.086
	60	0.079	0.043	-0.100	-0.029	0.022	0.007	-0.093	0.057	-0.294	0.036	0.029	-0.108
L1 main effect	15	1.830	-0.357	-0.759	-0.089	0.714	0.446	-0.491	0.893	-3.214	-0.893	0.179	0.179

bias	30	-0.938	-0.491	0.804	-1.071	-2.321	-0.089	0.089	0.848	-0.268	0.580	0.937	-0.179
	45	-0.268	0.491	0.625	0.759	-0.714	0.937	-0.089	-0.714	-1.071	-0.268	0.268	0.402
	60	1.429	0.759	0.446	0.446	0.893	-0.268	0.268	-0.357	0.491	-0.223	-0.268	-0.491
L2 main effect bias	15	1.480	2.224	0.519	0.187	0.458	0.372	0.701	0.803	3.431	0.501	-0.143	-0.330
	30	0.128	-0.629	-0.341	0.770	1.757	0.055	0.084	0.559	-1.108	1.416	1.018	0.458
	45	0.801	0.337	0.528	0.165	0.105	-0.325	0.440	1.177	2.536	0.105	-0.075	0.842
	60	1.713	0.963	-0.226	0.319	-0.576	0.526	0.033	0.004	3.589	1.164	0.374	-0.141
L1 factor loading $SE$	15	0.088	0.062	0.050	0.044	0.106	0.074	0.060	0.052	0.170	0.117	0.095	0.083
	30	0.064	0.045	0.036	0.031	0.077	0.054	0.044	0.038	0.123	0.087	0.070	0.061
	45	0.052	0.037	0.030	0.026	0.063	0.045	0.036	0.032	0.103	0.072	0.059	0.051
	60	0.045	0.032	0.026	0.023	0.055	0.039	0.032	0.028	0.090	0.064	0.052	0.045
L1 main effect $SE$	15	0.116	0.080	0.065	0.056	0.120	0.084	0.068	0.059	0.128	0.089	0.072	0.062
	30	0.081	0.057	0.046	0.040	0.084	0.059	0.048	0.041	0.089	0.063	0.051	0.044
	45	0.066	0.047	0.038	0.033	0.069	0.048	0.039	0.034	0.074	0.051	0.042	0.036
	60	0.058	0.041	0.033	0.029	0.060	0.042	0.034	0.030	0.064	0.045	0.036	0.031
L2 main effect $SE$	15	2.147	1.479	1.194	1.034	2.147	1.495	1.215	1.046	2.260	1.543	1.242	1.082
	30	2.011	1.397	1.137	0.986	2.000	1.406	1.142	0.990	2.058	1.445	1.169	1.015

	4	1.969	1.370	1.118	0.963	1.967	1.382	1.123	0.971	2.013	1.401	1.136	0.988
	5												
	6	1.915	1.361	1.101	0.957	1.942	1.362	1.104	0.963	1.958	1.375	1.119	0.972
	0												

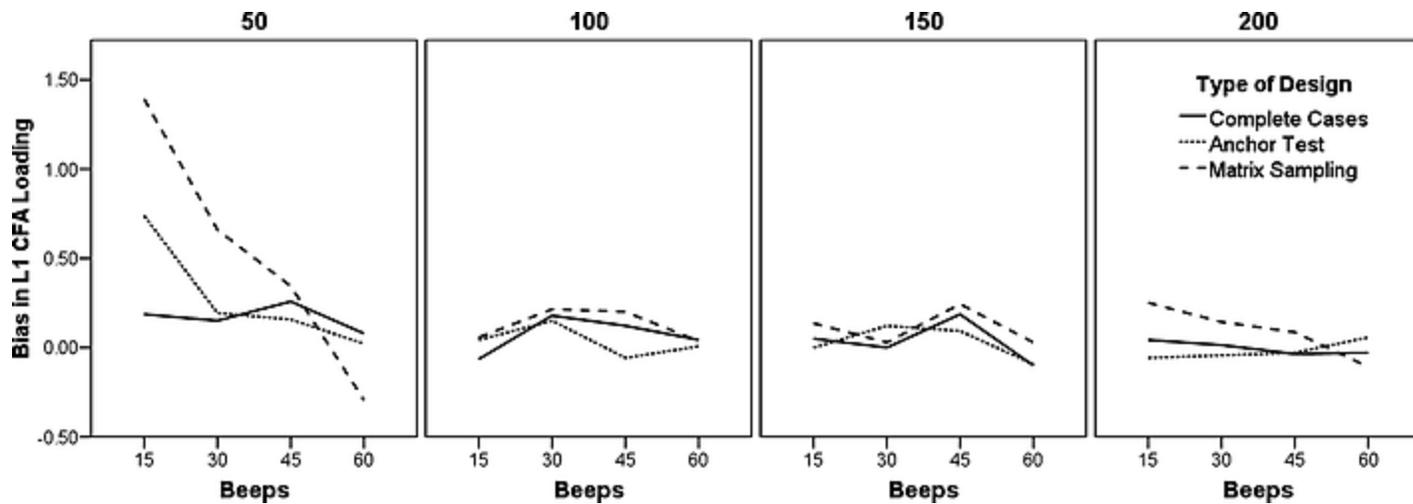
“Termination failures” are the percentages of the 1,000 samples on which the model did not terminate normally. “Bias” is the percentage difference between the simulated value and the population value. *SE* = standard errors. Factor loadings are the loadings of individual items on the Level 1 latent positive affect variable. L2 main effects are the effects of hypomania, the Level 2 predictor, on the latent positive affect variable



**Fig. 2** Effects of planned missing-data designs and sample sizes on termination failures

***Parameter bias***

How closely did the factor loadings and regression weights resemble the population values? Bias scores—differences between the simulated estimates and population estimates, expressed as percentages—are shown in Table 9. As a visual example, Fig. 3 illustrates the bias scores for the Level 1 CFA factor loading. Consistent with the broader missing-data literature, the level of bias was small across all three planned missing-data designs. Most of the bias scores (87 %) were less than  $\pm 1$  %, and the most biased estimates were nevertheless only  $-3.21$  % and  $3.59$  % different from the population values. The findings thus support the prior simulations, albeit in a different context.

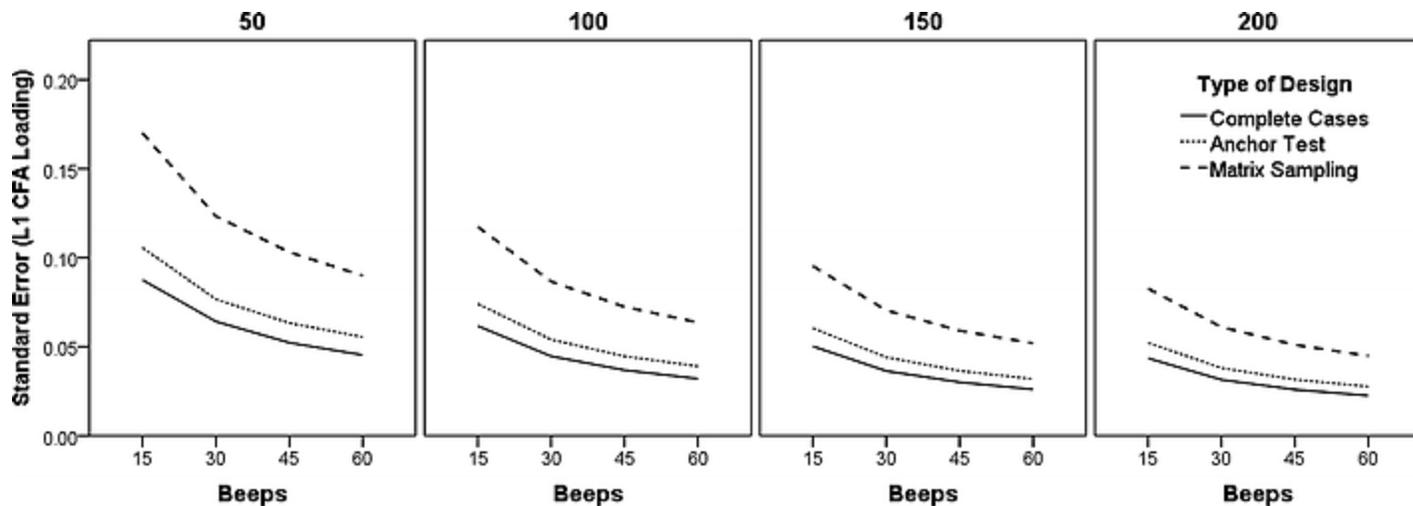


**Fig. 3** Effects of planned missing-data designs and sample sizes on Level 1 confirmatory factor analysis loading parameter biases

### *Standard errors*

How did the planned missing-data designs influence standard errors across different sample sizes? Table 9 displays the standard errors. First, it is apparent that the standard errors decline as both Level 1 and Level 2 sample sizes increase, consistent with the literature on power analysis in multilevel models (Bolger, Stadler & Laurenceau, 2012; Maas & Hox, 2005; Scherbaum & Ferrerter, 2009). More relevant for our purposes is how the standard errors differed across the different planned missing-data designs. Overall, standard errors were lowest for the complete-cases design, somewhat higher for the anchor test design, and notably higher for the matrix design; as one would expect, the standard errors increased as missingness increased.

Furthermore, the influence of missing data was restricted to the within-person aspects of the model: Missing data increased standard errors for the Level 1 factor loading and for the Level 1 main effect, but not for the Level 2 main effect. To illustrate this pattern, Fig. 4 displays the standard errors for the Level 1 CFA factor loading.



**Fig. 4** Effects of planned missing-data designs and sample sizes on Level 1 confirmatory factor analysis loading standard errors

### *Discussion*

The Monte Carlo simulations revealed several things about the behavior of planned missing-data designs. For model termination, all designs converged well, once acceptable sample sizes were reached (at least 100 people and 30 beeps), so the use of planned missing-data designs for within-person constructs doesn't create notable convergence problems, which were determined primarily by sample size. For parameter bias, the estimates of the factor loadings and regression weights were minimally biased. All of the designs accurately recovered the parameter values, even under sparse conditions (e.g., a matrix design with 50 people and 15 beeps). Although it is perhaps counterintuitive, this finding follows from the underlying statistical theory (Little & Rubin, 2002; Rubin, 1976) and replicates the broader simulation literature, which supports the effectiveness of maximum likelihood methods when data are MCAR, as in planned missing designs. Finally, for standard errors, the planned missing designs influenced the sizes of standard errors as one would expect: Standard errors increased as missingness increased. The anchor test design yielded somewhat higher standard errors than the complete-cases design; the matrix design, which has covariance coverage of only 16 %, yielded much higher standard errors. But in all cases, standard errors declined as both Level 1 and Level 2 sample sizes increased. The increase in standard errors due to planned missing data can thus be offset by larger samples at both levels.

### **A Monte Carlo simulation of variable response rates**

For our second Monte Carlo study, we wanted to explore the effects of variable response rates on the behavior of planned missing-data designs. As experience-sampling researchers know all too well, participants vary in how many surveys they complete, an irksome form of unplanned missingness. Some participants respond to every single beep, most participants respond to most

beeps, and some participants respond to few beeps. Part of conducting experience-sampling research is developing procedures to boost response rates, such as incentives (e.g., everyone who completes at least 70 % is entered in a raffle), establishing a rapport with participants, sending reminder texts or e-mails, or having participants return to the lab once or twice during the study (Barrett & Barrett, 2001; Burgin, Silvia, Eddington & Kwapil, 2013; Conner & Lehman, 2012; Hektner et al., 2007).

In our prior simulation, response rates were fixed: Each “participant” in a sample had the same number of Level 1 observations, which is akin to a study in which everyone responded to the same number of beeps. In a real study, however, researchers would observe a range of response rates, including low response rates for which planned missing designs might fare poorly. The present simulation thus varied response rates to examine their influence, if any, on the performance of different planned missing designs.

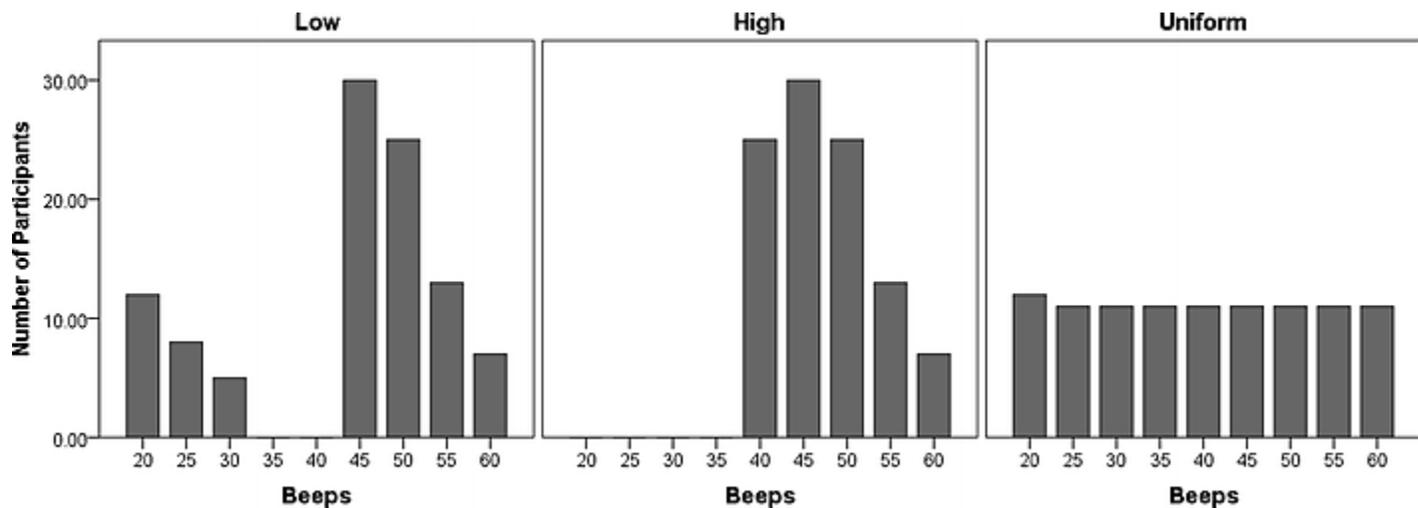
### ***Empirical context***

As with the prior Monte Carlo simulations, we used results from real experience-sampling data as population values. For these simulations, we shifted to a new context and sample data set. The data are from a study of social anhedonia and social behavior in daily life (Kwapil, Silvia, Myin-Germeys, Anderson, Coates, & Brown, 2009). Social anhedonia, viewed as a continuum, represents trait-like variation in people’s ability to gain pleasure from social interaction (Kwapil, Barrantes-Vidal & Silvia, 2008). People high in social anhedonia have a diminished need to belong, spend more time alone, and are at substantially higher risk for a range of psychopathologies (Brown, Silvia, Myin-Germeys & Kwapil, 2007; Kwapil, 1998; Silvia & Kwapil, 2011). In the original study, we measured social anhedonia as a Level 2 variable prior to the experience-sampling study. People were then asked to respond to questions about mood and social engagement several times a day across 7 days. For the present analyses, we focused on four items that people completed when they were with other people at the time of the beep: *I like this person (these people)*, *My time with this person (these people) is important to me*, *We are interacting together*, and *I feel close to this person (these people)*. The four questions, not surprisingly, covary highly, and thus load highly on a latent *social engagement* variable. The ICCs for the items ranged from .155 to .254; the ICC for the latent variable was .246. These items are good examples of items that could productively be used for a planned missing-data design: One would want to measure social engagement at each beep, but the four items are interchangeable.

For our simulations, we chose a simple Level 2 main effect model: *social anhedonia*, an observed Level 2 variable, had a main effect on the latent social engagement variable. No Level 1 predictors or cross-level interactions were present. As one would expect, as social anhedonia increased, people reported lower social engagement with other people,  $b = -.389$ ,  $SE = .064$ ,  $p < .001$ , which would be considered a large effect size ( $R^2 = 29\%$ ).

### Monte Carlo design

The Monte Carlo simulations explored the effects of four kinds of response rate variability. We created a hypothetical experience-sampling study in which 100 people were beeped 60 times over the course of a week. Three planned missing-data designs were examined, as in the prior simulations: a complete-cases design, an anchor test design, and a matrix sampling design. We then created four kinds of response rate patterns. In a *fixed* condition, response rates were fixed at 45 out of 60, to represent a case in which everyone completed 75 % of the beeps and to serve as a no-variability benchmark. In a *uniform* condition, response rates varied from 20 out of 60 (33 %) to 60 out of 60 (100 %) in increments of 5. (Twenty beeps served as a floor; experience-sampling research commonly drops participants whose response rates fall under a predefined threshold.) In a *high-response-rate* condition, 70 % of the sample completed at least 75 % of the beeps, with some variability around 75 %, and in a *low-response-rate* condition, 70 % of the sample again completed at least 75 % of the beeps, but 30 % of the sample piled up at the floor. These conditions are easiest to understand when they are visualized: Figure 5 illustrates the response rates for the low, high, and uniform conditions. For the simulation we thus used a 3 (type of design: complete cases, anchor test, matrix)  $\times$  4 (response rate: fixed, uniform, low, high) design. As before, we used Mplus 6.12 and 1,000 simulated samples in each condition.



**Fig. 5** Depiction of the low, high, and uniform response rates used in the simulations

We focused on two effects for which poor performance would be obvious: The factor loading for one of the Level 1 items (the item *I feel close to this person (these people)*, chosen randomly), and the regression weight for the main effect of social anhedonia on the latent social engagement variable. As before, we evaluated parameter bias (how much did the simulated factor loading and regression weight diverge from the population values?) and the standard errors (how much did the standard errors change as a function of response rates and missing-data designs?). We did not evaluate estimation failures because, given the simpler model and solid sample sizes at both levels, all samples terminated normally in all conditions.

### *Parameter bias*

For parameter bias, we again found minimal bias in the estimated Level 1 factor loading and Level 2 main effect. Table 10 displays the bias values, as percentages, for all conditions. Bias values were tiny: The most biased score was only  $-1.2\%$  different from the population value. These results indicate that parameter bias is essentially unaffected by the type of planned missing-data design and the types of response rates.

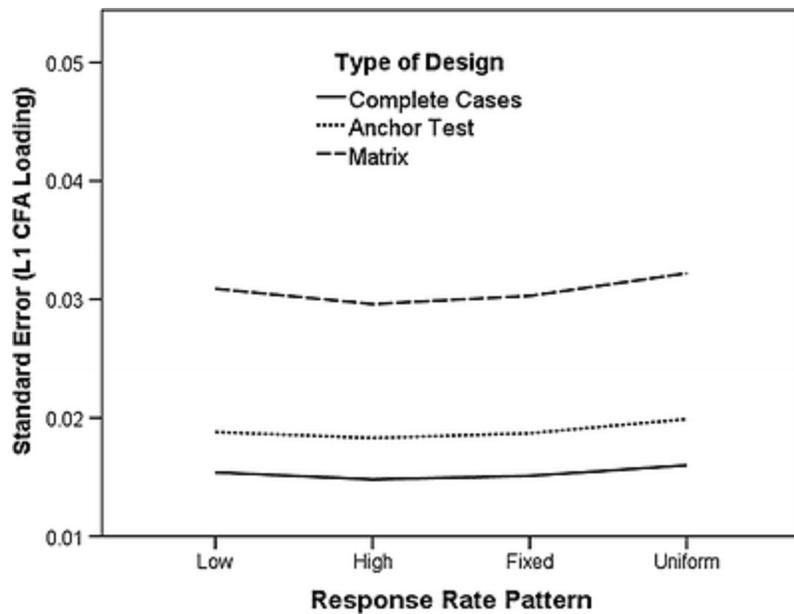
**Table 10** Effects of the planned missing-data design and response rates on parameter biases and standard errors

	Complete Cases				Anchor Test				Matrix			
	Low	High	Fixed	Uniform	Low	High	Fixed	Uniform	Low	High	Fixed	Uniform
Factor loading bias	0.049	0.056	0.021	-0.028	-0.035	0.042	0.070	0.049	-0.105	0.140	0.021	0.049
L2 main effect bias	-0.823	-1.208	-0.591	-0.694	0.925	-0.129	-0.334	0.000	0.848	0.334	0.977	0.360
Factor loading <i>SE</i>	0.0154	0.0148	0.0151	0.0160	0.0188	0.0183	0.0187	0.0199	0.0309	0.0296	0.0303	0.0322
Main effect <i>SE</i>	0.0650	0.0643	0.0640	0.0650	0.0652	0.0643	0.0650	0.0655	0.0654	0.0646	0.0648	0.0655

“Bias” is the percentage difference between the simulated value and the population value. *SE* = standard errors. Factor loadings are the loadings of individual items on the Level 1 latent social engagement variable. L2 main effects are the effects of social anhedonia, the Level 2 predictor, on the latent social engagement variable

### *Standard errors*

For standard errors, we again found an influence of the planned missing-data design. The standard errors for the Level 2 main effect were essentially identical across all conditions, as is shown in the last row of Table 10. However, the standard errors for the Level 1 factor loading (not surprisingly) varied across the conditions. These values are displayed in Table 10 and Fig. 6. As in our prior simulations, standard errors increased as the amount of missingness increased: They were lowest for the complete-cases design, somewhat higher for the anchor test design, and notably higher for the matrix design.



**Fig. 6** Effects of planned missing-data designs and response rates on Level 1 confirmatory factor analysis loading standard errors

The type of response rate had minor effects, at most: Standard errors were lowest for the high-response-rate conditions, but the variation between the types of response rates was trivial. As a result, variation in response rates appears to exert a minor influence, regardless of the kind of missing-data design.

### *Discussion*

The Monte Carlo simulations replicated the general findings from the prior simulation and from the broader literature on planned missing data. As before, missing data created minimal bias in the parameter estimates (factor loadings and regression weights), but they did increase the standard errors as the amount of missingness increased. In this simulation, the increase in standard errors was restricted to the Level 1 CFA. Beyond the prior study, we evaluated the possible influences of different patterns of response rates. Overall, these patterns created at most minor differences in standard errors. The trivial influence of different response-rate patterns should be reassuring, given that experience-sampling research always involves between-person variability in response rates.

### **General discussion**

In research, as in life, there is no free lunch, but there are occasionally good coupons. To date, experience-sampling research hasn't considered the value of planned missing-data designs, which have proven to be helpful in other domains in which time and resources are tight (Graham et al., 2006). Planned missing-data designs offer a way for researchers to ask more Level 1 items overall without presenting more items at each beep. But nothing comes for free—as the level of

missingness increases, the standard errors increase—so these designs should be applied only when the gains outweigh the somewhat reduced power.

The findings from our simulated designs resemble the findings from the broader missing-data literature (Davey & Savla, 2010; Enders, 2010; McKnight et al., 2007) and fit the underlying statistical theory (Little & Rubin, 2002; Rubin, 1976). Standard errors increased as the covariance coverage decreased. Anchoring the set of items by always asking one of the items yielded only slightly higher standard errors than always asking all questions. For all the designs, the inflated standard errors were restricted to the parts of the multilevel structural equation model that involved the items, as one would expect, so the designs did not cause a widespread inflation of error throughout the model. Higher standard errors, of course, reduce power to detect significant effects, so researchers should weigh the virtues of a planned missing design in light of its drawbacks.

### *Practical considerations*

When should researchers considering using a planned missing-data design in their experience-sampling study? We see planned missing-data designs as being particularly valuable when researchers intend to collect several items to measure a dependent variable and then average across them. This is common in experience-sampling research, particularly research on affect, social behavior, and inner experience. Instead of simply averaging to form the outcome, researchers can use a matrix or anchor test design and use shorter questionnaires. But like any method, planned missing-data designs should be used when their trade-offs are favorable, not because they seem novel, fancy, or interesting.

The primary payoff of these designs is space: Fewer items can be asked at each beep, so each daily-life survey will take less time. For example, consider a study with 20 items overall but only 15 items per beep, a savings of five items. The savings in space can be spent in several ways. First, it can be spent on lower burden. Asking fewer questions per beep reduces the participants' burden, which should lead to higher compliance rates. Deliberate nonresponse—ignoring the beeper, PDA, or phone—is a vexing source of missing data, and any simple methodological or procedural element that can boost response rates is worthwhile (Burgin et al., 2013). Simply keeping each survey short is its own virtue.

Second, the saved space can be spent on asking additional constructs. Researchers could add five new items, thus asking 20 items per beep out of a 25-item protocol. The amount of time that participants spend completing a survey will thus be the same, but a wider range of constructs can be assessed. In this case, somewhat higher standard errors would be traded for the ability to measure and analyze more constructs. Our intuition is that this option capitalizes the most on the strengths of a planned missing-data approach.

Third, the saved space can be spent on increasing the Level 1 sample size. As noted earlier, there is a tension between the number of items per beep and the number of beeps per day. Shorter

questionnaires can be administered more often. Instead of asking 20 items six times daily, for example, researchers might ask 15 items eight times daily. This would increase the Level 1 sample size, and thus power, all else being equal (Bolger et al., 2012). In some cases, having more Level 1 units will eliminate the effect of missing data on power.

Finally, the saved space can be spent to offset other influences on administration time. Many technologies aid with collecting responses (Conner & Lehman, 2012), but some take more time than others. In our lab, we have compared text surveys given via Palm Pilots to interactive voice response surveys using participants' own phones (Burgin et al., 2013). It took much less time to complete the same set of items on a Palm Pilot (around 55 s) than on a phone (around 158 s). In such cases, applying a planned missing design can reduce administration time, and thus presumably increase response rates.

Beyond saving space, a smaller virtue of planned missing-data methods is the reduction of automatic responding. When people get the same items in the same order dozens of times, they tend to respond automatically and to “click through” the survey. Planned missing designs, by presenting different item sets at each beep, disrupt the development of response habits, which should increase data quality.

We have focused on applying these designs to within-person constructs, but experience-sampling researchers could extend them to both levels of their multilevel designs. Planned missing-data designs are typically seen in cross-sectional designs, such as large-scale educational assessment (e.g., Willse et al., 2008) or psychological research in which time and resources are limited (e.g., Graham et al., 2006). For example, some Level 2 constructs might lend themselves to a matrix sampling or anchor test design. We should emphasize, of course, that planned missing-data designs involve a trade-off: Researchers are willing to accept higher standard errors in exchange for the ability to measure more things, thus allowing more questions to be asked in the same amount of time. We doubt that the trade-off usually favors planned missing designs for Level 2 constructs of core interest to a research project, but we should note that in theory, nothing prevents implementing these designs at both model levels.

### ***Planning for power***

When considering a study that would include planned missing data, researchers can estimate the influence of different designs on power by conducting small-scale Monte Carlo simulations of their own. Missingness increases standard errors, and hence reduces power, but this reduction decreases when the Level 1 and Level 2 sample sizes increase. Researchers can run a handful of simulations, using data from their prior work or from other research as population values to understand how different designs and sample sizes would influence power for their particular study. Muthén and Muthén (2002) provided a user-friendly description of how to use simulations to estimate power in Mplus, and several recent chapters have illustrated applications specific to experience-sampling designs (Bolger & Laurenceau, 2013; Bolger et al., 2012).

One virtue of using Monte Carlo methods to estimate power is that this can alleviate unrealistic fears of the effects of planned missing data. As the present simulations show, most of the influence of missing data is on standard errors, and this increase can be offset by higher Level 1 and Level 2 sample sizes. A few simulations will show researchers the sample points at which power is acceptably high.

### ***Future directions***

An important direction for future work will be to test the effects of planned missing-data methods on participants' behavior, such as response rates (e.g., number of surveys completed) and data quality (e.g., time taken to complete each survey, the reliability of the within-person constructs). Based on how experience-sampling researchers discuss the trade-offs between items-per-beep and beeps-per-day, one would expect shorter surveys to foster higher compliance. But the methodological literature in experience sampling is surprisingly small, and only recently have studies examined the factors—such as technical and design factors, personal traits of the participants, and aspects of the daily environment—that predict response rates and data quality (Burgin et al., 2013; Conner & Reid, 2012; Courvoisier, Eid & Lischetzke, 2012; Messiah, Grondin & Encrenaz, 2011; Burgin et al., 2013).

Collecting data on how planned missing designs influence compliance would also afford a look at how multiple types of missingness behave. Experience-sampling research has extensive missing data at Level 1. Scores are typically missing at the beep level—for instance, someone does not respond to a phone call or PDA signal, causing all items to be missing—but they are occasionally missing at the item level—such as when someone skips an item or hangs up early. Planned missing-data designs yield observations that are missing completely at random: The available data are a representative random sample of the total observations, and the Monte Carlo simulations reported here reflect this mechanism.

As McKnight et al. (2007) have pointed out, however, patterns of missingness in real data typically reflect several mechanisms. In an experience-sampling study, for example, some observations are missing completely at random (e.g., a technical failure in an IVR system or PDA that causes a missed signal), some are missing at random (e.g., people are less likely to respond earlier in the day, which can be modeled by using time as a predictor; Courvoisier et al., 2012), and some are missing not at random (e.g., in a study of substance use, polydrug users are probably less likely to respond while and after using drugs; Messiah et al., 2011). These mechanisms might interact in complex ways. For example, a planned missing-data design could reduce unplanned missingness by fostering higher compliance. Or, less favorably, when unplanned missingness is high, such as with volatile or low-functioning samples, the data could become sparse if planned missingness were added. In the absence of empirical guidance, researchers should take expected effective response rates into account when estimating power and include likely predictors of missingness in the statistical model (Courvoisier et al., 2012).

## References

- Arney, M. F., Crowther, J. H., & Miller, I. W. (2011). Changes in ecological momentary assessment reported affect associated with episodes of nonsuicidal self-injury. *Behavior Therapy, 42*, 579–588.
- Barrett, L. F., & Barrett, D. J. (2001). An introduction to computerized experience sampling in psychology. *Social Science Computer Review, 19*, 175–185.
- Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York: Guilford.
- Bolger, N., Stadler, G., & Laurenceau, J. P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285–301). New York: Guilford Press.
- Brown, L. H., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). When the need to belong goes wrong: The expression of social anhedonia and social anxiety in daily life. *Psychological Science, 18*, 778–782.
- Buckner, J. D., Zvolensky, M. J., Smits, J. J., Norton, P. J., Crosby, R. D., Wonderlich, S. A., & Schmidt, N. B. (2011). Anxiety sensitivity and marijuana use: An analysis from ecological momentary assessment. *Depression and Anxiety, 28*, 420–426.
- Burgin, C. J., Brown, L. H., Royal, A., Silvia, P. J., Barrantes-Vidal, N., & Kwapil, T. R. (2012). Being with others and feeling happy: Emotional expressivity in everyday life. *Personality and Individual Differences, 63*, 185–190.
- Burgin, C. J., Silvia, P. J., Eddington, K. M., & Kwapil, T. R. (2013). Palm or cell? Comparing personal digital assistants and cell phones for experience sampling research. *Social Science Computer Review, 31*, 244–251.
- Conner, T. S., & Lehman, B. J. (2012). Getting started: Launching a study in daily life. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 89–107). New York: Guilford.
- Conner, T. S., & Reid, K. A. (2012). Effects of intensive mobile happiness reporting in daily life. *Social Psychological and Personality Science, 3*, 315–323.
- Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009). Experience sampling methods: A modern idiographic approach to personality research. *Social and Personality Psychology Compass, 3*, 292–313.

- Courvoisier, D. S., Eid, M., & Lischetzke, T. (2012). Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics. *Psychological Assessment, 24*, 713–720.
- Davey, A., & Savla, J. (2010). *Statistical power analysis with missing data: A structural equation modeling approach*. New York: Routledge/Taylor & Francis Group.
- Duncan, T. E., Duncan, S. E., & Stryker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Mahwah: Erlbaum.
- Eckblad, M., & Chapman, L. (1986). Development and validation of a scale for hypomanic personality. *Journal of Abnormal Psychology, 95*, 214–222.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.
- Fleeson, W., & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. *Journal of Personality and Social Psychology, 97*, 1097–1114.
- Graham, J. M. (2008). Self-expansion and flow in couples' momentary experiences: An experience sampling study. *Journal of Personality and Social Psychology, 95*, 679–694.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.
- Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323–343.
- Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques* (2nd ed.). New York: Routledge.
- Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks: Sage.
- Kwapil, T. R. (1998). Social anhedonia as a predictor of the development of schizophrenia-spectrum disorders. *Journal of Abnormal Psychology, 107*, 558–565.
- Kwapil, T. R., Barrantes-Vidal, N., & Silvia, P. J. (2008). The dimensional structure of the Wisconsin schizotypy scales: Factor identification and construct validity. *Schizophrenia Bulletin, 34*, 444–457.
- Kwapil, T. R., Silvia, P. J., Myin-Germeys, I., Anderson, A. J., Coates, S. A., & Brown, L. H. (2009). The social world of the socially anhedonic: Exploring the daily ecology of asociality. *Journal of Research in Personality, 43*, 103–106.

- Kwapil, T. R., Barrantes-Vidal, N., Armistead, M. S., Hope, G. A., Brown, L. H., Silvia, P. J., & Myin-Germeys, I. (2011). The expression of bipolar spectrum psychopathology in daily life. *Journal of Affective Disorders, 130*, 166–170.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken: Wiley.
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86–92.
- McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford.
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*, 259–284.
- Messiah, A., Grondin, O., & Encrenaz, G. (2011). Factors associated with missing data in an experience sampling investigation of substance use determinants. *Drug and Alcohol Dependence, 114*, 153–158.
- Mooney, C. Z. (1997). *Monte Carlo simulation (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07–116)*. Thousand Oaks: Sage.
- Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15–40). New York: Routledge.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599–620.
- Oorschot, M., Kwapil, T., Delespaul, P., & Myin-Germeys, I. (2009). Momentary assessment research in psychosis. *Psychological Assessment, 21*, 498–505.
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*, 287–312.
- Rubin, D. E. (1976). Inference and missing data. *Biometrika, 63*, 581–592.
- Scherbaum, C. A., & Ferrerter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*, 347–367.
- Silvia, P. J., & Kwapil, T. R. (2011). Aberrant asociality: How individual differences in social anhedonia illuminate the need to belong. *Journal of Personality, 79*, 1315–1332.

Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: Dispositional and situational predictors of non-response in experience sampling research. *Social Science Computer Review*. doi:10.1177/0894439313479902

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Watson, D. (2000). *Mood and temperament*. New York: Guilford Press.

Willse, J. T., Goodman, J. T., Allen, N., & Klaric, J. (2008). Using structural equation modeling to examine group differences in assessment booklet designs with sparse data. *Applied Measurement in Education, 21*, 253–272.