Markov Methodology to Improve Precision in the Assessment of Data-Intensive Single-Case Designs

Vincent Berardi[1], Ricardo Carretero-González[2], Marc Adams[3], John Bellettiere[1], Suzanne Hughes[1], Melbourne Hovell[1]

[1]San Diego State University, Center for Behavioral Epidemiology and Community Health
[2]San Diego State University, Department of Mathematics and Statistics
[3] Arizona State University

Single case designs that employ intensive longitudinal measures of health behavior have the potential to enable precise, nuanced evaluations of how individuals respond to interventions. Existing statistical tools typically focus on blunt, large effect outcomes such as changes in the level and/or slope of dependent variables over different intervention phases to arrive at a binary determination of whether an intervention was effective or not. This paper introduces a Markov-based, empirical transition method that offers a more comprehensive appraisal of interventions. The procedure projects a univariate time-series into discrete states and empirically determines the probability of transitioning from one state to another. Probabilities of state transition are computed separately for each intervention phase and are summarized in intervention phase-specific transition matrices. Comparing transition matrices illuminates intricate differences among intervention phases that can be quantified to estimate statistical significance via bootstrapping techniques. Analyses used to evaluate the face validity of findings are also discussed. This technique is demonstrated using three case studies from a quasi-SCD secondhand smoke reduction trial utilizing real-time air-particle sensors to discourage indoor cigarette smoking. The results enabled the identification of complex phenomena such avoidance and escape behavior in response to punitive contingencies. The Markov approach's ability to elucidate subtle behavioral details has not typically been feasible with standard methodologies, mainly due to historical limitations associated with infrequent repeated measures. The results suggest that Markov analyses can enhance the precision of outcomes provided by data-intensive single case designs, ultimately enabling the development of interventions uniquely tailored to specific individuals.

Introduction

Single-Case Design

Single-case designs (SCDs) are a class of research trials in which the unit of analysis is a single individual. SCDs are typically viewed as a fundamental departure from the tradition of between-group research designs in psychological research, yet the foundation of psychology laid in the late 1800's to early 1900's typically employed one or few subjects (Kazdin, 2011). Historically, SCDs have been used widely in education and behavior analysis (Kazdin, 2011; Horner et al., 2005) and more recently in other disciplines including exercise science (Sniehotta, Presseau, Hobbs, & Araújo-Soares, 2012), physical therapy (Boersma et al., 2004), nursing (Hentz, 2007), and pain management (Onghena & Edgington, 2005). While this literature illustrates the construct validity of SCDs, their use remains questionable to specialists more familiar with traditional randomized control trials (RCTs) (Normand, 2016). This paper describes how SCDs designed for understanding and changing behavior may be tested more completely when sufficient repeated measures are available.

SCDs can be used to evaluate a treatment approach over time, often with the goal of assessing the effect of an intervention for a specific person. This is accomplished by ob-

Research reported in this publication was supported by NHLBI of the National Institutes of Health under award number R01HL103684. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was also supported by the San Diego State University Computational Science Research Center and the ARCS Foundation. Corresponding concerning this article should be addressed to Vincent Berardi at Center for Behavioral Epidemiology and Community, San Diego State University, 9245 Sky Park Court,#230, San Diego, CA 92123, USA. Email: vinnieberardi@gmail.com

serving an individual over two or more distinct phases (intervals of time) both with and without an active treatment in place. The rationale underlying SCDs is the same as for randomized control trials (RCTs) and other group experimental designs, namely to allocate independent variables to conditions in order to isolate an effect on a dependent variable. In contrast to RCTs where independent variables are allocated by randomly assigning individuals to groups and measuring dependent variables infrequently (e.g. pre- and post-test), in SCDs the independent variable is allocated to phases for a single individual or a small number of individuals and the dependent variable is measured more frequently. The basic framework of an SCD includes a baseline phase (A), comprised of several repeated observations of the dependent variable without an active treatment. This baseline phase is followed by an intervention phase (B), defined by a discrete point in time where the independent variable begins to be experimentally manipulated for the duration of the phase, again with concurrent repeated observations of the dependent variable. In this AB type of single case design, the baseline period serves as the control, a counterfactual condition informing what would have occurred without treatment. More complicated study designs can be implemented, such as ABA designs that aim to observe reversal effects (Kazdin, 2011).

An advantage of SCDs is that the dependent variable is observed with far greater frequency than in traditional group-based experiments. This distinction makes SCDs an attractive option for researchers aiming to observe the precise nuances of how people interact with treatment protocols, specifically those that incorporate highly-individualized shaping procedures. Treatments such as these address some of known deficiencies associated with RCTs, including the reliance on a common treatment for all individuals regardless of their similarity to the "average" subject and restrictive inclusion/exclusion criteria that may reduce external validity, even with large sample sizes (Normand, 2016). These concerns have lead SCDs to be described as the strongest study design for making individual treatment decisions (Kravitz & Duan, 2014). Another beneficial attribute of SCDs is the possibility of aggregating many studies in order to generate an overall estimate of treatment effect on populations even when they are widely divergent in characteristics, leading to greater precision and generalizability, possibly at lower costs than traditional experiments due to the recruitment of fewer subjects (Zucker, Ruthazer, & Schmid, 2010).

Early in the history of SCDs, visual analyses of data plots were the typical analytic tool with researchers arguing that an advantage of this approach was that only large effects could be observed (Parsonson & Baer, 1986). Visual methodologies were sometimes augmented with simple statistical techniques such as celeration lines, the two-standard deviation band and the C-statistic (Nourbakhsh & Ottenbacher, 1994; Lindsley, 1992). In the mid-1970's through the 1980's, researchers began to focus on measures of statistical significance of differences between study phases using methodologies such as ANOVA and non-parametric methods like the Mann-Whitney U test, Wilcoxen's Signed Rank Test, and the Wampold-Worshom test (Gage & Lewis, 2013). Due to concerns that the inferential approach would lead to the inclusion of non-robust significant variables in the literature, the significance testing approach eventually gave way to effect-size calculations where researchers use their expert judgment to define the effect size of important findings (Gage & Lewis, 2013). These types of analysis techniques can be broadly grouped into four classes: i.) standardized mean differences such as Cohen's d (Cohen, 1992); ii.) overlap methods such as percentage of non-overlapping data (Mastropieri & Scruggs, 1985) and percent-zero data (Scotti, Evans, Meyer, & Walker, 1991); iii.) regression methods such as interrupted time-series (Wagner, Soumerai, Zhang, & Ross-Degnan, 2002), hierarchical linear models (Shadish, Kyse, & Rindskopf, 2013), and generalized estimating equations (Zeger, Liang, & Albert, 1988); and iv.) Bayesian approaches such as Bayes factors (de Vries & Morey, 2013) and hierarchical Bayesian modeling (Schluter & Ware, 2005).

The analytic techniques listed above aim to identify changes in one or more of the following across study phases: level, trend, variability, overlap, and immediacy of effect. When considering the variety of ways that individuals can respond differently to an intervention and the variety of contexts within which a single individual can receive an intervention, these measures of association appear rather blunt and may not reveal critical interactions that define contexts in which a given treatment may or may not be effective. This is particularly true if the treatment condition is not associated with a single event, such as the promulgation of legislation, but rather comprises many discrete procedures (e.g. repeated counseling, continued exposure to media). Because the metrics described above (level, trend, etc.) only assess gross differences between the baseline and treatment phases, the nuances of individuals' interaction with treatment are often not examined. This manuscript describes the development of a Markov chain methodology that has the potential to serve as a more comprehensive and intricate assessment of SCDs and individuals' responses to treatment. The approach developed to achieve this task requires a high sampling frequency with approximately $> O(10^3)$ observations per SCD. A meta-analysis of 409 SCDs indicated that the average number of observations in the baseline phase was 10.22 (Smith, 2012), meaning that the methodology under consideration is not appropriate for most current SCDs. However, high sampling frequency is becoming increasingly more common with sensors that record data on a near continuous basis (Madden & Franklin, 2002) in various contexts and certain treatments have already begun to implement these types

of tools (Klepeis et al., 2013; Adams et al., 2013). It is these types of studies that the procedures in this manuscript were developed to analyze. The variance within the large number of measurements generated enables techniques typically associated with dynamical systems to be employed allowing for far more complex patterns than simple changes in level, slope, etc. to be observed.

Project Fresh Air

While the methodologies detailed below are generalizable to any study with a sufficient number of observations, Project Fresh Air (PFA), a secondhand smoke (SHS) reduction trial meeting the data intensity requirements, was used as a prototype. This study aimed to ameliorate SHS exposure by leveraging punitive contingencies, which are defined as behavior(s) followed by a contingent aversive stimulus that makes the behavior(s) less likely to be emitted in the future. Approximately 300 homes were enrolled in this trial, each of which contained at least one adult who generated SHS (typically via indoor cigarette smoking) and at least one exposed child living in the home. To monitor indoor air quality, two Dylos DC1700 air particle quality monitors were installed inside of each home. The monitors were calibrated to detect particles with sizes ranging from 0.5 to 2.5 microns, which is consistent with SHS as well as non-tobacco aerosol sources such as cooking (Klepeis et al., 2013). One monitor was installed in the main smoking room and another was placed in the child's bedroom; measurements from only the main room monitor are included in the current analysis. The monitors measured the air particle concentration every ten seconds. In approximately half of the homes, the air particle monitors were fitted with devices that were programmed to deliver aversive visual and auditory feedback (yellow/red lights and tones) when air particle concentrations exceeded 60 $\frac{\mu g}{m^3}$. The intensity of the aversive feedback increased (Bellettiere et al., 2014) if a second 120 $\frac{\mu g}{m^3}$ threshold was breached. For these homes, the trial was stratified into two phases: 1.) *Baseline* – a period during which feedback was not active and 2.) *Treatment* – the period during which the feedback was activated, representing an AB SCD logic. The aim of the analyses presented below is to investigate the effect of the first (60 $\frac{\mu g}{m^3}$) threshold on a small subset of homes in PFA. This threshold was focused on since it was breached more frequently relative to the upper threshold. Additionally, there was no opportunity for the feedback associated with the lower threshold to be confounded. Conversely, the upper threshold feedback was preceded, and possibly confounded, by the lower threshold feedback.

Markov Chains and Transition Matrices

The proposed SCD analysis methodology presented herein is based on Markov chains. Markov chains (or processes) are discrete systems that, at any given time, can be characterized as being in a particular state. They are memoryless, meaning that the probability distribution of the system's next state is determined entirely by the current state. If there are n states, the transition matrix T is an n-by-n structure summarizing these probabilities over all states. $T_{i,j}$, the element of matrix T corresponding to the i^{th} row and j^{th} column, is the probability that a system in State i at a given time will be in State j at the next time step. For instance, $T_{1,1}$ is the probability that a system in State 1 remains in State 1 at the next time. $T_{1,2}$ is the probability that a system in State 1 moves to State 2, and so on. One row for each state is constructed and, by the law of total probability, the sum of each row must be 1. As an example, consider the matrix

$$T = \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.25 & 0.25 \end{pmatrix}.$$

In this case, when in State 1 at time t_i, the system will move to State 2 at time t_{i+1} with probability 1. When in State 2 at time t_i, the system has a 0.5 probability of moving to State 1 and a 0.5 probability of moving to State 3 at time t_{i+1}. At time t_i, if the system is in State 3, then at time t_{i+1}, the system will move to State 1 with probability 0.5, to State 2 with probability 0.25, and remain in State 3 with probability 0.25. The probabilities that populate the transition matrix can be determined either theoretically or empirically from data. Often, analyses focus on examining the eigenvalues and other properties of the transition matrix to determine the long-term behavior of the system, (Suzuki, 1995).

Empirical Markov Model for Project Fresh Air

Empirical Transition Matrices

The proposed SCD approach proceeds by empirically calculating transition matrices for each phase (A and B) of the intervention. In terms of PFA, one transition matrix is calculated for both the Baseline and Treatment phases of a single home as follows. Each data point in the time series of measurements generated by the air particle monitor was assigned to one of six states corresponding to the air particle concentration ranges summarized in Table 1. States are denoted as S_j for $j = 1 \ldots 6$; S_4 represents a particle concentration that has breached the first threshold that activates aversive feedback (i.e., 60 $\frac{\mu g}{m^3}$). Our interest is in particle generating behavior which, by definition, is associated with elevated air particle concentrations. Generally, low-level measurements ($<\approx 30\frac{\mu g}{m^3}$) are associated with background particle concentrations rather than particle generating events; therefore, they can safely be ignored and were not included in this analysis. Since the analysis focuses on the lower 60 $\frac{\mu g}{m^3}$ threshold, the maximum value of S_6 is 90 $\frac{\mu g}{m^3}$, the midpoint between the lower and upper feedback thresholds. Table 1 summarizes the number of observations in each state

State	Conc. Range $\left(\frac{\mu g}{m^3}\right)$	# Observations		
		HM162	*HM209*	*HM184*
$< S_1$	< 30	715,806	405,519	809,962
S_1	30 – 40	6,881	10,837	2,644
S_2	40 – 50	1,790	3,488	589
S_3	50 – 60	685	1,509	224
S_4^*	60 – 70	396	868	158
S_5	70 – 80	486	544	112
S_6	80 – 90	277	276	173
$> S_6$	> 90	654	1,478	1,399

Table 1
Range of particle concentrations for each of the states used in the Markov analysis along with the number of observations in each state for each of the three homes under consideration (HM162, HM209, HM184). S_4 has an asterisk since it represents the first state where a particle concentration has breached the first threshold and activated aversive feedback.

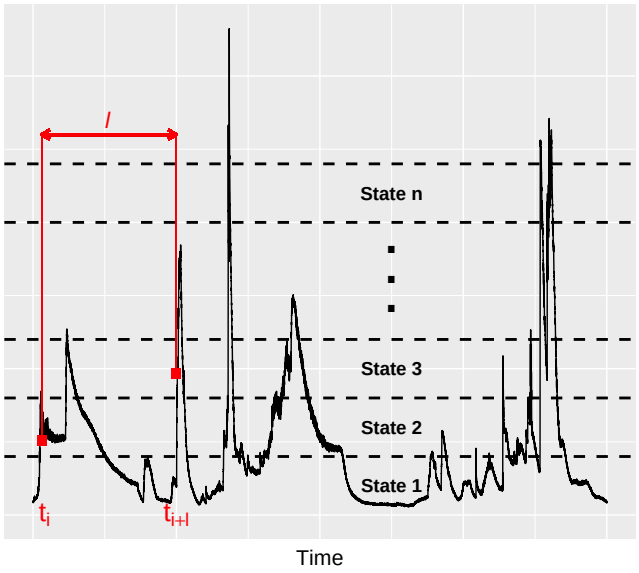

Figure 1. Schematic of the mechanism used to populate Markov process transition matrices based on an air particle time series. Time is shown on the x-axis and particle concentration is shown on the y-axis. At time t_i, the system is in state S_2. After a lag of l time units, denoted by the red double arrow, the system is in S_3. This process is performed for every observation so that the probability of moving from every state to every other state can be summarized.

for three homes that are discussed in detail below. The number of states and boundaries between states are parameters of this methodology that must be specifically determined for each study. States should be narrow enough to allow for frequent movement between them but wide enough so that a non-trivial number of observations fall within each state. Exploratory analyses indicated that the states defined in Table 1 achieved this goal for data from the PFA study. The conclusions discussed below were qualitatively similar when using up to twice as many states and/or expanding the maximum and minimum of the data values considered by up to $20\frac{\mu g}{m^3}$.

To populate the transition matrix, a *lag*, l, is selected. s_i was defined as the state of the system at time t_i and s_{i+l} was defined as the state of the system l units later at time t_{i+l}. (The last $l-1$ observations were eliminated from this analysis since the system's state l units later was not observed.) We call the s_i's *source* states and the s_{i+l}'s *destination* states. A schematic of this process is depicted in Fig. 1 for a single point where $s_i = S_2$ and $s_{i+l} = S_3$. For each S_j, consider all of the s_i's such that $s_i = S_j$, i.e., all of the observations in S_j. We then determined s_{i+l}, the destination state, for each of these observations. The raw counts were divided by $|S_j|$, the total number of observations contained in state j, to convert them into probabilities describing the transition from each state into every other state. This information was summarized in an empirical transition matrix, which described the probability of moving from one state to another at the next time step.

Define T_B as the empirically-determined transition matrix for the baseline phase of the intervention when the visual and audio feedback was not yet activated and T_T as the empirically-determined transition matrix for the Treatment phase once the feedback had been activated. If present, intervention effects should manifest themselves as differences between these two matrices, which can be summarized by $T_\Delta \equiv T_T - T_B$. T_B, T_T, and T_Δ were calculated for HM162, HM209, and HM184, three households in the PFA study. As will be demonstrated below, these homes were chosen since they exemplify different analytic results of interest. In each case, a lag of $l = 6$ measurements, or one minute, was used. Once the outcome measures have been fully described, the rationale for this choice of lag is discussed in the next section.

In Fig. 2, the first and second column of panels depict T_B and T_T, respectively, while the third column of panels depicts T_Δ. Generally speaking, the transition matrices have the largest probabilities along the diagonal, meaning that the source state s_i and the destination state s_{i+l} are the same. In other words, one minute after any given measurement, the concentration is most likely to have not changed by an amount large enough for it to have transitioned into another state. For HM162 and HM209, the intervention had the effect of increasing the probabilities in the subdiagonal entries, which represent, on average, a decrease in particle concentrations after l time units have elapsed. These increased subdiagonal values came at the expense of decreasing probabilities on the diagonal and superdiagonal. This result can be seen by

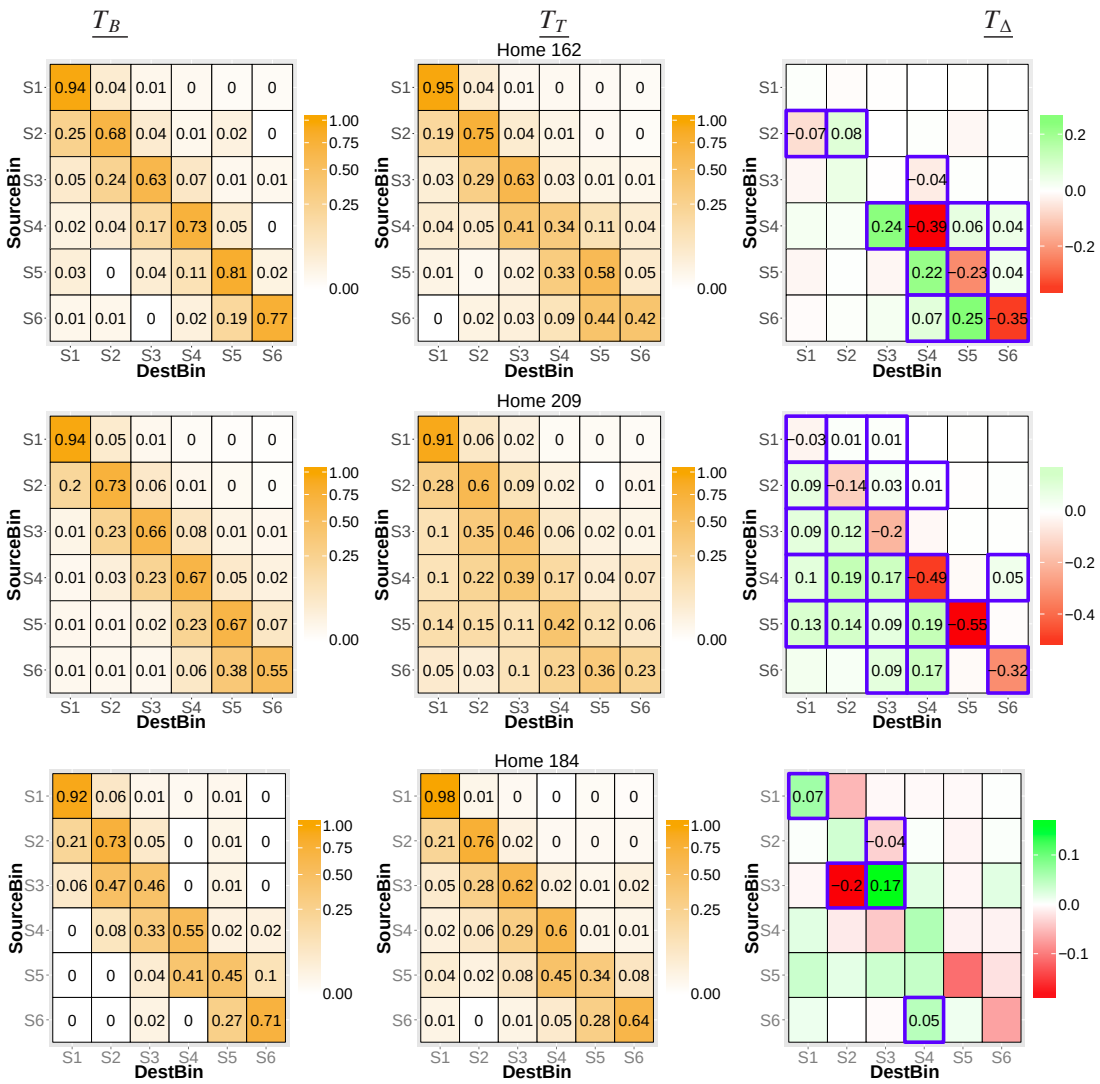

Figure 2. Transition matrices for three homes in PFA. The rows of each matrix represent the source states and the columns represent the destination states. The value of $T_{i,j}$ represents the empirically-calculated probability of the system being in State S_j when it was in S_i l units earlier. The left column of the figure contains baseline transition matrices T_B, the center column contains treatment transition matrices T_T, and the right column contains T_Δ, the difference between these two matrices. The rows of panels correspond, from top to bottom, to HM162, HM209, and HM184. For T_Δ, only those values determined to be statistically significant are shown. These cells are highlighted in blue.

examining the transition matrices T_B and T_T themselves or by observing the sign and location of significant differences highlighted in T_Δ. For these homes, the intervention had the effect of creating downward pressure on states, where, once a state had been reached, air particle concentrations were more likely to decrease in the Treatment phase compared to the Baseline phase. In HM184, there was no systematic pattern and only five of the source-destination pairs in T_Δ were significant.

Evaluation of Lag

The values along the diagonal of the transition matrices provide a metric by which to gauge the optimal value for the lag l. For smaller values of l, the probabilities on the diagonal are expected to be larger. Taken to the extreme, if l was chosen to be one measurement (10 sec for PFA), the system will be highly autocorrelated since there has likely been an insufficient amount of time for air particle concentrations to change; therefore, changes to a new state will be rare. As a result, transitions between states could be drowned out by a large number of source-destination observations remaining

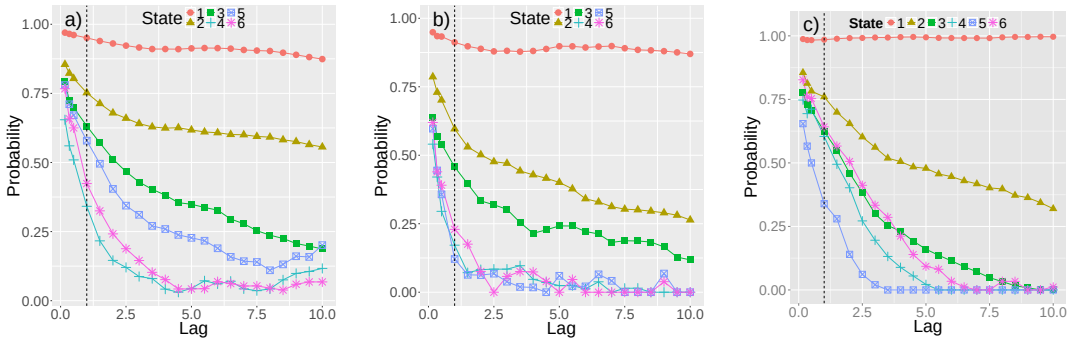

Figure 3. The value of the diagonal element corresponding to each of the 6 states for values of l ranging from 10 sec–10 min for Homes 162, 209, and 184, from left to right. The vertical dashed line represents a lag of 1 min.

in the same class. With larger values of l, the effect of the autocorrelation will decrease and for sufficiently large values there will be an essentially random relationship between the concentration at t_i and t_{i+l}. These two time points would be separated by such a long period as to not capture the same particle-generating dynamics (i.e. causal independent variables). The relationship would instead reflect an overall distribution of the different states. The choice of l must strike a balance between the extremes of the maximally autocorrelated and the overall distribution scenarios.

Figure 3 illustrates the values of each diagonal element of T_T for lags ranging from 1 to 60 measurements (10 seconds to 10 minutes). T_T was used for this calculation since this phase was associated with the greatest number of observations. The smallest variations occurred for S_1 through S_3, since these states were most likely to be associated with background measurements that have little variance. For S_4 through S_6 in each home, the diagonal probability value associated with a lag of $l = 1$ min was relatively small (< 0.5 in most cases) compared with the probabilities for the $l = 10$ sec case (> 0.9 in most cases). This indicates that for $l = 1$ min, the degree of autocorrelation between source and destination observations was reduced to the point that there were substantial numbers of transitions among states, possibly enabling patterns to emerge. But the $l = 1$ min value was also higher than the horizontal asymptote, which represents the mean, global distribution of probabilities associated with large lags. This means that there was meaningful correlation between source and destination states and the mapping captured more than just the overall distribution of the states. Because the $l = 1$ min probabilities balanced the competing interests of the asymptotic and pure autocorrelation scenarios, it was selected for all analyses. Exploratory analyses outside the scope of this paper indicated that lags ranging from 30 seconds to 3 minutes yielded results that were quantitatively similar to those summarized in Fig. 2. observed.

Significance of the Differences between T_B and T_T

The matrices T_Δ in Fig. 2 illustrate the difference between T_B and T_T. It is important to determine whether these differences were significant and reflective of distinctive dynamics in different intervention phases. Bootstrapped confidence intervals for each element of T_Δ were calculated to make this assessment. This procedure took advantage of the fact that each row of the transition matrix defines a multinomial distribution with n categories (each of the n destination states). The probability of the system moving into a given category (typically called a successful trial in a multinomial distribution) was given by the empirically-calculated, discrete distribution summarized in the row. It is possible to easily generate any number of values at random from a multinomial distribution. For the multinomial distribution defined by the row associated with a given S_j, $|S_j|$ values were randomly generated. These random values were then grouped according to the destination state into which they fell. Dividing the number of elements in each group by $|S_j|$ turns the randomly-generated values into a randomly-generated multinomial probability distribution. This procedure was repeated $B \equiv 10,000$ times for both the Baseline and Treatment phase matrices and for each sample the difference between the distributions was calculated, resulting in B T_Δ-like difference matrices. For these matrices, each element of a given row (or state/category of the multinomial distribution) was considered separately and was sorted from lowest to highest. The $0.025B^{\text{th}}$ and $0.975B^{\text{th}}$ value define a 95% confidence interval for a particular element. If the resulting confidence interval did not contain zero, then the change associated with this element in T_Δ was considered significant. Only significant changes are illustrated in the T_Δ matrices in Fig. 2; these probabilities are outlined in blue.

Behavioral Interpretation of Results

From the standpoint of behavior science, more-detailed conclusions about the dynamics of a household can be extracted from the transition matrices. The treatment included

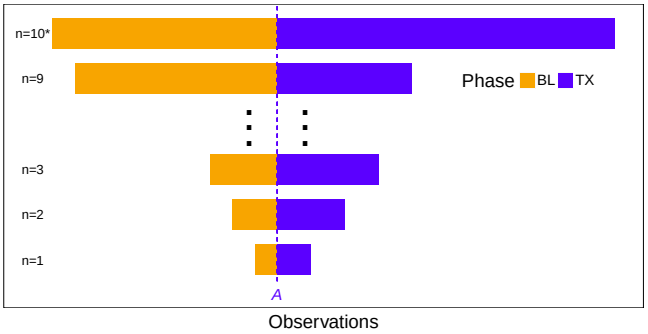

Figure 4. Schematic of the procedure used to evaluate the number of measurements for the Markov analysis to converge. The dashed line A indicates the onset of the intervention characterized by the availability of aversive monitor feedback. For $n = 1$, T_B was formed using only a small subset of the data adjacent and just prior to A. This is denoted by the small red rectangle on the lowest row. The small blue rectangle on the lowest row illustrates the small subset of data just after A that was used to construct T_T. For $n = 2$, the amount of data used to populate the transition matrices was expanded by 10% of the total number of measurements in each phase. This process continued until $n = 10$, when all of the data in each phase of the intervention was used to construct the transition matrices, which is the case that is illustrated in Fig. 2.

lights and tones emanating from the monitor once air particle concentrations were detected. This behavior-stimulus association is an example of an aversive/punishing contingency. By definition, an aversive/punishing contingency occurs when the presentation of a stimulus made contingent on a specific behavior results in a reduction of the behavior that led to the generation of the stimulus. This reduction can occur in two ways, *escape behavior* or *avoidance behavior*. In the escape paradigm, an individual performs behaviors to immediately alleviate the aversive stimulus associated with its action. For example, in PFA, once the aversive alarm has sounded, an individual may respond by extinguishing a cigarette or by moving outside in order reduce air particle concentrations triggering the sound. Avoidance behavior, on the other hand, is when an individual has discriminated the condition that led to the aversive stimulus and avoids the behavior or environment altogether. In PFA, after several instances of being exposed to aversive monitor feedback as a result of particle-generating behavior, an individual may move outdoors after lighting a cigarette so that particle concentrations do not trigger the monitor's feedback.

The T_Δ matrices appear to indicate that Home 162 exhibited escape behavior while Home 209 exhibited avoidance behavior. For Home 162, the reduction in the value of the diagonal probabilities and associated increase in subdiagonal probabilities only occurred for states S_4 through S_6, pre-

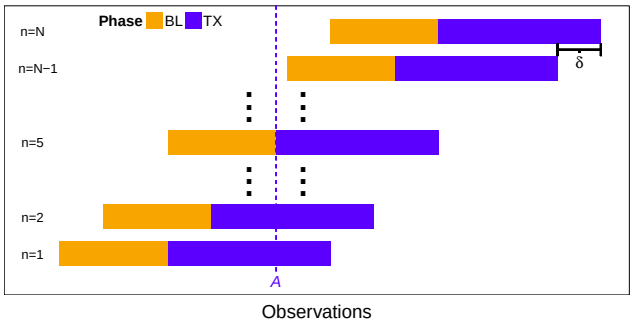

Figure 6. Schematic of the procedure used to evaluate optimal boundary between the Baseline and Treatment transition matrices. For the first iteration, the 1st through $(m_B + m_T)^{\text{th}}$ observations were used, with a boundary at m_B. For each subsequent iteration, the window of observations considered is shifted by δ, which is chosen so that at the fifth iteration, the boundary between the two phases aligns with A. For each iteration, transition matrices are empirically calculated and T_Δ, the difference between these matrices, is also determined. n is the iteration number.

cisely those states that triggered the monitor feedback. That is, household members in HM162 appear to be seeking relief from the aversive stimuli. In Home 209, though, the effect was present along diagonal and subdiagonal entries for all states, including those prior to the activation of feedback. This can be interpreted inferentially as the household adjusting their behavior in order to avoid triggering the alarm rather than reacting to the alarm once it has become engaged.

Validity of Results

The compelling results summarized in Fig. 2 reflect the stratification of measurements based on whether they are in the Baseline or Treatment phase of the study followed by an analysis of the difference between transition matrices associated with each phase. These results cannot be fully accepted, though, without investigating whether delineating the data by some other, possibly random, criteria would demonstrate similar or larger changes between transition matrices. If the largest changes are indeed associated with stratifying by intervention phase, this will increase the evidence for interpreting the results as a demonstration of the effectiveness of the intervention

This analysis proceeds by calculating i.) the appropriate number of measurements required for the system to converge to 'mean' behavior and ii.) the effect of stratifying the data based on intervention phase versus other stratification criteria.

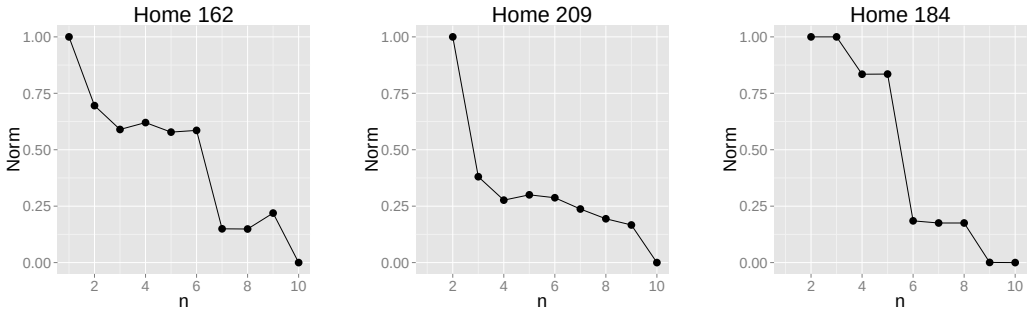

Figure 5. $\|T^*_\Delta - T^n_\Delta\|_F$ for $n = 1 \ldots 10$ for the three representative homes. In the two left-most cases, note the convergence to the results associated with the full amount of data as larger and larger subsets of data are used. Results were normalized by the largest value to make a visual comparison more tractable.

Variable	*Definition*	*Home 162*	*Home 209*	*Home 184*
T_B	Baseline transition matrix	-	-	-
T_T	Treatment transition matrix	-	-	-
T_Δ	T_T - T_B	-	-	-
T^n_Δ	Sequence of increasingly large T_Δ matrices	-	-	-
N_B	# obs. in Baseline phase	179,806	251,461	330,649
N_T	# obs. in Treatment phase	546,622	173,058	484,611
N	# obs. total	726,428	424,519	815,260
A	Boundary btwn. Baseline and Treatment phases	179,807	251,462	330,650
m_B	# obs. required for Baseline results to converge	125,864	176,022	231,453
m_T	# obs. required for Treatment results to converge	382,635	121,141	339,228
δ	Shift size for evaluating optimal boundary	13,486	18,860	24,799

Table 2
Variable definitions and, where appropriate, values for representative homes. See text for details.

Convergence of the Markov System

To assess the minimum number of observations required for the results to converge, for a given home, define the A^{th} observation as the first time point when the monitor feedback was activated and available should the air particle concentration exceed the lower threshold. $N_B \equiv A - 1$ is the number of observations in the Baseline phase and $N_T \equiv N - A + 1$ is the number of observations in the Intervention phase, where $N \equiv N_B + N_T$ is the total number of observations. (Table 2 illustrates the values of these variables for each of the homes under consideration.) The empirically-calculated probabilities summarized in Fig. 2 were based on using all N_B measurements in the Baseline phase for T_B and all N_T measurements in the Intervention phase for T_T. To investigate the convergence properties of our system the two sequences $B_n = N_B - n\frac{N_B}{10} + 1$ and $T_n = A + n\frac{N_T}{10}$ for $n = 1 \ldots 10$ were formed and the analysis summarized in Fig. 2 was then repeated for each value of n where T_B was calculated using the measurements from B_n to N_B and T_T was calculated using the measurements from A through T_n. As shown in Fig. 4, this had the effect of performing the Markov analysis for successively larger transition matrices with a boundary at A. $n = 10$ corresponds to the full analysis summarized in Fig. 2.

We now determine the number of observations required for the system to converge to the results obtained when using all N_B and N_T observations. For each value of n, denote the difference between the two transition matrices T_B and T_T as T^n_Δ. To determine the concordance of analysis results generated by the use of all the data (T^{10}_Δ) with analysis results generated by the use of subsets of the data of varying lengths, $\|T^{10}_\Delta - T^n_\Delta\|_F$ was calculated for each n. $\|Y\|_F$ represents the Frobenius norm of matrix Y, which is the L^2

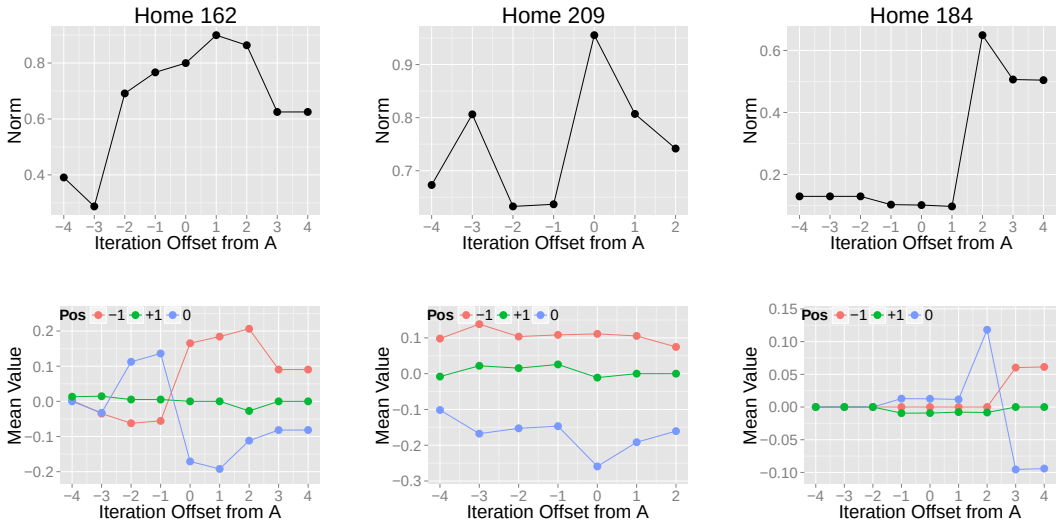

Figure 7. The top panel illustrates $\|T_\Delta\|_F$, which summarizes the difference between the two transition matrices that are being compared for various boundaries separating the data. The x-axis represents the number of shifts away from A, the boundary between intervention phases. The bottom panel illustrates the mean subdiagonal, diagonal, and superdiagonal values for each of the boundaries under consideration.

norm of a vectorized version of the matrix. When calculating T_Δ^n, all elements with a non-significant difference were set equal to zero. The results of this analysis are shown in Fig. 5 for each of the three homes under considerations. For Home 162, the norm of the difference between T_Δ^n and T_Δ^{10} approached convergence for all $n \geq 7$. $n = 7$ corresponds to using 125,864 measurements to calculate T_B and 382,635 measurements to calculate T_T. These values are denoted as m_B and m_T, respectively. For Homes 209 and 184, the system begins to converge at $n = 4$ and $n = 6$, respectively. For consistency with the previous case and for requirements for additional analyses to be discussed in the next section, we consider both of these this system to have converged for $n = 7$.

Optimal Boundary between T_B and T_T

The effect of shifting the boundary between the two empirical matrices was explored by implementing the following iterative procedure. Define m_B and m_T as the number of measurements in the Baseline and Treatment phases, respectively, required for the system to converge, as identified in the previous section. For the first iteration, the transition matrix T_B was generated from the 1^{st} through m_B^{th} observations in the study and T_T was generated from the $(m_B + 1)^{\text{th}}$ through the $(m_B + m_T)^{\text{th}}$ observations. The associated T_Δ matrix summarizing the difference between these two transition matrices was also calculated. In a sliding window-type procedure, the observation indices were then shifted by some value δ so that T_B was defined by the δ^{th} through $(m_B+\delta)^{\text{th}}$ observations and T_T was defined by the $(m_B+1+\delta)^{\text{th}}$ through the $(m_B+m_T+\delta)^{\text{th}}$ observations for the second iteration. T_Δ was again calculated. This procedure was repeated (i.e. δ was added again) as long as there were a sufficient number of observations to accommodate shifting the data window by δ. δ was chosen such that after five shifts, the boundary between T_B and T_T exactly matched A, the boundary between the Baseline and Treatment phases ($\delta \equiv \frac{A-m_B}{5}$). A schematic of this process is illustrated in Fig. 6.

Figure 7 illustrates $\|T_\Delta\|_F$ for various stratification points between the two transition matrices for the three representative homes under consideration. Larger norms indicate a greater difference between T_B and T_T and, therefore, a greater difference between the dynamics summarized by the matrices. A, the boundary between the Baseline and Treatment phases, is denoted as iteration 0 and all other windows are reported in terms of their iteration offset from A. For Home 209, the maximum of the norm is exactly at A, while for Home 162, the maximum occurs when the breakpoint is slightly offset from A. For Home 184, the pattern was slightly different and the norm was relatively low until it spiked when the breakpoint was offset by two iterations past A. In these cases, the norms are larger for stratification points that, generally speaking, most closely align with the breakpoint between the Baseline and Treatment phases. This indicates that stratifying by the intervention phase led to larger difference in dynamics than alternate boundaries, demonstrating the influence of the intervention in affecting household dynamics and adding validity to the findings.

The Frobenius norm provides a measure of the total action of a matrix, but yields no information about the structure of

the values within a matrix. It is possible for two matrices to have the same norm and completely different structures (As a trivial example, the identity matrix and a matrix of the same dimensions with 1's on the anti-diagonal and 0's elsewhere have the same norm, but, at most, one element in common). It only makes sense to compare norms across different boundaries between transition matrices if the structure of the underlying matrices are similar. The T_Δ matrices that have been discussed thus far have been characterized by negative values on the diagonal, positive values on the subdiagonal, and negative values on the superdiagonal. To assess the structure of the matrices calculated during the iterative, sliding-window procedure, the mean subdiagonal, diagonal, and superdiagonal values were recorded for each of the boundary iterations. These results are shown in the bottom panels of Fig. 7. For Home 209, the characteristic signs for the 3 interior diagonals are present across all boundaries. For Home 162, the expected structure of a negative diagonal and positive superdiagonal does not emerge until the boundary between the two matrices aligns with A. This is additional evidence that the intervention was critical in changing the dynamics of the home in a manner that is face valid. The results for Home 184 do not follow the same pattern. Its largest norm occurs at an offset of 2; however the expected pattern does not emerge until offsets of 3 and 4 from A. This is likely a function of the small effect size of the intervention for this home.

Discussion

This paper describes a Markov procedure that empirically evaluated the effects of an intervention with a level of detail not possible with typical methodologies. While in the traditional paradigm differences in the level, slope, or variation of a dependent variable are often used to arrive at a binary determination of whether an intervention was efficacious or not, this methodology offers the ability to identify much more nuanced, informative effects. In other words, whereas traditional SCD methodologies focus on "if" an intervention affects behavior, the empirical Markov methodology assesses "how" an intervention changes behavior. Our example resulted in outcome measures that can be inferred to correspond to avoidance and escape behaviors. It is likely that different behavioral phenomena will be observed when applying the empirical Markov procedure to other studies. For all cases, though, the precise level of detail offered by this approach provides an opportunity to assess interventions in a manner that is much more consistent with established behavioral theory. This should ultimately lead to more effective interventions.

To allow for detailed behavioral descriptions, this methodology requires a large number of repeated measures for a single individual. However, such intensive longitudinal data have not been the norm in behavioral interventions thus far. This trend is changing and more studies will have the requisite data measurement intensity (Walls & Schafer, 2006; Nahum-Shani, Hekler, & Spruijt-Metz, 2015; Spruijt-Metz et al., 2015). Wearable devices such as the Fitbit, Apple Watch, and a myriad of similar technologies enable a large number of physiological variables to be assessed continuously in near real-time. Big data from smart homes, networked cars, and the Internet of Things, which uploads data from sensors on physical devices (thermostats, washers/dryers, etc.) to networks, allow for additional behaviors and even the context in which they occur to be considered (Chui, Löffler, & Roberts, 2010). Real-time data generated by such devices enables a more comprehensive assessment of individuals than has ever been possible. The current trend of quantified self tracking, where individuals record certain aspects of their daily life with great precision, will only add to this ability (Swan, 2013). As this technology becomes more ubiquitous, there will be a greater number of opportunities to provide the type of personalized, data-intensive health interventions that are amenable to methods such as the empirical Markov approach (Khoury & Ioannidis, 2014). This process has already begun to take hold in studies concerning, for example, physical activity (Adams et al., 2013), dietary intake (Shroff, Smailagic, & Siewiorek, 2008), cigarette smoking cessation (Rodgers et al., 2005), and drug abuse (Butler et al., 2008). It has been suggested that current analytic approaches are not compatible with the intensive data streams generated by these studies and, therefore, new methodologies are needed (Riley et al., 2011). The empirical Markov model described herein has the potential to play this much-needed role.

The preponderance of data-centric studies in conjunction with analytic methodologies that are capable of elucidating detailed accounts of behavior (such as the empirical Markov methodology) might have the effect of increasing the prevalence of studies implementing single case designs as opposed to traditional between-subject designs. In such studies, an intervention effect may be sufficiently strong as too leave little doubt as to the efficacy of a treatment, especially if replicated in multiple individuals. This development would have the effect of moving the evaluation of treatment away from descriptions of "average" outcomes, which may conceal important functional relations, towards inductive assessments of actual observations (Normand, 2016). This development is likely to lead to more easily interpretable results that can better inform treatment decisions for patients.

Limitations in this approach are now outlined. Procedures for the management of missing/corrupt data must be developed. Furthermore, this analysis was only performed for three homes so the generalizability of the results to other homes is yet to be accomplished. Applying the empirical Markov methodology to a large number of homes or different types of behavior should be addressed in the future. A

key step in this process is the development of succinct metric to summarize the T_Δ matrices as opposed to the ad-hoc descriptions that were used above. Possibilities under consideration are a principal components decomposition of the matrix as well as pattern recognition approaches. For hybrid multiple-baseline-RCTs, where several SCDs are performed on a large scale, such as PFA, such a metric can be used to aggregate results in a way that would allow the overall intervention efficacy to be assessed with an uncommon level of precision at the individual level. This could function as an alternative or validation tool for more traditional statistics such as hierarchical linear models and generalized estimating equations.

Bibliography

References

Adams, M. A., Sallis, J. F., Norman, G. J., Hovell, M. F., Hekler, E. B., & Perata, E. (2013). An adaptive physical activity intervention for overweight adults: a randomized controlled trial. *PloS one*, *8*(12).

Bellettiere, J., Hughes, S. C., Liles, S., Boman-Davis, M., Klepeis, N., Blumberg, E., ... others (2014). Developing and selecting auditory warnings for a real-time behavioral intervention. *American journal of public health research*, *2*(6), 232.

Boersma, K., Linton, S., Overmeer, T., Jansson, M., Vlaeyen, J., & de Jong, J. (2004). Lowering fear-avoidance and enhancing function through exposure in vivo: a multiple baseline study across six patients with back pain. *Pain*, *108*(1), 8–16.

Butler, S. F., Budman, S. H., Licari, A., Cassidy, T. A., Lioy, K., Dickinson, J., ... Katz, N. (2008). National addictions vigilance intervention and prevention program (navipproâĎć): a real-time, product-specific, public health surveillance system for monitoring prescription drug abuse. *Pharmacoepidemiology and drug safety*, *17*(12), 1142–1154.

Chui, M., Löffler, M., & Roberts, R. (2010). The internet of things. *McKinsey Quarterly*, *2*(2010), 1–9.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological methods*, *18*(2), 165.

Gage, N. A., & Lewis, T. J. (2013). Analysis of effect for single-case design research. *Journal of Applied Sport Psychology*, *25*(1), 46–60.

Hentz, P. B. (2007). Case study: The method. *PL Munhall (Ed.), Nursing research: A qualitative perspective*, 359–384.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional children*, *71*(2), 165–179.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings*. Oxford University Press.

Khoury, M. J., & Ioannidis, J. P. (2014). Big data meets public health. *Science*, *346*(6213), 1054–1055.

Klepeis, N. E., Hughes, S. C., Edwards, R. D., Allen, T., Johnson, M., Chowdhury, Z., ... Hovell, M. F. (2013). Promoting smoke-free homes: a novel behavioral intervention using real-time audio-visual feedback on airborne particle levels. *PloS one*, *8*(8), e73251.

Kravitz, R., & Duan, N. (2014). Design and implementation of *n*-of-1 trials: a userâĂŹs guide. *Agency for healthcare research and quality, US Department of Health and Human Services*.

Lindsley, O. R. (1992). Precision teaching: Discoveries and effects. *Journal of Applied Behavior Analysis*, *25*(1), 51–57.

Madden, S., & Franklin, M. J. (2002). Fjording the stream: An architecture for queries over streaming sensor data. In *Data engineering, 2002. proceedings. 18th international conference on* (pp. 555–566).

Mastropieri, M. A., & Scruggs, T. E. (1985). Early intervention for socially withdrawn children. *The Journal of Special Education*, *19*(4), 429–441.

Nahum-Shani, I., Hekler, E. B., & Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology*, *34*(S), 1209.

Normand, M. P. (2016). Less is more: Psychologists can learn more by studying fewer people. *Frontiers in Psychology*, *7*.

Nourbakhsh, M. R., & Ottenbacher, K. J. (1994). The statistical analysis of single-subject data: a comparative examination. *Physical therapy*, *74*(8), 768–776.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical journal of pain*, *21*(1), 56–68.

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In *Research methods in applied behavior analysis* (pp. 157–186). Springer.

Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., & Mermelstein, R. (2011). Health behavior models in the age of mobile interventions: are our theories up to the task? *Translational behavioral medicine*, *1*(1), 53–71.

Rodgers, A., Corbett, T., Bramley, D., Riddell, T., Wills, M., Lin, R.-B., & Jones, M. (2005). Do u smoke after txt? results of a randomised trial of smoking cessation using mobile phone text messaging. *Tobacco control*, *14*(4), 255–261.

Schluter, P., & Ware, R. (2005). Single patient (*n*-of-1) trials with binary treatment preference. *Statistics in medicine*, *24*(17), 2625–2636.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation*.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological methods*, *18*(3), 385.

Shroff, G., Smailagic, A., & Siewiorek, D. P. (2008). Wearable context-aware food recognition for calorie monitoring. In *2008 12th ieee international symposium on wearable computers* (pp. 119–120).

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510.

Sniehotta, F. F., Presseau, J., Hobbs, N., & Araújo-Soares, V. (2012). Testing self-regulation interventions to increase walk-

ing using factorial randomized *n*-of-1 trials. *Health Psychology*, *31*(6), 733.

Spruijt-Metz, D., Hekler, E., Saranummi, N., Intille, S., Korhonen, I., Nilsen, W., ... others (2015). Building new computational models to support health behavior change and maintenance: new opportunities in behavioral research. *Translational behavioral medicine*, *5*(3), 335–346.

Suzuki, J. (1995). A markov chain analysis on simple genetic algorithms. *Systems, Man and Cybernetics, IEEE Transactions on*, *25*(4), 655–659.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, *1*(2), 85–99.

Wagner, A. K., Soumerai, S. B., Zhang, F., & Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical pharmacy and therapeutics*, *27*(4), 299–309.

Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data*. Oxford University Press.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049–1060.

Zucker, D. R., Ruthazer, R., & Schmid, C. H. (2010). Individual (n-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *Journal of clinical epidemiology*, *63*(12), 1312–1323.