Appendix S1

Data Pre-processing

The data required for analysis and modeling of the individual's behavior is in the form of four csv files; a table containing "availability time" of the participants, a table containing "date, time and step counts", a table containing "date, time and heart rate", and a table containing "messages, information about messages, weather indices such as temperature, real feel temperature, wind speed, precipitation, location of the participant, etc.".

The date and time format in the above mentioned csv files are different. So, the first step in preprocessing the data is to define a unique date and time format for all data. In this work, we use the 24 hour format for the time. It should be noted that due to the daylight saving, the time zone changes during the six months duration of the study occurred for most participants. In formatting the date and time data, this fact also was considered.

According to the results in the dynamical models reported by Conroy et al. (2019), the stepping response behavior is mostly different during the weekends in comparison to the weekdays. In light of this reasoning, we separate the data, into weekend and weekdays based on the day of the week. Then, we obtain the participant's availability time for weekends and weekdays from table containing availability time of the participants. As mentioned earlier, the participant specified the start time and end time of their availability to receive intervention messages during weekends and weekdays. The data which is used for modeling purposes, occur from an hour before participant's start time to one hour after participant's end time of availability. The reason for considering one additional hour after the participant's end time of availability is to have enough data for taking into account the effects of treatment messages that have been possibly sent at the end of the mentioned availability period by participant. Also, we keep the data from one hour ahead of participant's start time of availability so that we can incorporate it in the regression model of physical activity.

Fitbit data were reviewed in 1-minute epochs to classify device non-wear. Minute-level step data were valid if greater than zero or heart rate data were available. Device non-wear was classified when both heart rate recordings were missing and no steps were recorded for a minute epoch. Accordingly, the associated step data for those minutes are removed.

In the messages table, we have the report of scheduled, sent, received and acknowledged date and time for each message, but we may not have all of them for each scheduled/sent message. If the message has not been received by the participant (for example due to the smart phone being off, or poor Internet connection), it does not have value in the modeling of effects of treatment. So, we exclude the messages that have not been received by the participants. Similarly, because decisions to send messages in future interventions will be made with imperfect knowledge about whether a given message will be read, we included as inputs all messages that were received by the participants (not just those that were acknowledged as read).

The next step is to combine the table of date, time and step counts with the messages table (time associated with the "received" messages is considered). This is done for the both data of weekdays and weekends.

When the whole data related to each participant is gathered in two databases (for weekdays and weekends), it is time to consider a sample time of a set duration of time. For the numerical results in this paper, the sample time is considered to be 15 minutes. In order to apply the sample time to the data, for each consecutive 15 minutes in each day, we sum the values of step counts, move more messages, sit less messages, quotes, and get the average for temperature, precipitation, etc. It should be noted that applying the sample time is done on a daily basis. In other words, we consider the fact that effects of treatments in each day is for that day only and does not continue to the next day.

System Identification via Linear Regression Model with Multiple Variables

This method is used to estimate the parameters of dynamical models of physical activity for the case of modeling the effects of each type of treatment message to the stepping counts during weekdays and weekends. The linear regression model with multiple variables and noise is of the form

$$y(kd) = \quad a_0 + \sum_{i=1}^{5} a_i y(kd - id) + \sum_{j=1}^{3} \quad \sum_{i=0}^{5} b_{ij} u_j(kd - id) + \varepsilon(kd) \tag{1}$$

where $y(kd)$ is the system output at time kd which is the number of step counts at time kd, $u_j(kd - id)$ are the inputs (separate variables for move more, sit less, and inspirational quotes messages) to the system at time $kd - id$, d is the sampling time, $\varepsilon(kd)$ is noise at time kd, and a_0, a_i, b_{ij} are the unknown coefficients of the model.

In the model above the order is considered to be 5. The choice of the model order is done by performing a trade-off between model complexity and size of the model error. To identify the coefficients of model, we use least square method, which is a common approach in regression analysis to estimate the parameters of the model. Identifying the unknown parameters in the least square method is done by minimizing the sum of the squares of the residuals. The residuals are computed as

$$r(kd)= \quad y(kd) - a_0 - \sum_{i=1}^{5} a_i y(kd - id) - \sum_{j=1}^{3} \quad \sum_{i=0}^{5} b_{ij} u_j(kd - id) \tag{2}$$

And the least square is done by minimizing the sum of the square root of residuals as

$$\min_{a_0, a_i, b_{ij}} \sum_{k} r(kd)^2$$

$$subject\ to \quad r(kd)= \quad y(kd) - a_0 - \sum_{i=1}^{5} a_i y(kd - id) - \sum_{j=1}^{3} \quad \sum_{i=0}^{5} b_{ij} u_j(kd - id) \tag{3}$$

By solving the optimization problem in (4), the unknown parameters of the model are identified.

System Identification via Linear Parameter-varying Model

A linear parameter varying (LPV) system is a model whose dynamics vary as a function of several time-varying parameters. This method is used for estimating the effects of various temperatures in an individual's physical activity. The LPV model with noise is of the form

$$y(kd) = \quad a_0\big(p(kd)\big) + \sum_{i=1}^{5} a_i\big(p(kd)\big)y(kd-id) + \sum_{j=1}^{3} \quad \sum_{i=0}^{5} b_{ij}\big(p(kd)\big)u_j(kd-id) + \varepsilon(kd) \tag{4}$$

where $a_0\big(p(kd)\big), a_i\big(p(kd)\big), b_{ij}\big(p(kd)\big)$ are the unknown coefficients of the model that vary with parameter p at time k. The residual for this model is computed as

$$r(kd) = \quad y(kd) - a_0\big(p(kd)\big) - \sum_{i=1}^{5} a_i\big(p(kd)\big)y(kd-id) - \sum_{j=1}^{3} \quad \sum_{i=0}^{5} b_{ij}\big(p(kd)\big)u_j(kd-id) \tag{5}$$

Identifying the unknown parameters of model is done by solving the optimization problem as follows

$$\begin{gathered} \min_{a_0(p(kd)),a_i(p(kd)),b_{ij}(p(kd))} \sum_k r(kd)^2 \\ subject\ to\ \ r(kd) = \quad y(kd) - a_0\big(p(kd)\big) - \sum_{i=1}^{5} a_i\big(p(kd)\big)y(kd-id) - \sum_{j=1}^{3} \quad \sum_{i=0}^{5} b_{ij}\big(p(kd)\big)u_j(kd-id) \end{gathered} \tag{6}$$

In this work, the parameter p is considered to be temperature and the coefficients $a_0\big(p(kd)\big), a_i\big(p(kd)\big), b_{ij}\big(p(kd)\big)$ are considered to be quadratic function of parameter p at time kd, which is the interpolated temperature at time k, i.e.

$$\begin{aligned} a_0\big(p(kd)\big) &= \quad a_{00} + a_{01}p(kd) + a_{02}p(kd)^2 \\ a_i\big(p(kd)\big) &= \quad a_{i0} + a_{i1}p(kd) + a_{i2}p(kd)^2 \\ b_{ij}\big(p(kd)\big) &= \quad b_{ij0} + b_{ij1}p(kd) + b_{ij2}p(kd)^2 \end{aligned} \tag{7}$$

Validity

The error bound is computed to investigate the validity of the responses to each type of message as

$$conf_{forced-response} = \frac{mean\big(y_{forced-response}\big) * conf}{mean\big(y_{forced-response}\big) + mean\big(y_{natural-response}\big)} \tag{8}$$

where $y_{natural-response}$ is the natural response of system (zero input) and $y_{forced-response}$ is the forced response of system (with input; i.e. move more, sit less, and inspirational quotes messages). The confidence interval of $conf = \mu(r) \pm 2\sigma(r)$ is computed for residual margin, where r is the residual vector, $\mu(r)$ is the estimated mean of residual and $\sigma(r)$ is the estimated standard deviation of residual.