Appendix

Efficacy of contextually-tailored suggestions for physical activity:
A micro-randomized trial of HeartSteps

Appendix

I. Introduction to the Statistical Method

The analyses used in the paper use the methods in Boruvka et al. (1) as well as the required adjustments needed to form SEs and confidence intervals. This method accomodates the within person correlation of the time-varying proximal outcome (here 30 minute step count subsequent to a decision point) by using a robust standard error. Additionally this method centers the treatment indicators (akin to contrast coding) so that the estimator of the treatment effects used to form the test statistic are robust to misspecification of the model involving the control covariates.

In the case of time-invariant randomization probabilities, as was used in HeartSteps, the Boruvka et al. (1) method reduces to a GEE (2) with a centered treatment indicator with no specification of a correlation matrix. The latter is often misleadingly referred to as use of a "working independence correlation matrix," even though there is no assumption of independence between proximal outcomes at different time points. The use of the robust standard error accomodates the correlation between the proximal outcomes at different time points. Boruvka et al. (1) also include a number of small sample corrections used to improve the accuracy of the critical value in the test statistic and thus obtain the desired Type I and Type II error rates with smaller numbers of participants.

The rationale for use of this approach to construct the test statistics is due to the series of findings in the statistical literature reporting and proving bias in estimation of effects when a model for the correlation structure is included in the estimation and covariates are endogeneous (3,4,5). Here both availability as well as pre-decision point step count are likely impacted by prior treatment, and are thus endogenous. See the remarks in section 3 of Boruvka et al. (1) for further details. The assumptions underlying this hypothesis test are that the relevant moments are finite and the covariates are not collinear.

II. Primary Statistical Analyses

As discussed in the main paper, the primary analyses include data from the 37 participants. These participants provided at least 36 days out of 42 days of data, totaling 7540 decision points of which 6061 decision points were available. Recall availability is determined prior to randomization at a decision point. Furthermore, recall that a participant is unavailable at a decision point if the participant is currently driving, walking or running, or has just finished an activity bout in the previous 90 seconds. HeartSteps determined availability by using Android activity detection to assess participants' current activity at each decision point. Finally, since suggestion delivery required internet connectivity, participants were unavailable when their phones were offline.

The estimated effect, averaged over time in study and availability, of delivering a contextually-tailored activity suggestion vs. not delivering a suggestion is given by β_0 in the following regression. The regression is fit across all available participant-decision points for Y_{t+1},

$$\alpha_0 + \alpha_1 Z_1 + \beta_0(A_t - 0.6) \tag{1}$$

where

- Y_{t+1} is log-transformed total Jawbone step count in the 30 minutes following the t^{th} decision point,
- Z_t is log-transformed total Jawbone step count in the 30 minutes prior to the t^{th} decision point, and
- A_t is an indicator of whether or not treatment was provided at decision point *t*.

The analysis, using the methods of Boruvka et al. (2017), results in Table 1.

	Estimate	95%	95%	SE	Hotelling	p-value
		LCL	UCL		T-squared	
α_0	1.78	1.54	2.03	0.12	217.3	0.00
α_1	0.41	0.35	0.48	0.03	181.2	0.00
β_0	0.13	-0.01	0.27	0.07	3.8	0.06

Table 1: Fitted coefficients and univariate Hotelling's T tests for (1).

The estimated effect, averaged over availability, using a linear adjustment for day in study, of delivering a contextually-tailored activity suggestion vs. not delivering a suggestion on the first day of the study is given by β_0 in the following regression. The time trend, with study day, in this effect is given by β_1 in the following regression. The regression is fit across all available participant-decision points for Y_{t+1},

$$\alpha_0 + \alpha_1 Z_1 + \beta_0(A_t - 0.6) + \beta_1 d(t)(A_t - 0.6) \tag{2}$$

where

- Y_{t+1} is log-transformed total Jawbone step count in the 30 minutes following the t^{th} decision point,
- Z_t is log-transformed total Jawbone step count in the 30 minutes prior to the t^{th} decision point,

- $d(t)$ is the index of the day on which the t^{th} decision point occurred, ranging from 0 to 41, and
- A_t is an indicator of whether or not treatment was provided at decision point t.

The analysis, using the methods of Boruvka et al. (2017), results in Table 2.

	Estimate	95% LCL	95% UCL	SE	Hotelling T-squared	p-value
α_0	2.00	1.77	2.24	0.12	294.7	0.00
α_1	0.41	0.35	0.47	0.03	189.6	0.00
α_2	-0.01	-0.02	-0.00	0.01	5.1	0.03
β_0	0.51	0.20	0.81	0.15	11.4	0.00
β_1	-0.02	-0.03	-0.01	0.01	9.2	0.01

Table 2: Fitted coefficients and univariate Hotelling's T tests for (2)

III. Secondary Statistical Analyses

As discussed in the main paper, the secondary analyses include data from the 37 participants. These participants provided at least 36 days out of 42 days of data, totaling 7540 decision points of which 6061 decision points were available. Recall availability is determined prior to randomization at a decision point. Furthermore, recall that a participant is unavailable at a decision point the participant is currently driving, walking or running, or has just finished an activity bout in the previous 90 seconds. HeartSteps determined availability by using Android activity detection to assess participants' current activity at each decision point. Finally, since suggestion delivery required internet connectivity, participants were unavailable when their phones were offline.

These analyses use the methods in Boruvka et al. (2017) as well as the required adjustments to form SEs and confidence intervals.

The estimated effect, averaged over time in study and availability, of delivering a contextually-tailored walking suggestion vs. not delivering a suggestion is given by β_{01} and the estimated effect, averaged over time in study and availability, of delivering a contextually-tailored anti-sedentary suggestion vs. not delivering a suggestion is given by β_{02} in the following regression. The regression is fit across all available participant-decision points for Y_{t+1},

$$\alpha_0 + \alpha_1 Z_1 + \beta_{01}(A_{1,t} - 0.3) + \beta_{02}(A_{2,t} - 0.3) \tag{3}$$

where

- Y_{t+1} is log-transformed total Jawbone step count in the 30 minutes following the t^{th} decision point,

- Z_t is log-transformed total Jawbone step count in the 30 minutes prior to the t^{th} decision point,
- $A_{1,t}$ is an indicator of whether or not a walking suggestion was provided at decision point t, and
- $A_{2,t}$ is an indicator of whether or not an anti-sedentary suggestion was provided at decision point t.

Note that $A_{1,t} = A_{2,t} = 0$ when the participant is randomized to no suggestion at decision point t. The analysis, using the methods of Boruvka et al. (2017), results in Table 3.

	Estimate	95% LCL	95% UCL	SE	Hotelling's T-squared	p-value
α_0	1.78	1.53	2.02	0.12	217.5	0.00
α_1	0.41	0.35	0.48	0.03	182.0	0.00
β_{01}	0.21	0.04	0.39	0.09	6.0	0.02
β_{02}	0.03	-0.15	0.21	0.09	0.1	0.75

Table 3: Fitted coefficients and univariate Hotelling's T tests for (3).

The estimated effect, averaged over availability, using a linear adjustment for day in study, of delivering a contextually-tailored walking suggestion vs. not delivering a suggestion on the first day of the study is given by β_{01} in the following regression. The time trend, with study day, in this effect is given by β_{11}. The estimated effect, averaged over availability, using a linear adjustment for day in study, of delivering a contextually-tailored anti-sedentary suggestion vs. not delivering a suggestion on the first day of the study is given by β_{02} in the following regression. The time trend, with study day, in this effect is given by β_{12}. The regression is fit across all available participant-decision points for Y_{t+1},

$$\alpha_0 + \alpha_1 Z_1 + \alpha_2 d(t) + \beta_{01}\left(A_{1,t} - 0.3\right) + \beta_{11} d(t)\left(A_{1,t} - 0.3\right) + \beta_{02}\left(A_{2,t} - 0.3\right) + \beta_{12} d(t)\left(A_{2,t} - 0.3\right) \quad (4)$$

where

- Y_{t+1} is log-transformed total Jawbone step count in the 30 minutes following the t^{th} decision point,
- Z_t is log-transformed total Jawbone step count in the 30 minutes prior to the t^{th} decision point,
- $d(t)$ is the index of the day on which the t^{th} decision point occurred, ranging from 0 to 41,

- $A_{1,t}$ is an indicator of whether or not an active suggestion was provided at decision point *t*, and
- $A_{2,t}$ is an indicator of whether or not a sedentary suggestion was provided at decision point *t*.

Note that $A_{1,t} = A_{2,t} = 0$ when the participant is randomized to no suggestion at decision point *t*. The analysis, using the methods of Boruvka et al. (2017), results in Table 3.

	Estimate	95% LCL	95% UCL	SE	Hotelling's T-squared	p-value
α_0	1.99	1.75	2.23	0.12	292.8	0.00
α_1	0.41	0.35	0.47	0.03	190.4	0.00
α_2	-0.01	-0.02	-0.00	0.01	4.7	0.04
β_{01}	0.73	0.42	1.04	0.15	23.2	0.00
β_{02}	0.24	-0.18	0.66	0.20	1.4	0.25
β_{11}	-0.03	-0.04	-0.01	0.01	18.1	0.00
β_{12}	-0.01	-0.03	0.01	0.01	1.5	0.23

Table 4: Fitted coefficients and univariate Hotelling's T tests for (4).

IV. Details Regarding Why Some of the 37 participants have fewer than 42 days of data

A participant is defined to provide a day of data if the sensing/software is able to determine availability on this day and randomize at a decision point (if available), both of which required the participant's phone to have an internet connection. Of the 37 participants, 7 participants had less than 42 days of data. This happened for two reasons:

1. Participants had to end the study early (N=2). In the first case, the participant found out during the study they were going abroad, resulting in needing to end the study 6 days early. In a second case, due to scheduling availability, the exit interview was done on the 41st rather than 42nd study day.
2. When the remaining (N=5) participants made the study team aware of travel and/or technical issues that resulted in a lack of connectivity and thus required multiple days of excluded data (days>=3), the study team attempted to extend their participation beyond the original 42nd study day to compensate for these excluded days. However the number of additional days of data still fell shy of the number of excluded days. In one case, the participant had 20 days of data excluded due to myriad of technical issues that were not discovered until very late. This participant agreed to participate an additional 2 weeks to compensate, but would not agree to go beyond 2 weeks. In the end this participant provided 36 days of data. The other 4 participants were within 2-3 days of providing the 42 days of data.

V. Sensitivity Analyses

Imputing Missing Step Counts Using Google Fit

As discussed in the paper, we singly imputed a zero step count for any minute when the Jawbone tracker did not record steps; recall a step count might not be recorded because the participant is sedentary or because the participant is not wearing the Jawbone tracker. To assess sensitivity of the above analyses to this imputation we reconducted the analyses singly imputed missing Jawbone step data by using data from Google Fit, an Android application that uses the phone accelerometer to track steps.

Recall there are 6061 available decision points among 37 participants. Jawbone (JB) step count missingness and availability of Google Fit (GF) step counts are summarized below in Figure 1 and Table 10.

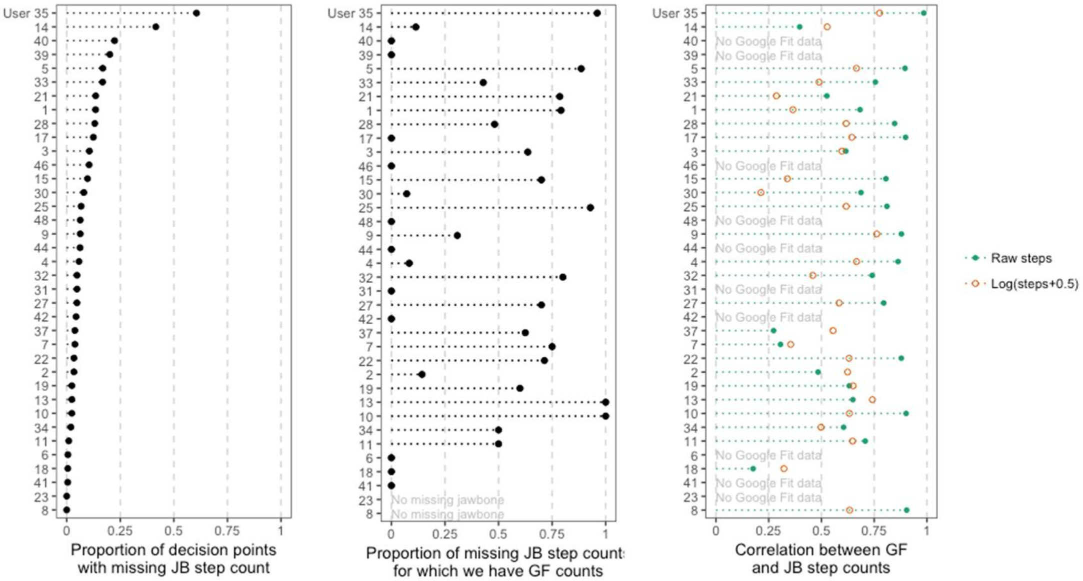

Figure 1. Missingness of Jawbone and Google Fit Step Counts by Participant

	Missing JB	Missing JB, available GF	Missing JB and GF
Num. available decision points	552	271	281
Proportion of all available & included decision points (6061)	0.09	0.05	0.05

Table 10. Summary of missing JB and GF step counts out of 6061 available & included, decision points

Single Imputation procedure

- When the Jawbone step count is missing (552 available decision points), use the Google Fit step count. 271 decision points are imputed in this manner.

- When both JB and GF step counts are missing, impute 0 for the step count. (281 decision points)
- Transform the result with log(step count + 0.5)

-

	Estimate	95% LCL	95% UCL	SE	Hotelling T-squared	p-value
α_0	1.94	1.73	2.14	0.10	367.5	0.00
α_1	0.39	0.33	0.44	0.03	194.8	0.00
β_0	0.13	-0.00	0.27	0.07	4.0	0.05

Table 11: Primary Analysis of Display 1,Table 1 imputing Google Fit step counts

	Estimate	95% LCL	95% UCL	SE	Hotelling T-squared	p-value
α_0	2.10	1.87	2.35	0.12	316.5	0.00
α_1	0.38	0.33	0.44	0.03	200.6	0.00
α_2	-0.01	-0.02	-0.00	0.00	4.0	0.06
β_0	0.52	0.21	0.82	0.15	11.6	0.00
β_1	-0.02	-0.03	-0.01	0.01	9.5	0.00

Table 12: Primary Analysis of Display 2, Table 2 imputing Google Fit step counts

	Estimate	95% LCL	95% UCL	SE	Hotelling's T-squared	p-value
α_0	1.93	1.73	2.14	0.10	367.9	0.00
α_1	0.39	0.33	0.44	0.03	195.4	0.00
β_{01}	0.21	0.04	0.38	0.09	6.3	0.02
β_{02}	0.03	-0.15	0.22	0.09	0.1	0.72

Table 13: Secondary Analysis of Display 3, Table 3 imputing Google Fit step counts

	Estimate	95% LCL	95% UCL	SE	Hotelling's T-squared	p-value
α_0	2.09	1.85	2.33	0.12	315.2	0.00
α_1	0.38	0.33	0.44	0.03	201.5	0.00

α_2	-0.01	-0.02	-0.00	0.00	3.6	0.07
β_{01}	0.74	0.43	1.04	0.15	24.1	0.00
β_{02}	0.25	-0.17	0.67	0.21	1.5	0.24
β_{11}	-0.03	-0.04	-0.01	0.01	19.0	0.00
β_{12}	-0.01	-0.03	0.01	0.01	1.5	0.23

Table 14: Secondary Analysis of Display 4, Table 4 imputing Google Fit step counts

Analyses Using Subsets of the 35 Participants Who Contributed More Than 38 Days of Data

We reconducted the primary and secondary analyses restricting attention to participants who contributed data for a longer number of days. 35 participants provided data for at least 38 days and 31 participants provided data at least 41 days.

38 days required

Effective sample size for these analyses is **35** participants, and **7187** participant-decision points.

	Estimate	95% LCL	95% UCL	SE	Hotelling's T-squared	p-value
α_0	1.78	1.52	2.03	0.13	200.4	0.00
α_1	0.41	0.34	0.47	0.03	161.2	0.00
β_0	0.12	-0.03	0.26	0.07	2.7	0.11

Table 21: Primary Analysis of Display 1, Table 1 Requiring 38 Days of Data

	Estimate	95% LCL	95% UCL	SE	Hotelling's T-squared	p-value
α_0	1.97	1.73	2.21	0.12	292.7	0.00
α_1	0.41	0.34	0.47	0.03	166.7	0.00
α_2	-0.01	-0.02	0.00	0.01	4.0	0.05
β_0	0.49	0.16	0.81	0.16	9.4	0.01
β_1	-0.02	-0.03	-0.01	0.01	8.0	0.01

Table 22: Primary Analysis of Display 2, Table 2 Requiring 38 Days of Data

	Estimate	95% LCL	95% UCL	SE	Hotelling	p-value
α_0	1.77	1.52	2.03	0.13	200.5	0.00
α_1	0.41	0.34	0.47	0.03	161.8	0.00
β_{01}	0.18	0.00	0.36	0.09	4.2	0.05
β_{02}	0.03	-0.17	0.23	0.10	0.1	0.76

Table 23: Secondary Analysis of Display 3, Table 3 Requiring 38 Days of Data

	Estimate	95% LCL	95% UCL	SE	Hotelling's T-squared	p-value
α_0	1.95	1.72	2.19	0.12	291.0	0.00
α_1	0.41	0.34	0.47	0.03	167.3	0.00
α_2	-0.01	-0.02	0.00	0.01	3.5	0.07
β_{01}	0.71	0.39	1.03	0.16	20.2	0.00
β_{02}	0.21	-0.24	0.66	0.22	0.9	0.34
β_{11}	-0.03	-0.04	-0.01	0.01	17.4	0.00
β_{12}	-0.01	-0.03	0.01	0.01	1.0	0.34

Table 24: Secondary Analysis of Display 4, Table 4 Requiring 38 Days of Data

41 days required

Effective sample size for this analysis is **32** individuals, and **6631** participant-decision points.

	Estimate	95% LCL	95% UCL	SE	Hotelling's T-squared	p-value
α_0	1.81	1.54	2.07	0.13	192.3	0.00
α_1	0.40	0.33	0.47	0.03	142.7	0.00
β_0	0.10	-0.059	0.25	0.07	1.9	0.18

Table 31: : Primary Analysis of Display 1, Table 1 Requiring 41 Days of Data

	Estimate	95%	95%	SE	Hotelling's	p-value

		LCL	UCL		T-squared	
α_0	2.00	1.74	2.255	0.12	258.1	0.00
α_1	0.40	0.33	0.463	0.03	147.6	0.00
α_2	-0.01	-0.02	0.000	0.01	3.9	0.06
β_0	0.55	0.21	0.884	0.16	11.1	0.00
β_1	-0.02	-0.03	-0.009	0.01	12.2	0.00

Table 32: Primary Analysis of Display 2, Table 2 Requiring 41 Days of Data

	Estimate	95%	95%	SE	Hotelling's	p-value
		LCL	UCL		T-squared	
α_0	1.80	1.54	2.07	0.13	193.3	0.00
α_1	0.40	0.33	0.47	0.03	143.1	0.00
β_{01}	0.16	-0.03	0.35	0.09	2.9	0.10
β_{02}	0.03	-0.18	0.24	0.10	0.1	0.76

Table 33: Secondary Analysis of Display 3, Table 3 Requiring 41 Days of Data

	Estimate	95%	95%	SE	Hotelling's	p-value
		LCL	UCL		T-squared	
α_0	1.99	1.73	2.24	0.12	258.7	0.00
α_1	0.40	0.33	0.46	0.03	147.9	0.00
α_2	-0.01	-0.02	0.00	0.01	3.5	0.07
β_{01}	0.73	0.38	1.08	0.17	18.8	0.00
β_{02}	0.32	-0.12	0.76	0.21	2.2	0.15
β_{11}	-0.03	-0.04	-0.01	0.01	19.0	0.00
β_{12}	-0.01	-0.03	0.00	0.01	2.8	0.11

Table 34: Secondary Analysis of Display 4, Table 4 Requiring 41 Days of Data

References

1. Boruvka, A., Almirall, D., Witkiewitz, K., Murphy, S.A. Assessing Time-Varying Causal Effect Moderation in Mobile Health (in press), *to appear in the Journal of the American Statistical Association*. Accepted author version posted online: 31 Mar 2017

2. Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22.

3. Pepe, M. S. and G. L. Anderson (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics - Simulation and Computation 23 (4), 939–951.

4. Tchetgen Tchetgen, E. J., M. M. Glymour, J. Weuve, and J. Robins (2012). Specifying the correlation structure in inverse-probability-weighting estimation for repeated measures. Epidemiology 23 (4), 644–646. Letter to the Editor.

5. Vansteelandt, S. (2007). On confounding, prediction and efficiency in the analysis of longitudinal and cross-sectional clustered data. Scandinavian Journal of Statistics 34 (3), 478–498.