Journal of Methods and Measurement in the Social Sciences,
Vol.10, No. 2, 95-110, 2019

Observation Effect in Ecological Momentary Assessments: A Study of Sun Protection Practices

Elizabeth Schofield **Jennifer L. Hay**
Yuelin Li
Memorial Sloan Kettering Cancer Center, New York, New York.

Daily diaries and ecological momentary assessments are plagued by the assessment itself becoming an intervention, known as the observation effect. Bayesian hierarchical level modeling is a technique to analyze repeated measures or multiple outcomes. In a study of twice-daily self-reporting of sun protection behavior among high-risk individuals, we investigate observation effects, agreement between retrospectively self-reported reminder effect and observation effect, differential observation effects, and consistency of behaviors. Participants who retrospectively reported no reminder effect showed a decrease in protective behaviors over time, whereas those who reported they were reminded showed sustained use. Advantages of the Bayesian methodology are demonstrated for assessing consistency of behaviors. Although we cannot observe prior behavior, we theorize that individuals experience an initial elevation at the onset of observation, though this unobserved increase is only sustained for a subset who later attribute this sustained behavior to a reminder effect. Implications for study designs with repeated observations are discussed.

Keywords: observation effect; ecological momentary assessment; mere measurement; Bayesian HLM

In behavioral outcomes research, self-report of behaviors plays an important role in data capture when an unobtrusive and objective measure is otherwise unavailable. Recall bias may be more prevalent over longer periods of report, such that ecological momentary assessments (EMA) (Moskowitz & Young, 2006) or similar use of daily diaries may reduce recall bias by limiting the amount of time on which the patient is reporting (Shiffman, Stone, & Hufford, 2008). Current cell phone and other mobile technology, including interactive voice response systems, can be harnessed to simplify such EMA data collection or produce a similarly developed intervention, further making these study designs more attractive (Heron & Smyth, 2010; Mundt, Perrine, Searles, & Walter, 1995).

The use of EMA data has led some to question whether the observation is itself an intervention when the outcome is a behavior, where the EMA serves as a reminder or otherwise provokes a change in behavior, especially in the case where some behaviors may be more socially desirable (Edwards, 1953) or desirable to researchers (McCambridge, Witton, & Elbourne,

2014). The extent to which measuring intentions affects behavior has been deemed the "mere measurement" (Morwitz & Fitzsimons, 2004) or "question-behavior" effect (Fitzsimons & Williams, 2000), while the effect that observation influences behavior is deemed the "Hawthorne" effect (Landsberger, 1958). This is also seen in operations management, where measurement can be, and often is, used to drive performance. For example, Lied and Kazandjian (1998) coined the term "Hawthorne strategy" in harnessing observational effects to improve clinical performance and quality among healthcare workers. Student achievement is impacted by a "repeated testing" effect (Campbell & Stanley, 1963). In fact, Samuel Messick's work investigated purposeful educational assessments - harnessing positive outcomes such as increased motivation, within the realm of "consequential validity" (Messick, 1989). Behavior and observation effects as exhibited via daily diaries, however, fundamentally differ from performance in repeated testing since the mechanisms through which repeated testing drives performance are retrieval practice and other practice effects (Roediger & Karpicke, 2006), not a reminder effect as hypothesized for EMA. Thus, the threat to internal validity inherent in "repeated testing" as described by Campbell & Stanley is not directly applicable. In a recent study using a mobile health app to serve as a reminder to implement specific health behaviors, Pirolli et al. (2017) found that daily reminders for 28 days did significantly and positively impact the execution of those health behaviors. Despite these concerns, findings have been mixed with respect to evidence of behavioral change due to EMA data collection. In a recent meta-analysis of observational effects on behavioral outcomes, results differed by behavior, where behaviors such as hygiene and physical activity showed small increases while behaviors such as blood donation, alcohol consumption, diet, and sexual behavior showed no change (Rodrigues, O'Brien, French, Glidewell, & Sniehotta, 2015). In that same meta-analysis, outcomes did not differ significantly by whether the behaviors were self-reported, such as in EMA data collection, or when measured objectively, such as with vaccination uptake or cancer screening. Similarly, French and Sutton (2010) describe a wide range of studies reporting changes in cognitive, emotional, and behavioral outcomes, but ponders whether other studies exist with null findings.

According to the *integrated behavior model* (IBM) (Fishbein & Yzer, 2003) of psychology, salience of a behavior in and of itself is not enough to drive behavior, an individual also must have intention. Similarly to the *theory of planned behavior* (Ajzen, 1991), from which the IBM evolved, intention is then driven by attitudes, norms, and perceived control. Thus, an observation effect may be more prominent when the participant already has positive feelings about the behavior being reported, is influenced by any perceived norms, or believes they are especially capable of performing the behavior. For example, in 1987 Greenwald and colleagues (Greenwald,

Carnot, Beach, & Young, 1987) reported a marked increase in voter turnout simply by measuring intentions and self-efficacy of voters prior to voting day; however, this finding could not be replicated in a second experiment (Smith, Gerber, & Orlich, 2003). There are further effects of skills, habit, and environmental constraints on the extent to which intentions are executed; thus, smaller effects for studies of blood donation, with more logistical issues and constraints, and larger effects for studies involving simpler tasks such as walking more steps per day, may be attributed to this theoretical model.

Bayesian analytic techniques, and Bayesian hierarchical level models (HLM) in particular, provide certain advantages and richness in a situation with a relatively small number of clusters (e.g., persons with longitudinal data). Frequentist HLM models with dichotomous outcomes and small numbers of clusters may suffer from non-identifiability (i.e., the model cannot be estimated) (McNeish & Stapleton, 2016), and there is some evidence that Frequentist p-values are not reliable for some HLM models due to misspecification of degrees of freedom (Kruschke & Liddell, 2018; Luke, 2017). Further, the theoretical advantages of Bayesian models and resulting credibility intervals, as a replacement to traditional results producing p values, have been documented in recent literature (Wagenmakers et al., 2018). The functional differences between a Bayesian analysis and a traditional Frequentist analysis are the inclusion of Bayesian priors on the model parameters, and the resulting full posterior distribution, rather than only estimating a few model parameters. In short, a Bayesian prior is a distributional assumption about parameters to be estimated; for example, we may assume that the odds ratio for gender on sunscreen use is a value with uniform probability between -20 and +20. Even minor constraints on the parameters, by way of Bayesian priors, help to identify the model with more stability so that complex models or models with a large number of clusters are estimable when parameters might be unidentifiable in traditional Frequentist models (Gershman, 2016). Further, pooling information across behaviors by way of a multilevel or HLM model allows for interpretation of the overall behavioral trends when each specific behavior contributes to a shared pattern. This “partial pooling” provides more accurate estimates of associations than not pooling and instead stratifying the analysis by behavior (Gelman, 2006). Finally, Bayesian models yield credible intervals of parameter estimates, for example, the 95% credible interval being the central portion of the estimated posterior distribution that contains 95% of the values. More specifically, one type of credible interval is the Highest Density Region (Gelman et al., 2020) such that all values within this interval have a higher probability density than values outside the interval. This differs from the Frequentist confidence intervals. The functional differences listed above are due to a philosophical difference in beliefs where Bayesian framework treats

parameters as being random and traditional Frequentist statistics treat parameters as fixed. The Bayesian credibility intervals have a more intuitive interpretation in that they describe the beliefs about the true association is likely one of the values within the range. Credible intervals in a Bayesian hierarchical model also lessen the problem of multiple comparisons (Gelman, Hill, & Yajima, 2012), a concern sometimes ignored in Frequentist models. The current study is a secondary analysis using Bayesian HLM analysis to assess the extent of an observation effect during a 14-day study of sun protection behaviors utilizing EMA data collection. The study addresses whether there was an overall increase in use of sun protection over time within the study, agreement of self-reported level of reminder at the conclusion of study to real-time self-reported behavior over the course of the study, whether increase in use over time was related to self-reported reminder or other demographics (i.e., age and gender), and the within-person consistency (i.e., variance and covariance) of the four behaviors.

Methods

Recruitment

First-degree relatives of melanoma patients were identified and recruited for participation in a longitudinal study of sun protection behaviors and the decision factors related to them (Shuk et al., 2012). Specifically, the study assessed "Did you use sunscreen", "seek shade", "use a long-sleeved shirt or pants", or "use a hat". The longitudinal surveys were collected twice per day, assessing morning and then afternoon sun protection, via interactive voice system (IVS) on a call to the participant's cell phone. Fifty-nine first-degree relatives of melanoma patients completed at least one telephone call and 53 (77%) were retained for the entire 14-day study.

At each morning and afternoon assessment over 14 days, the survey evaluated use (yes/no) of each of the four sun protection behaviors (i.e., sunscreen, shade-seeking, protective clothing, and hat), along with presence or absence (yes/no) of 21 specific environmental and situational decision factors such as "was it sunny or hot outside?" and "did you want to dress nicely?" found to be critical to decision making in formative work (Shuk et al., 2012). The twice-daily IVS surveys also assessed perceived personal risk of melanoma, self-efficacy related to sun protective behaviors, perceived efficacy of sun protective behaviors in preventing melanoma, and satisfaction related to their current sun protective behavior. At the conclusion of the study, a self-reported variable indicating degree to which the participant felt that the survey served as a reminder for use of sun protection was collected; specifically, participants were asked, "Did the surveys act as a reminder for sun protection" with response options of "no,

never", "yes, but only some of the time", and "yes, almost all of the time". Additional basic demographics were also collected for each participant during enrollment.

Statistical Analysis

In this paper, we investigate three aspects of a possible observation effect using these data. First, we assess whether there was any overall trend in use of sun protection, or behavior-specific time trends, over the course of the study. Next, we investigate differential time trends by self-reported level of reminder in the study. Finally, the potential reminder effect is assessed by participant age and gender, both overall and for specific sun protection behaviors.

Growth models testing time effect for each behavior. To assess potential increases in use of sun protection over 14 days (up to 28 observation times), we use a Bayesian HLM to analyze the repeated dichotomous assessments for the sun protection behaviors. In a simplified case with a single sun protection behavior, modeled over time, we might regress the outcome on a random intercept per-person, and include a main effect for time, where the time parameter would be used to test for an increase indicating a reminder effect. In the case of multiple behaviors, we extend that simple model by adding behavior-specific effects to both the mean and the time effect. The model thus follows a logit specification for the kth behavior of the ith participant at the jth time point such that

$$\Pr(y_{ijk} = \text{Yes}) \sim \text{logit}(\alpha_i + \beta_1^T \cdot I_k + \beta_2 \cdot time_{ij} + \beta_3^T \cdot I_k \cdot time_{ij})$$
$$\alpha_i \sim N(\alpha_{0i} + \alpha_{1i}^T \cdot I_k, \sigma)$$
$$\alpha_{1i} \sim MVN(\alpha_1, \Sigma)$$

where I_k is an indicator vector with length three, such that each element is an indicator for the three non-sunscreen behaviors (i.e., sunscreen is the referent group). A significant non-zero finding for an element of β_3^T indicates differential time effects by behavior, compared to sunscreen; a significant non-zero finding for β_2 indicates an overall time trend potentially attributable to a general observation effect. Random intercepts for each participant are estimated through the α_i term such that each participant has an individual mean use of each sun protection behavior. Non-linear time effects may also be included in models to account for time effect that may not be constant across the 14 days of study (e.g., utilization may increase over time within study until a point when stable utilization is attained).

Agreement of self-reported reminder effect and use. Next, engagement with each sun protection behavior was tested between participants who said they were "Always reminded", "Sometimes reminded", or "Never reminded"

about sun protection from the IVS surveys. To the simple, single behavior outcome model previously described, we would want to test whether the time effect was modified by reminder status via an interaction term. In our HLM model, we can assess not only whether self-reported reminder status affects longitudinal behavior, but whether this effect is differential by type of sun protection behavior. Thus, we have a two-way interaction of reminder status with time, and also a three-way interaction of reminder with time and behavior, with the "never reminded" category as referent. Specifically, sun protection behavior was modeled as

$$\Pr(y_{ijk} = \text{Yes}) \sim \text{logit}(\alpha_i + \beta_1^T \cdot I_k + \beta_2 \cdot time_{ij} + \beta_3^T \cdot I_k \cdot time_{ij} + \beta_4^T \cdot reminder_i \cdot time_{ij} + \beta_5^T \cdot I_k \cdot time_{ij} \cdot reminder_i)$$

with random effects as described above, and reminder treated as a vector. Significant positive ($\beta_4 > 0$) two-way interaction effects indicate that those who found the survey to be a reminder did indeed experience a general measurement effect over time in the study. Significant three-way interaction effects indicate differential time trends by sun protection behavior.

Models testing differential reminder by demographic group: interactions of gender, etc., by time. Finally, sun protection behavior is modeled as a function of the interaction between time and gender, and between time and age. This final model, using the same notation as above, incorporates the right-hand side of the previous equation but adds parameters for differential reminder effects by either gender or age as follows, where *X* is the gender or age stratum:

$$\Pr(y_{ijk} = \text{Yes}) \sim \text{logit}(\alpha_i + \beta_1^T \cdot I_k + \beta_2 \cdot time_{ij} + \beta_3^T \cdot I_k \cdot time_{ij} + \beta_4^T \cdot reminder_i \cdot time_{ij} + \beta_5^T \cdot I_k \cdot time_{ij} \cdot reminder_i + \beta_6^T \cdot X_i \cdot time_{ij} + \beta_7^T \cdot X_i \cdot I_k \cdot time_{ij} + \beta_8^T \cdot X_i \cdot I_k \cdot time_{ij} \cdot reminder_i)$$

Differential interactions of time and gender or age across behavior types (represented as β_6 and β_7), would indicate differential time trends for age groups or gender. Non-zero effects on the 4-way interaction, β_8, may be evidence that the measurement in the survey only serves as a reminder under certain conditions (e.g. gender and when the participant reports that they were reminded).

Within-person variation and covariation of behaviors. Using the base time effect model from the previous analysis, median within-person variance, covariance, and correlation of the four behaviors are extracted from the Bayesian simulation results. Credibility intervals are also calculated from the resulting full posterior distributions.

Analyses were conducted in R version 3.3.2 and primarily using STAN, implemented via the rstanarm package, for Bayesian analysis (Stan Development Team, 2016). Prior to the development of the rstanarm package, Bayesian methods required extensive programmatic coding, including maximization algorithms for manually coded likelihood formulas both for the complete data and for the individual regression parameters. With the introduction of the rstanarm package, the analyst need only be comfortable with basic R commands. For example, the *stan_lm*, *stan_lmer*, and *stan_glm* functions can be used for standard linear models, mixed effects models, and generalized linear models (e.g., logistic or Poisson regression), respectively, with arguments similar to those required for the more familiar *lm*, *lmer*, and *glm* regression functions. The functions require such minimal arguments as formula (i.e., the regression model), family (i.e., the parametric distribution of the outcome variable) and data, although informative prior density functions may also be provided by the user. Fit statistics such as the leave-one-out information criteria (LOOIC) (Vehtari, Gelman, & Gabry, 2016) are implemented via the *loo* function. The full posterior samples for all parameters are called using the *as.matrix.stanreg* function. Although other statistical software such as SPSS have begun to implement components of Bayesian analysis in more recent releases, the functionality is often limited to a Bayes factor or other summary that does not include the full posterior distribution of all parameters.

Bayesian model results for parameters of interest are reported as posterior medians with credibility intervals, with each model estimated using 1000 iterations on each of 4 Hamiltonian Monte Carlo Markov Chains and the first 500 iterations of each chain excluded as warmup. In general, default uninformative priors were used and a logit link for the binomial outcomes (i.e., used or did not use the behavior during the reporting interval). LOOIC is used for comparing model performance.

Results

This study cohort of 59 participants included 22 males and 37 females, ranging in age between 18 and 82 years (mean age of 49 years). Data included between 1 and 28 observations (each either morning or afternoon) per participant, for a total of 1,312 observations with up to four sun protection outcomes each. For HLM analyses, each of the four behaviors at each timepoint are treated as individual outcomes, so that a transposed dataset with a total of 5,248 records is modeled. Further demographic information is reported by Hay et al. (2017).

Growth models testing time effect for each behavior. The overall effect of time within the study appears slightly, though not significantly, negative in direction (posterior median OR = 0.97). Main effects of shade-seeking, hats, and protective clothing indicate that rates of shade-seeking are

comparable to sunscreen usage, while use of hats and protective clothing are less frequent. Interaction effects of time within the study and sun protection behaviors indicate that any time trends are comparable for sunscreen, shade-seeking, and hats, though use of protective clothing shows a slight increase (posterior median OR = 1.08) when compared to sunscreen use over time. Combined effects of the time main effect and the time by behavior interactions indicate that that there is an overall decrease in use of sunscreen, shade-seeking, and hats during the course of the study, but a slight overall increase in use of protective clothing. Large intervals for the posterior estimates of main effects of each behavior indicate a large source of variability may be unaccounted for in these models, even after inclusion of random per-person behaviors. A model including a quadratic time effect was assessed but deemed not to appreciably enhance model fit. Model results are presented as Model 1 in Table 1.

Agreement of self-reported reminder effect and use. At the conclusion of the study, 60% of participants reported that the survey "almost always" reminded them to use sun protection, while 17% reported it "never" did, with the remaining 23% reporting "sometimes". As shown in Table 2, completion rates, measured as the number of twice-daily reports received for each participant, did not significantly vary by reminder status ($F(2,50)=0.99$; $p=0.38$), and sun protection behaviors did not differ on either the first, midpoint, or the last report for any given behavior across reminder status groups. Averaging over all behaviors and all reporting timepoints using the Bayesian HLM models, participants who later reported that the survey served as a reminder did not differ significantly in behavior (posterior median OR = 0.82). However, interaction terms of the continuous time variable and the categorical reminder variable showed that participants who reported the survey always reminded them to use sun protection had a differential time trend (posterior OR = 1.23) than those who were not reminded. These complex findings are depicted in Figure 1, where protection is aggregated over all behaviors. Counter to an expected increase in use over time by those who reported a reminder, those who were not reminded showed a decrease over the week following study enrollment, while those who reported being reminded showed more consistent use (e.g., no increase and less decrease). Results are presented as Model 2 in Table 1.

Models testing differential reminder by demographic group: interactions of gender, etc., by time. As reported previously for this study, differences in behaviors were seen by gender and age group; for example, men and participants over 50 years of age were more likely to use hats than females or those under 50 (Hay et al., 2017). In this study, however, we assessed differential time and reminder effects by gender and age, across behaviors. The overall time trend did not differ markedly for those over 50

Table 1

Posterior median and 95% credibility intervals (CrI) for odds ratios (ORs) in Bayesian HLM analysis

Model	Predictor	OR (95% CrI)	Predictor	OR (95% CrI)
1	Time	0.97 (0.93, 1.00)	Shade x Time	1.01 (0.96, 1.06)
	Shade	1.04 (0.50, 2.20)	Hat x Time	1.00 (0.95, 1.05)
	Hat	0.47 (0.21, 1.02)	Clothing x Time	1.08 (1.02, 1.14)
	Clothing	0.39 (0.16, 0.98)		
2	Time	0.99 (0.79, 1.23)	Remind: Always x Time	1.23 (1.01, 1.44)
	Time^2	1.00 (0.98, 1.01)	Remind: Sometimes x Time	1.05 (0.83, 1.33)
	Shade	1.68 (0.70, 3.97)	Shade x Time^2	1.02 (1.01, 1.04)
	Hat	0.71 (0.23, 1.79)	Hat x Time^2	1.02 (1.01, 1.04)
	Clothing	0.62 (0.22, 1.63)	Clothing x Time^2	1.02 (1.00, 1.03)
	Remind: Always	0.82 (0.33, 1.89)	Remind: Always x Time^2	0.99 (0.97, 1.00)
	Remind: Sometimes	0.82 (0.32, 2.11)	Remind: Sometimes x Time^2	1.00 (0.98, 1.01)
	Shade x Time	0.74 (0.60, 0.91)		
	Hat x Time	0.77 (0.61, 0.92)		
	Clothing x Time	0.86 (0.69, 1.06)		
3	Male x Time	1.00 (0.76, 1.29)	Time x Male x Shade x Remind : Always	1.06 (0.88, 1.29)
	Time x Male x Shade	0.92 (0.67, 1.24)	Time x Male x Hat x Remind : Always	1.11 (0.84, 1.53)
	Time x Male x Hat	0.93 (0.63, 1.34)	Time x Male x Clothing x Remind : Always	0.88 (0.71, 1.10)
	Time x Male x Clothing	1.10 (0.79, 1.48)	Time x Male x Shade x Remind : Sometimes	0.92 (0.72, 1.16)
			Time x Male x Hat x Remind : Sometimes	1.39 (1.01, 2.02)
			Time x Male x Clothing x Remind : Sometimes	0.78 (0.61, 1.01)
4	Age x Time	1.05 (0.89, 1.27)	Time x Age x Shade x Remind : Always	0.74 (0.61, 0.90)
	Time x Age x Shade	1.29 (0.98, 1.60)	Time x Age x Hat x Remind : Always	0.76 (0.57, 0.98)
	Time x Age x Hat	1.30 (0.95, 1.74)	Time x Age x Clothing x Remind : Always	0.87 (0.70, 1.06)
	Time x Age x Clothing	1.08 (0.85, 1.37)	Time x Age x Shade x Remind : Sometimes	0.92 (0.72, 1.16)
			Time x Age x Hat x Remind : Sometimes	1.39 (1.01, 2.02)
			Time x Age x Clothing x Remind : Sometimes	0.78 (0.61, 1.01)

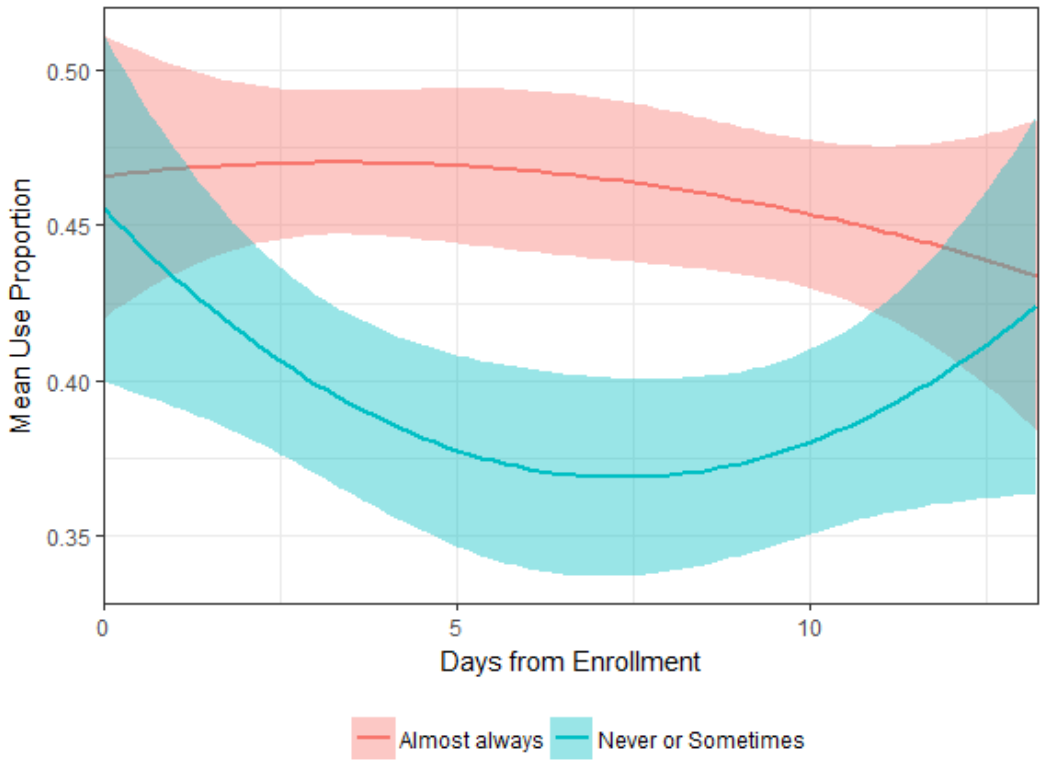
(posterior median OR = 1.05) or males (posterior median OR = 1.00), compared to participants under 50 years of age and females. However, participants over 50 years did show differential time trend for shade-seeking (OR = 1.29) and hats (OR = 1.30), compared to those under 50 years of age. Differential reminder effects on these two behaviors were also apparent for those over 50, where the 4-way interaction for those reporting the survey "almost always" reminded them was lower (OR for shade-seeking = 0.74; OR for hat = 0.76) than that for participants under 50 or who reported it "never" reminded them. Similar effects were seen for those reporting that the survey "sometimes" reminded them. Results are presented as Model 3 in Table 1.

The overall time trend also did not differ between males and females for sunscreen use (posterior median OR = 1.00), though some minimal gender-time effects were seen for other behaviors. For example, males had a slightly more positive time trend for use of protective clothing (OR = 1.10), indicating a more positive, or less negative time trend than females. The gender-time trend was modified by the self-reported reminder; the increased odds for males to use protective clothing is offset if they reported they were "almost always" reminded (OR = 0.88). This pattern is similar for those that reported a "sometimes" reminder. Results are presented as Model 4 in Table 1.

Table 2
Baseline (day 1), midpoint (day 7) and last day use of each sun protection behavior, by reminder status

	Never n = 9	Sometimes n = 12	Almost always n = 32	χ^2 (*df*)	*p*
Baseline					
Sunscreen	4 (44%)	5 (42%)	16 (50%)	0.28 (2)	0.87
Shade	5 (56%)	7 (58%)	15 (47%)	0.55 (2)	0.76
Hat	2 (22%)	4 (33%)	13 (41%)	1.08 (2)	0.58
Clothing	4 (44%)	4 (33%)	13 (41%)	0.30 (2)	0.86
Midpoint					
Sunscreen	4 (44%)	5 (42%)	17 (53%)	1.96 (2)	0.38
Shade	3 (33%)	5 (42%)	16 (50%)	0.87 (2)	0.65
Hat	2 (22%)	3 (25%)	14 (44%)	2.21 (2)	0.33
Clothing	3 (33%)	6 (50%)	14 (44%)	0.59 (2)	0.75
Last day					
Sunscreen	3 (33%)	5 (42%)	16 (50%)	0.87 (2)	0.65
Shade	5 (56%)	6 (50%)	12 (38%)	1.01 (2)	0.60
Hat	2 (22%)	2 (17%)	14 (44%)	3.52 (2)	0.17
Clothing	6 (67%)	6 (50%)	15 (47%)	1.11 (2)	0.58
n of obs, *M* (*SD*)	23.9 (5.1)	23.4 (3.4)	25.0 (3.3)		

Figure 1. Fitted quadratic curves of mean sun protection use, aggregated over all four behaviors, by day and participant-reported reminder status.



Within-person variation and covariation of behaviors. Using Model 1, the median within-person variance estimates across simulations for sunscreen, shade, hat, and protective clothing are 3.94, 6.99, 6.47, and 9.07, respectively. This indicates that an individual's use of protective clothing is generally much more variable than the most consistent behavior, sunscreen. Further, we can see by the full results in Table 3 that use of sunscreen is negatively correlated with the three other behaviors, such that shade-seeking and wearing hats or protective clothing are more likely to be used in conjunction or not at all, while sunscreen may serve as an alternative and be used instead of these other behaviors.

Table 3
Covariance and correlation matrices of within-person behaviors, medians with 95% credibility intervals

Measure	Behavior	Sunscreen	Shade	Hat	Clothing
Covariance	Sunscreen	3.94(2.6, 6.1)	-4.37(-6.8, -2.7)	-2.68(-4.9, -1.3)	-3.24(-5.4, -1.7)
	Shade		6.99(4.7, 10.4)	3.29(1.3, 6.0)	3.47(1.5, 6.1)
	Hat			6.47(4.2, 9.9)	2.66(0.6, 5.5)
	Clothing				9.07(6.3, 13.6)
Correlation	Sunscreen	1	-0.84(-0.9, -0.7)	-0.54(-0.7, -0.3)	-0.55(-0.7,-0.3)
	Shade		1	0.50(0.2, 0.7)	0.44(0.2, 0.6)
	Hat			1	0.36(0.1, 0.6)
	Clothing				1

Discussion

Observation effects may occur even in one-time, cross-sectional surveys, whereby research participants change their subsequent behavior after assessment of their behaviors or attitudes. Yet with frequent or repeated assessments, as utilized with EMA and mobile health (mHealth) applications, the potential for observation effects increases, given that the salience and self-monitoring of participants' behavior may increase with the use of EMA. That is, the repeated observations serve as a reminder about the behavior. This may be more of a concern with socially desirable or "healthier" behaviors. As such, measuring the extent to which this occurs, and under what conditions it occurs, is important methodologically for the planning of future EMA studies across many spheres of risk behavior assessment, and to clarify when assessment may function, inadvertently, as intervention. In the current study, despite being phoned twice daily for 14 days and asked about engagement of four different sun protection behaviors and environmental and social contexts related to sun exposure, one out of every six participants reported that the survey never served as a reminder for sun protective behaviors. Although there was no overall increase among all participants in sun protective behaviors over the course of the study, initial behaviors were sustained for those who reported being reminded, while initial behaviors were not sustained for those who reported they were not reminded. This pattern may imply that an observation effect is already present by day one, perhaps due to the novelty of knowing that observation is about to begin, or some level of social desirability effect, and that the inflated behaviors are only sustained for some individuals. A social desirability effect is especially plausible in this study as participants are at increased risk for skin cancer due to family history, and the behaviors are commonly-known sun protective practices. A baseline self-report of usual sun protection behavior over some period leading up to the study would be useful in confirming this theory and also possibly disentangling the observation effect at onset from meaningful behavior change.

While it is possible that for some individuals a twice daily phone call did not serve as a reminder to be more conscientious about sun protection, it is also possible that lack of behavioral response to the twice daily reminders led individuals to infer that they were not reminded. That is, although the behavior was made salient via the reminders, for some individuals this in itself was not enough to perpetuate the behavior, and thus these individuals then report that the survey did not remind them. This is similar to the self-perception theory, whereby humans use their behavior as clues to their own affect, or in this case whether or not the study served as a reminder (Bem, 1967). Perhaps because other criteria of the Integrated Behavior Model were not met, such as intentions aligned to the behavior, the behavior could not be sustained over the course of the study and thus these individuals then mis-attribute the role of the survey as a reminder.

The reminder effect may be more of a maintenance effect for a class of individuals who were susceptible to a twice-daily prompt, as these are the individuals that were able to perpetuate the initial increase in behavior. Similar to Messick's (1989) purposeful assessments of students, uptake in the behavior may occur in anticipation of the first assessment, which would lead to an unobserved initial elevation. A study by Cizza et al. inadvertently demonstrated what would have been an unobserved initial increase in behavior, thereby highlighting the importance of a behavioral assessment at screening, prior to the

study's reporting period (Cizza, Piaggi, Rother, & Csako, 2014). Such an initial elevation bias that then attenuates over time has recently been reported elsewhere (Shrout et al., 2018).

Differential reminder or maintenance effects were seen by age group for hats and shade seeking, as well as gender and long sleeve or protective clothing. This may be an indication that intentions, accessibility, and social norms differ for these behaviors by demographic groups. For example, if the average woman has a larger collection of protective clothing available to them than the average man, it stands to reason that even if they both had a behavioral response on the first day of observation and opted for protective clothing, the behavior could more easily be maintained by the female than the male. Thus, for future studies a baseline measure of self-efficacy for the sun protection behaviors would be useful. Notably, no significant difference was found between overall use of protective clothing by gender, rather this was a subtle effect related to time trajectories by gender and self-reported reminder status.

Despite both the increasing availability of Bayesian analytic tools and the widespread interest in applying Bayesian analysis to behavioral research, this work is among the minority in bringing such analysis into the peer-reviewed behavioral research literature. In our case, participants who reported no reminder effect in the survey had no differential behaviors at baseline, last day, or overall, only a subtle differential time effect can be extrapolated from the data. This effect is best captured via a multilevel model, which in turn enables us to estimate an even more powerful model – the multilevel model that incorporates information from all four sun protection behavior outcomes. A traditional multilevel model this complex with limited samples size may have identification problems and therefore may not even be estimable due to the number of parameters, but our use of Bayesian analysis sidesteps this problem. That is, Bayesian models are much more stable when the number of parameters estimated increases and the sample size is relatively small, even when traditional methods fail. The provision of the full Bayesian posterior distribution of the variance-covariance parameters is also a unique advantage to the Bayesian methodology and provides additional inference from the within-person contribution of this type of study design. The within-person covariance captures the correlations between behaviors on how one behavior affects another (e.g., need for sunscreen is reduced if the person is in the shade). Further, results from the Bayesian analysis are comparable in presentation (i.e., odds ratios) to a traditional analysis and this brings a level of familiarity to what may sound to some a new frontier in data analysis. The availability of packages such as rstanarm which make Bayesian models more accessible and less labor intensive enabled us to execute this work in a shorter time frame and without concern for errors in derivation and syntax from coding full likelihood functions and prior densities.

Although we have attempted to demonstrate the existence of an observation effect using this longitudinal study of sun protection practices using a Bayesian HLM framework, there are some limitations to this work. First, we have used what we believe to be post-exposure observations to deduce the existence of an observation effect induced response. As we do not have pre-study measures of participants' sun protection usage to serve as a true baseline, we may only speculate using a dropoff from initial usage to infer such a phenomenon. Other than the timing (only post-reaction) of observations, the dichotomization of behavioral outcomes at each interval also introduces a potential loss of statistical

power when quantifying sun protections. Although statistical power and sample size are less of a concern for Bayesian analysis than for traditional analysis due to the difference in interpretation of p-values, there is still a loss of granularity with this measurement. For example, shade-seeking or wearing a hat for just a small portion of an afternoon of full sun exposure is measured equivalently to an afternoon of actively and continuously utilizing sun protection, when with regards to risk it might need to be more closely aligned to non-utilization. An outcome with better gradation would be useful both in understanding predictors of sun protective behavior and also quantifying observation or mere measurement effect. Also, behaviors in the EMA-style are self-reported; an objective assessment of behavior, which would also allow for a control group, would allow a more direct study of the observation effects of interest. Further, dropout and intermittent missingness over the course of the study may well be informative due to the social desirability of sun protection practices for this population, first degree relatives of melanoma patients. Finally, if either an initial uptake in behavior or a maintenance effect of that uptake exists, it may be in part attributable to other factors besides the self-report of behaviors. For example, the twice-daily assessments included items on perceived risk, self-efficacy, efficacy of sun protective behaviors for melanoma prevention, and satisfaction with recent behavior. It's possible that the effects in our Bayesian HLM model which suggest a class of individuals susceptible to a maintenance effect may be due to a combination of these influences in addition to the reminder based on reporting behavior.

In addition to future work utilizing assessment of pre-study behavior, designs that alter reminders and measurement factorially may be helpful next steps in teasing apart these potentially distinct effects, as well as examining interactions. In our study, there were no reminders apart from assessments. In a comparable two-week study of protective health behavior, reminders could be varied by frequency (i.e., none, daily, weekly), or by level of harmonization with measurement (i.e., at the end of each repeated assessment or delivered randomly). Measurement could also be varied in terms of frequency (i.e., hourly, daily, weekly, or randomly). Such designs may clarify whether dose effects of either variable, or whether harmonization of variables, leads to the greatest increases in or maintenance of health behavior.

Author note: We acknowledge the funding support of National Institutes of Health (NIH) Grant R21 CA137532 (Jennifer L. Hay, Principal Investigator) and NIH Support Grant P30CA08748-48.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211. doi: 10.1016/0749-5978(91)90020-T

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonnance phenomena. *Psychological Review, 74*, 183-200. doi: 10.1037/h0024835

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Miffton Company.

Cizza, G., Piaggi, P., Rother, K. I., & Csako, G. (2014). Hawthorne effect with transient behavioral and biochemical changes in a randomized controlled sleep extension trial

of chronically short-sleeping obese adults: Implications for the design and interpretation of clinical studies. *PLoS ONE, 9*, e104176

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology, 37*, 90-93. doi: 10.1037/h0058073

Fishbein, M., & Yzer, M. C. (2003). Using theory to design effective health behavior interventions. *Communication Theory, 13*, 164-183. doi: 10.1111/j.1468-2885.2003.tb00287.x

Fitzsimons, G. J., & Williams, P. (2000). Asking questions can change choice behavior: Does it do so automatically or effortfully? *Journal of Experimental Psychology: Applied, 6*, 195-206. doi: 10.1037/1076-898X.6.3.195

French, D. P., & Sutton, S. (2010). Reactivity of measurement in health psychology: How much of a problem is it? What can be done about it? *British Journal of Health Psychology, 15*, 453-468. doi: 10.1348/135910710X492341

Gelman, A. (2006). Multilevel (hierarchical) modeling: What It can and cannot do. *Technometrics, 48*, 432-435. doi: 10.1198/004017005000000661

Gelman, A., Carlin, J., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. (2020). *Bayesian Data Analysis* (3rd ed.). New York, NY: Chapman & Hall.

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness, 5*, 189-211. doi: 10.1080/19345747.2011.618213

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71*, 1-6. doi: 10.1016/j.jmp.2016.01.006

Greenwald, A. G., Carnot, C. G., Beach, R., & Young, B. (1987). Increasing Voting Behavior by Asking People if They Expect to Vote. *Journal of Applied Psychology, 72*, 315-318. doi: 10.1037/0021-9010.72.2.315

Hay, J. L., Shuk, E., Schofield, E., Loeb, R., Holland, S., Burkhalter, J., & Li, Y. (2017). Real-time sun protection decisions in first-degree relatives of melanoma patients. *Health Psychology, 36*, 907-915. doi: 10.1037/hea0000523

Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology, 15*(Pt 1), 1-39. doi: 10.1348/135910709x466063

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review, 25*, 178-206. doi: 10.3758/s13423-016-1221-4

Landsberger, H. A. (1958). *Hawthorne Revisited*. Ithaca, NY: Cornell University.

Lied, T. R., & Kazandjian, V. A. (1998). A Hawthorne strategy: implications for performance measurement and improvement. *Clinical performance and quality health care, 6*, 201-204.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*, 1494-1502. doi: 10.3758/s13428-016-0809-y

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology, 67*, 267-277. doi: 10.1016/j.jclinepi.2013.08.015

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review, 28*, 295-314. doi: 10.1007/s10648-014-9287-x

Messick, S. (1989). Validity. In R.L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 13 – 103). New York: MacMillan.

Morwitz, V. G., & Fitzsimons, G. J. (2004). The Mere-Measurement Effect: Why Does Measuring Intentions Change Actual Behavior? *Journal of Consumer Psychology, 14*, 64-74. doi: 10.1207/s15327663jcp1401&2_8

Moskowitz, D. S., & Young, S. N. (2006). Ecological momentary assessment: What it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience, 31*, 13-20.

Mundt, J. C., Perrine, M. W., Searles, J. S., & Walter, D. (1995). An application of interactive voice response (ivr) technology to longitudinal studies of daily behavior. *Behavior Research Methods, Instruments, & Computers, 27*, 351-357. doi: 10.3758/BF03200429

Pirolli, P., Mohan, S., Venkatakrishnan, A., Nelson, L., Silva, M., & Springer, A. (2017). Implementation Intention and Reminder Effects on Behavior Change in a Mobile Health System: A Predictive Cognitive Model. *Journal of Medical Internet Research, 19*, e397. doi: 10.2196/jmir.8217

Rodrigues, A. M., O'Brien, N., French, D. P., Glidewell, L., & Sniehotta, F. F. (2015). The question-behavior effect: genuine effect or spurious phenomenon? A systematic review of randomized controlled trials with meta-analyses. *Health Psychology, 34*, 61-78. doi: 10.1037/hea0000104

Roediger, H. L., 3rd, & Karpicke, J. D. (2006). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science, 1*, 181-210. doi: 10.1111/j.1745-6916.2006.00012.x

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1-32

Shrout, P. E., Stadler, G., Lane, S. P., Joy McClure, M., Jackson, G. L., Clavél, F. D., . . . Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences of the United States of America, 115*, E15-E23. doi: 10.1073/pnas.1712277115

Shuk, E., Burkhalter, J. E., Baguer, C. F., Holland, S. M., Pinkhasik, A., Brady, M. S., . . . Hay, J. L. (2012). Factors associated with inconsistent sun protection in first-degree relatives of melanoma survivors. *Qualitative Health Research, 22*, 934-945. doi: 10.1177/1049732312443426

Smith, J. K., Gerber, A. S., & Orlich, A. (2003). Self-prophecy effects and voter turnout: An experimental replication. *Political Psychology, 24*, 593-604. doi: 10.1111/0162-895X.00342

Stan Development Team. (2016). *rstanarm: {Bayesian} applied regression modeling via {Stan}*. Retrieved from http://mc-stan.org/

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35-57. doi: 10.3758/s13423-017-1343-3