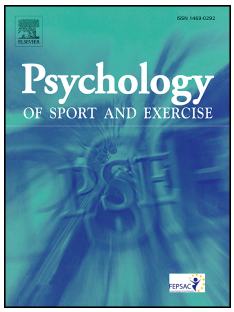


N-of-1 methods: A practical guide to exploring trajectories of behaviour change and designing precision behaviour change interventions

Dominika Kwasnicka, Felix Naughton

PII: S1469-0292(18)30797-0

DOI: https://doi.org/10.1016/j.psychsport.2019.101570

Reference: PSYSPO 101570

To appear in: *Psychology of Sport & Exercise*

Received Date: 20 December 2018

Revised Date: 29 July 2019

Accepted Date: 4 August 2019

Please cite this article as: Kwasnicka, D., Naughton, F., N-of-1 methods: A practical guide to exploring trajectories of behaviour change and designing precision behaviour change interventions, *Psychology of Sport & Exercise* (2019), doi: https://doi.org/10.1016/j.psychsport.2019.101570.

Title: N-of-1 methods: A practical guide to exploring individual trajectories and antecedents of behaviour

Dominika Kwasnicka and Felix Naughton

Author Note

Dominika Kwasnicka, SWPS University of Social Sciences and Humanities, Aleksandra Ostrowskiego 30b, 53-238, Wrocław and School of Public Health, Curtin University, Kent Street, Perth, 6102, Western Australia, +61 8 9266 2559, dkwasnicka@swps.edu.pl; and Felix Naughton, School of Health Sciences, University of East Anglia, Norwich NR4 7UL, United Kingdom, + 44 (0) 1603 593459, f.naughton@uea.ac.uk.

We would like to acknowledge all N-of-1 volunteers who took part in data collection and allowed their data to be used for training and demonstration purposes. Thank you to Diane Dixon and Carin Schroder for making their pre-whitening guide available. Finally, special thanks go to Peter Verboon for putting together the R script for this paper.

Correspondence concerning this article should be addressed to Dominika Kwasnicka, Faculty of Health Sciences, School of Psychology, Curtin University, Bentley Campus, Perth, Western Australia, 6845. E-mail: dominika.kwasnicka@curtin.edu.au.

Title: N-of-1 methods: A practical guide to exploring trajectories of behaviour change and designing precision behaviour change interventions.

Abstract

Objectives: (1) To introduce N-of-1 methods and how they can help the researchers identify predictors of behavioural outcomes, (2) to provide examples of studies that test individual theory-based predictions of physical activity and/or exercise; (3) to provide a practical example dataset to illustrate how to design and undertake a basic analysis for an N-of-1 study; and (4) to suggest a future agenda for N-of-1 physical activity and exercise research.

Design: Factors for consideration when designing an N-of-1 study include variability of predictors and outcomes, assessment frequency and appropriate analysis methods. Existing literature and piloting can help inform these aspects.

Methods: We use a dataset of 24 individuals who collected data over 28 days to illustrate example analysis procedures. Data, guidance and associated SPSS and R syntax are made available to provide researchers with tools to learn about and practice N-of-1 analysis.

Results: Guidance on dealing with missing data, looking at graphical representations of N-of-1 data, managing autocorrelation using the prewhitening method and analysing N-of-1 datasets is provided. Using the example dataset, we demonstrate how to identify antecedents of physical activity (steps) to assess directionality of associations. We also include an overview of aggregating N-of-1 datasets using multilevel modelling.

Conclusions: N-of-1 methodology provides a means of tracking individual patterns of behaviour and identifying potential antecedents of physical activity and exercise to help determine causality. Assisted by mobile technologies, there is great potential to enrich our understanding of movement behaviour using this approach to inform interventions.

Keywords: N-of-1, idiographic methods, within person design, N-of-1 analysis, R, SPSS, statistics

Background

In the context of behaviour change, a 'traditional' scientific model principally makes the underlying assumption that behaviour change interventions and treatments work in a similar way in all people, where researchers calculate an average effect across individuals. Therefore, individuals are considered, to some extent, interchangeable, meaning the identification of behavioural cause and effect in one person would apply to other people. If we anticipate individual differences, then we can proceed to the identification of subgroups of individuals for whom the assumption will be accurate. Once we have identified a subgroup of people that the given individual is part of, then we can apply the relevant intervention or treatment to them. However, this conventional scientific model is not a true representation of a personalised or person specific approach. Most intervention development frameworks and approaches that incorporate user perspectives as part of the design process (Bartholomew, Parcel, & Kok, 1998; Yardley, Morrison, Bradbury, & Muller, 2015) do not lead to interventions that provide truly individualised interventions. Such frameworks typically lead to interventions for an average person or at best averages within sub-groups of people. While tailored intervention development frameworks (Dijkstra & De Vries, 1999) can facilitate the generation of truly individualised interventions, most often the logic developed for these interventions are based on group level data (Naughton et al., 2014). Therefore, the intervention that is effective for some people may not necessarily work for others and may even be harmful for some.

In line with a truly person-specific approach, people are not considered interchangeable and correct identification of behavioural predictions and associated outcomes in one group of people or one subgroup of people, may not apply to the individual that we want to provide a treatment for. Applying a person-specific approach, researchers need to identify person-specific predictions that are relevant to outcomes for the person that they will

treat or intervene on. In order to develop a person-specific treatment or intervention, the researchers would need to assess which treatment/intervention with what content, intensity and delivery mode is the most suitable to that given person. Such a high degree of personalisation for behavioural interventions has various challenges that we will mention throughout this article.

Between-subjects randomised controlled trials (RCTs) are considered to be on top of the hierarchy of evidence (Lilienfeld, McKay, & Hollon, 2018). When conducting an RCT the researchers are testing a treatment or intervention between individuals looking at the difference in average effects between a comparator and a treatment group. A valid problem with RCT design, as well as with other nomothetic approaches (i.e., group level aggregated approaches), is that even with a successful treatment/intervention group there are people who do not respond to the treatment or even for whom the treatment is harmful. Reporting only the average effects, the researchers often loose vast amounts of information about the treatment effectiveness and suitability of the treatment effects within the individuals.

Another issue with conventional randomised designs is that these designs often rely on assessments of cognitions and outcomes at specific time points, e.g., baseline and follow up. This does not enable the assessment of how predictor and outcome variables may vary over time, e.g., a person can report low stress levels today but it does not mean that on average this person is not stressed. When designing exercise promotion interventions, intervention participants are often asked to wear a pedometer for a week before (T0), after the intervention (T1) and then at the follow up (T2). Variability in the outcome of interest throughout the intervention and throughout the non-assessment period is usually not considered and other potentially influential effects not accounted for (e.g., social desirability bias, life events).

N-of-1 – idiographic methods

To overcome the aforementioned problems, researchers can employ idiographic designs (i.e., within person designs) to (1) better understand trajectories of predictor and outcome variables over time; (2) to explore association between the predictor and outcome variables and also (3) to test and evaluate treatment or treatments within individuals and (4) to test theories within individuals. Idiographic designs are often called N-of-1 studies, single case studies, within-person studies etc. N-of-1 studies test hypotheses within individuals based on repeated measurement of variables within the individual over time. N can refer to an individual but also to a family, school or geographical region. N is a unit that the assessment is relevant to and repeated on, so for instance the researchers can assess different schools that take part in an exercise promotion program and they can compare how each school performs over time. Studies may include just one unit of interest (e.g., one person or one school) but researchers can also look into multiple units of interest and sometimes they aggregate these to identify predictors of outcomes and intervention effects.

Intra-individual effects may differ from those found in between-participant studies. For example, on average an intervention can be successful in increasing physical activity of individuals but looking more closely into N-of-1 data we can explore trajectories of change in participants who did not change their activity levels or even decreased their activity over time despite engaging with the intervention. In the idiographic study, it is enough to have just one participant or one study unit (e.g., one school) as power of the study is determined by the number of repeated observations not by the number of study participants or study units, although as several parameters need to be considered when undertaking a power calculation, further guidance should be sought (Bolger, Stadler, & Laurenceau, 2012; Kwasnicka et al., 2019). A fully powered N-of-1 study may include one person that was repeatedly assessed 50 times or even 300 times. This is in some ways comparable to a study that has 50 or 300

participants respectively. However, the issues of data autocorrelation need to be considered, as data points are no longer independent observations like in RCTs; we will further describe issues of autocorrelation when we discuss an example dataset.

Behavioural theories apply to individuals; however, they are usually tested in nomothetic approaches in groups of individuals. This mismatch between the aim of theory and application and testing in behavioural studies is problematic (Johnston & Johnston, 2013). To best understand predictions of behaviour and to personalise interventions and treatments, we need to understand mechanisms of action within individuals (Nielsen et al., 2018). Idiographic design has been used in health psychology to a fairly limited extend. For instance, McDonald, et. al., (2017) identified only 39 studies that used N-of-1 design in the health psychology and behavioural science field and most of them relied on fairly limited statistical methods and did not use appropriate N-of-1 types of approaches (e.g., N-of-1 RCT). In this review 14 studies were relevant to physical activity (McDonald et al., 2017). Another recent systematic review of N-of-1 RCTs suggested that this methodology could be the next major advance in health psychology and behavioural science for precision medicine (Shaffer, Kronish, Falzon, Cheung, & Davidson, 2018); however, the studies published so far often lack methodologic and statistical rigour and are not always transparently and fully reported. Idiographic design allows developing and conducting precision behaviour change studies; however, it is underutilised in psychology and studies published so far do not always follow best practice.

Examples of N-of-1 exercise and physical activity studies

Two main types of N-of-1 design are observational and experimental. Observational N-of-1s are usually purely exploratory in nature and the repeated assessment is used to understand patterns of cognitions, predictor variables and outcome variables and temporal associations between them. For instance, a recent observational study with healthy young

adults, (N = 79) who reported only intermittent exercise explored if stress causes decreases in levels of exercise, or if exercise causes decreases in stress levels or if the relationship was bidirectional (Burg et al., 2017). For 12 months participants engaged in stress monitoring by Ecological Momentary Assessment (EMA; at the beginning, during and end of the day) and continuous activity monitoring using Fitbit. A random coefficients linear mixed model was applied to predict end-of-day stress from the occurrence/lack of exercise that day; a logistic mixed model was used to predict the occurrence/lack of exercise from ratings of anticipated stress; separate regressions were performed for each participant. The results were a significant average negative effect of exercise on stress and of stress on exercise. However, there was between-person variability across 69 participants; exercise was associated with a stress reduction for 15, a stress increase for 2 and no change in stress for 52. An increase in anticipated stress reported the previous night or that morning was associated with a significant 20–22% decrease (OR = 0.78–0.80) in the odds of exercising that day across the whole group of participants. Again, when looking at the 69 participants individually, this increase in stress reduced the likelihood of exercise for 17, increased the odds for 1, and had no effect for 51. The authors concluded that the relationship of stress to exercise can be uni- or bi-directional and varies from person to person. The study highlighted the importance of assessing within person predictions of exercise and temporal associations.

Another recent observational N-of-1 study explored the relationship between theoretical predictors and outcomes looking at predictors of physical activity, adherence to weight loss plan and weight change (Kwasnicka, Dombrowski, White, & Sniehotta, 2017). The authors used idiographic methods to explore the predictive variables associated with weight loss maintenance. Eight people who intentionally lost 5% and more of body weight took part in the study and for 6 months daily collected objective measures of physical activity through Fitbit and weight through Wi-Fi connected scales. They completed EMA surveys

twice a day exploring theory-based predictors of behaviour change maintenance and their personal self-selected predictors. They also engaged in proactive experience sampling (i.e., participant-initiated, event-contingent sampling) – collecting contextual information regarding their activity and weight changes (pictures and notes). Each participant's data was treated as a separate data-set and first analysed separately (details of analysis mentioned here will be further explained in the *practical guide* section of this article); data pre-whitening, controlling for lag 1 and 7, time series analysis, i.e., assessment of correlations between predictors and 3 outcome variables. Patterns of theoretical variables of behaviour maintenance contributing to the prediction and amount of variability accounted for, differed between participants for weight loss maintenance plan adherence and physical activity. The authors identified theoretical predictors that were the most predictive of physical activity increase and decrease in each person. Identifying which factors show the strongest correlations with assessed outcomes may allow the design of follow-up interventions that relate to the most predictive outcomes, applied at the time when they are the most needed.

Such personalised interventions can be tested using N-of-1 experimental design which involve experimental manipulation to assess the effect of intervention/treatment on a behavioural outcome(s). N-of-1 trials are regarded as the gold standard for generating evidence for individual treatment decisions (Guyatt, Meade, Jaeschke, Cook, & Haynes, 2000) over and above systematic reviews of RCTs. This is because the results from groups of participants are not going to be as relevant to an individual as the results from an RCT where they are the only participant.

Within experimental N-of-1s, there are multiple design types: AB, ABA, ABCBC, varying baselines etc; McDonald et al. (2017) provides a detailed overview of different design types and described examples of each type. Arguably the most sophisticated N-of-1 design is an N-of-1 RCT, i.e., a crossover experiment conducted with a single participant

who acts as their own control. N-of-1 RCTs usually provide repeated and randomly allocated periods of treatment to participants with sufficient frequency to minimise any chance of confounding influences on the outcome. Due to carry over effects, not all behaviour change techniques (BCTs) and interventions can be tested in N-of-1 RCTs. BCTs that are particularly suitable are the ones that are time specific, e.g., setting plans for a given day, as compared to setting long term plans (Kwasnicka et al., 2019). For example, a recent factorial N-of-1 RCT evaluated and compared the effectiveness of different BCTs to increase physical activity in older people comparing goal-setting with self-monitoring for a given day (Nyman, Goodwin, Kwasnicka, & Callaway, 2016). Eight adults age 60–87 were randomised to a 2 (goal- setting vs. active control) × 2 (self-monitoring vs. active control) factorial RCT over 62 days; with 31 days of data for each condition per participant (on some days participants received both interventions, on some days no interventions and on other days only one out of the two interventions). The time series data were prewhitened (where significant autocorrelations were identified) and analysed for each single case using linear regressions. The results showed that compared to control days, goal-setting increased walking in four out of eight participants and self-monitoring increased walking in seven out of eight participants, two participants had a significant but small linear decrease in walking over time.

As demonstrated by Nyman et al. (2016), idiographic methods can be applied to test which BCTs are most suitable for which individuals. Recent technology developments such as mobile devices allow us to deliver interventions and collect relevant data in an automated way, allowing us to evaluate and compare interventions with each other and to control arms. Different elements of the intervention can be separated and tested on different days and the effectiveness of each can be assessed and compared. The same principles of intervention design can be used to separate and compare different intensities (e.g., short messages versus long elaborated stories), different modes of intervention provision (e.g., text versus video)

and different elements of the interventions (not only separating different BCTs but also comparing different forms of the same BCT). While selection of variables of interest or interventions requires care when planning an N-of-1 study, it is of high importance that an appropriate design and method of data analysis is applied.

Practical guide to N-of-1 design and analysis

Several issues need to be considered when designing N-of-1 study, namely variability of predictors and outcomes, most suitable assessment frequencies and most appropriate analysis methods. McDonald et al. (2017) reported in their systematic review of behavioural N-of-1 studies that out of 39 studies, only 11 studies used statistical methods, 21 used visual analysis and 7 used descriptive statistics. It has been noted that statistical analysis in N-of-1 studies have historically lacked rigour and reporting transparency (Shaffer et al., 2018). Tate et al. (2013) proposed a quality rating scale for single-case experimental designs and N-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. The most current guidelines for best practice in N-of-1 reporting are: single-case reporting guideline in behavioural interventions (SCRIBE) (Tate et al., 2016) and the CONSORT extension for reporting N-of-1 trials (CENT) Statement (Vohra et al., 2015). Here we discuss issues relevant to design and analysis in a form of a practical step-by-step guide to N-of-1 study design. We are also providing a dataset that interested readers can use to practice the suggested analysis methods (https://osf.io/9psf2/). While we talk the reader through analysis using SPSS in this paper, we also provide an R script to carry out the same approach.

Variability of predictor and outcome variables

First of the issues to consider when designing N-of-1 study is variability of the included measures. The researchers can only assess behavioural predictions and outcomes of interest if the predictors and outcomes vary over time. This is usually the case for objectively monitored physical activity (e.g., assessed with accelerometery) but it may not be the case for

bouts of exercise, e.g., if assessed person does not engage in any exercise. Equally if testing the relationship between self-reported self-efficacy and physical activity in the individual who always has high levels of self-efficacy to be active (e.g., 10 out of 10) then the predictor (self-efficacy) will not predict the outcome (physical activity) as there would be insufficient variability in the predictor. It might be that for some variables that predict physical activity, variation in these occur over a longer timeframe and so a daily repeated measure for two months may not be a long enough timeframe to identify relevant variation. Statistical approaches for estimating intra-individual variability include intra-individual standard deviation, coefficient of variation and mean successive squared differences (Barbot, & Perchec, 2015). In order to capture variability in predictors and outcomes, the researchers need to make decisions about the frequency of the assessments.

Frequency of the repeated assessments

The most common approach used for data collection is through EMA (Stone & Shiffman, 1994). Frequency of EMA is influenced by how data is requested; EMA assessments can be researcher prompted (known as signal-contingent, e.g., by a daily text message sent to a participant's phone with a link to an online survey) or they can be participant initiated (known as event-contingent, e.g., every time you finish a gym session log it on your mobile phone app). In an N-of-1 study, frequency of the predictor variables will need to be mapped to the frequency of outcome variables so the relationship between predictor and outcome can be assessed, e.g., through time series cross-correlations. For example, if looking at the impact of motivation to exercise (assessed daily) on exercise bouts, then conventionally we will look at the scores for motivation on the given day (e.g., Likert type scale 0-5) as compared to the number of exercise bouts on the same day. We can also look at temporal predictions and time lags. A time lag refers to an interval of time between two related assessed variables (as an antecedent and its effect). Time lag 0 means correlation

between the predictor and outcome variable at the same time (e.g., on the same day), lag 1 means that one variable precedes the other one by the unit of time (e.g., stress level yesterday has an impact on exercise level today if the unit of time is one day). In terms of the number of data points needed for a viable statistical analysis in an N-of-1 study, there are no rules that will be appropriate for all studies. As with all quantitative studies, the number of data points depends on the statistical power required to identify a hypothesised relationship. However, additional parameters not usually encountered in between-subjects designs need to be estimated when undertaking power analysis for an N-of-1 study, such as effect heterogeneity (Kwasnicka et al., 2019). In order to assess variability in the predictor and outcome variables and to decide on the frequency of N-of-1 assessments, it is best to pilot the procedures before commencing an N-of-1 study.

Example dataset used to illustrate analysis methods - data structure

To illustrate analysis methods, we use a dataset of 24 individuals who collected data on themselves for 28 days as part of an N-of-1 special interest project. They were asked to provide daily responses to questions regarding several health behaviours including their fruit and vegetable consumption (number of portions of fruit and vegetables eaten each day), alcohol consumption (number of standard units consumed each day), numbers of steps as objectively measured with a pedometer, self-reported number of minutes of any other physical activity that could not be quantified as steps, levels of stress and happiness each day (measured on a 0-10 scale, 0 – low, 10-high) and perceived sleep quality (adapted from the Pittsburgh Sleep Quality Index, Buysse, Reynolds III, Monk, Berman, & Kupfer, 1989). As data was time stamped and data collection confined to a specific geographical area, we could also check meteorological data for the given day and see if variables such as air temperature, humidity, rain, wind had any impact on the daily measured outcomes, e.g., is the given person happier on the sunnier days. Meteorological data was added to the dataset for each

day participants collected data. The dataset can be inspected to gain a sense of what data was collected – each day in the dataset is represented by one row and a variable indicating the day sequence (from 1-28) was created for the rows for each participant.

Dealing with missing data

There are different approaches to dealing with N-of-1 missing data, though these are largely the same as dealing with missing data from any dataset (Kwasnicka et al., 2019). The first step is to visually inspect data and also (if available) to look into any additional qualitative data gathered that may explain missing data. Looking at time series plots, we are assessing if there are any obvious patterns of missing data for each person, e.g., prolonged periods of continuous missing data at the end of the data collection period may be explained in terms of participant attrition due to repetitive study procedures. If this is established as a reason for missing data the dataset could be shortened (Kwasnicka et al., 2017), although this could introduce bias. If the distribution of missing data appears to be random, then researchers may consider imputing missing data, e.g., using appropriate bootstrapping techniques to impute missing values. For instance, Amelia II (www.gking.harvard.edu/amelia) can be used for N-of-1 datasets, which performs multiple imputation and has been shown to reduce bias and increase efficiency as compared to listwise deletion (Honaker, King, & Blackwell, 2011). In the example dataset, as there was very little missing data, we imputed data where missing using a simple averaging approach using the adjacent data points either side of the missing data. However, there are limits to how much missing data can be dealt with by simple averaging; usually no more than 5-10% of randomly distributed data would be adequate.

Graphical representations of N-of-1 data

First, to gain a better understanding of data patterns and data distribution, you can start with plotting your data over time. In SPSS you can plot your data through: *Analyse* –

Forecasting – Sequence charts; then selecting the relevant variable, e.g., steps, entering time or date into axis label. Syntax for SPSS and R are provided in the OSF project (https://osf.io/9psf2/). Figure 1 shows plots for two participants, where we plotted their perceived happiness over the study period. In this Figure, we demonstrate how participants can often vary on repeated measures; participant 4 shows substantial variability in their perceived happiness whereas participant 5 shows almost no variability on this construct. Where there is little variation in either predictor or outcome variable, it is unlikely that an association between predictor and outcome variable can be identified.

Plots can also provide a sense of whether there might be a temporal trend in the data. Repeated or cyclical changes (seasonality) can be observed, such as differences in activity levels at the weekend versus the week. Longer term trends without a cyclical nature within the data can be interpreted as non-stationary data, where the mean, variance and autocorrelation structure changes over time, such as changes in physical activity due to seasonal transition e.g. winter to spring. Although time trends would need to be explored statistically for confirmation. Plots can be produced where two or more variables are plotted simultaneously – this can identify potential associations between different variables.

Please insert Figure 1 here

However, it is important that hypotheses about the potential association between variables are planned before exploration of data if undertaking confirmatory analyses, and, if not, it is explained what led to the hypotheses if generated after data exploration and that the analyses are exploratory. An overview of visual analysis in single case experimental design studies and a step-by-step guide for conducting a visual analysis of graphed data is provided by Lane and Gast (2014).

Autocorrelation of data points

Autocorrelation may be present in time series data-sets, where a measurement point is correlated with previous measurement points because they are collected relatively close in time. For example, your mood yesterday may predict your mood today. Statistical methods exist to remove (Naughton & Johnston, 2014) and to model (Vieira, McDonald, Araújo-Soares, Sniehotta, & Henderson, 2017) autocorrelation in idiographic data sets. Recent N-of-1 physical activity studies have used a *prewhitening method* to remove autocorrelation when data points were autocorrelated (Hobbs, Dixon, Johnston, & Howie, 2013; Kwasnicka et al., 2017) so each participant measurement point could be treated as an independent data point. Approaches which model and incorporate autocorrelation, e.g., Auto-Regressive Integrated Moving Average (ARIMA, Box & Pierce, 1970), ARIMAX (dynamic regression) or Generalised Additive Mixed Models (GAMM) are alternative methods which can model autocorrelation. To practise dynamic regression modelling, Vieira et al. (2017) provides an example dataset with R syntax: https://zenodo.org/record/580028#.W_os2ugza70. To practice prewhitening we have made available our example dataset (https://osf.io/9psf2/) with accompanying SPSS and R syntax.

The prewhitening process below works for single participants. Prewhitening essentially removes from a time-series any correlation between a data point and a specific lagged data point for the same variable (e.g., lag 1 is the previous day, lag 2 is two days previous etc.). Typically, the outcome variable would be examined and have autocorrelation removed. To assess if a specific variable demonstrates autocorrelations in SPSS go to *Analyse – Forecasting – Autocorrelations*, selecting the variables that you want to check for autocorrelation, e.g., happiness, stress, steps. In the SPSS display window select Autocorrelations and Partial autocorrelations and inspect the graphs. For a first order (one time point) autocorrelation check if the autocorrelation graph Lag1 is beyond the confidence interval line in the graph. If so, this indicates a significant association between these data

points – a *p* value for this association is provided in the accompanying table. Partial autocorrelation graphs adjust for shorter lags, e.g., an autocorrelation value at lag 2 would indicate an association when lag 1 is controlled for (though no significance test is provided by SPSS). In other words, partial autocorrelation graphs essentially adjust for lower-order lags to help identify where an autocorrelation occurs (1st order, 2nd order etc.). For example, an autocorrelation value at lag 2 would indicate an association when lag 1 is controlled for (though no significance test is provided by SPSS). If autocorrelation appears not to be present, it may not be necessary to adjust the outcome variable by itself at an earlier time point. However, there may be insufficient power to identify it so a conservative approach is to adjust for it if there is indication of autocorrelation but it does not reach statistical significance.

Prewhitening method

To prewhiten a variable to remove autocorrelation, you need to first create a lagged variable for the corresponding autocorrelation lag. Go to: *Transform – Create time series – select Function – Lag* – 1 (for 1st order autocorrelation) and select/drag across variable of interest and press OK. This creates a lagged variable, i.e., data moved by the lag specified (e.g., one time point for a lag 1). If you create a lagged variable when you have more than one participant in the dataset, the final data point for a participant will be lagged (i.e., shifted down one row) and will replace the first value for the next participant. To avoid this, either create a lagged variable for each participant separately or use the *Split file* command before using the *Shift values* command under *Transform*. ~~You should create a lagged variable only when you have a single participant in the dataset, otherwise the final data point for one participant will be lagged (i.e., shifted down one row) and replace the first value for the next participant.~~ To then create a prewhitened variable, go to: *Analyse – Regression – Linear* and in the dialog box select the dependent variable (DV) as your original variable before it was

lagged and your independent variable (IV) as the lagged version. You then need to select *Save* and tick the *Save unstandardized residuals* box and run the analysis. This newly created residuals variable is the new prewhitened variable. If you wanted to check if this process has removed any autocorrelation you can re-run the autocorrelation charts with the prewhitened variable, following the instructions from the graphical representations of N-of-1 data section above.

The prewhitened variable can be used as the DV in routine analyses (e.g., regression). We have undertaken the sequence described above for participant 7 in the training dataset and annotated the appropriate syntax (see OSF project) to investigate the association between the daily number of steps taken (independent variable) and happiness within the last day. Table 1 and Figure 2 show the autocorrelation table and plot demonstrating a significant 1st order autocorrelation for happiness.

Please insert Table 1 and Figure 2 here

When we run a regression to see if the number of steps predicts (prewhitened) happiness, we find a significant association at time 0. In other words, the number of steps taken for a given day predicts happiness for that day (standardised beta 0.60, p=0.001). However, this analysis only tells us if these variables are associated at the same time period, it does not test whether physical activity (number of steps as a proxy) might prospectively predict happiness or vice versa. To determine this, we would need to lag the IV.

Taking into account temporality to identify potential antecedents

To assess if an IV prospectively predicts the DV, we can simply create a lagged version of the IV using the same process as above (e.g., lag 1 if wanting to assess one measurement point back in time as a predictor, lag 2 for two measurement points back and so on). If a predictor analysis is then undertaken with the lagged variable, you are assessing if the IV from one measurement point back (e.g. yesterday) predicts the DV at time 0 (e.g.,

today). If it is likely that the IV is autocorrelated, then it is recommended to include the IV at lag 0 in the model or alternatively consider prewhitening the IV. For our example, we then investigated if yesterday's physical activity (steps) predicts today's happiness for participant 7. Steps showed no evidence of autocorrelation and so did not need adjusting for in this analysis. The analysis indicated no association between these time-bounded variables (standardised beta -0.06, p=0.75). SPSS can produce a cross-correlation chart where the association between different lags for two variables of interest are presented: *Analyse – Forecasting – Cross-correlations*. This represents an exploratory analysis and so should ideally be undertaken after any *a priori* hypotheses are generated or tested. The cross-correlation plot (Figure 3) indicates that steps and happiness are only associated for the same day, as only at lag 0 does the bar go over the confidence interval line. However, if the bar at lag 1 or higher reached the confidence interval line, this would indicate that the first variable entered into the cross-correlation (in this case steps) precedes the second (happiness), supporting the first as an antecedent to the other. If the bar at lag -1 or lower reached the confidence interval line, this would support the second variable being the antecedent. Of course, finding two variables associated only at lag 0 does not mean one is not the antecedent of the other. It might be that the frequency of measurement is too far apart to identify the point where a change in one variable precedes a change in the other, if a true causal relationship exists for that individual. The prewhitening offers a simple method to deal with autocorrelation; however, if the effect assessed is a slow change then prewhitening can remove the desired effect from the data. Such slow effects might well be seen as non-stationarity in the data and can be dealt with by fitting appropriate regression lines to the data before dealing with auto-correlation (Huitema & Mckean, 2000). Prewhitening therefore requires the assumption of stationarity. If there is evidence of non-stationarity, then prewhitening is unlikely to be suitable for the reason above; removing a genuine effect

through autocorrelation adjustment. The prewhitened variable can be used as the DV in routine analyses (e.g., regression, multivariate analyses etc); however, further more advanced methods exist to model a network of multivariate time series (Yang et al., 2018).

In this practical guide we have elaborated on the methods that align predictors and outcomes to assess the relationship between them. However; dynamic systems models can be used to capitalise on the rich information that also occurs between dynamic measurement points (i.e., continuous physical activity data) and self-reported data, which have been applied to physical activity phenomena (Ashour et al., 2016; Phatak et al., 2018; Riley et al., 2015; Spruijt-Metz et al., 2015; Timms, Martín, Rivera, Hekler, & Riley, 2014).

Please insert Figure 3 here

Aggregating data – multilevel modelling

For several reasons it can be appropriate to combine N-of-1 datasets into an aggregated analysis, such as if an association is expected to be similar between participants or when wanting to explore what factors may explain differences between individuals in associations. When this is done, it is often to examine whether the direction and strength of associations are similar between participants, e.g., as in the aforementioned stress and exercise study (Burg et al., 2017). A common method of undertaking an aggregated analysis of N-of-1 datasets is by using multilevel modelling/mixed models. In simple terms, the DV, which is the repeated measure (e.g., happiness measured every day), is a level 1 variable and any IV(s) or control variables (e.g., steps, hours of daily sunshine) that are also repeated measures at the same frequency as the DV are entered as level 1 factors (fixed effects). ~~Any factors relevant to the grouping level of units (e.g., gender of participants) are entered at level 2 (random effects), with any further grouping being entered at level 3 and so on.~~ The repeated measure at level 1 will be nested within level 2 factors (random effects), which are invariant

characteristics at the grouping level of units (e.g., gender of participants). Further grouping (e.g., hospitals where participants work) would be entered at level 3 and so on. This example analysis could be assessing whether the association between physical activity (steps) and happiness differs between men and women, when adjusting for how sunny the weather is. Autocorrelation of the DV can be incorporated within multilevel models, although autocorrelation is handled differently compared to prewhitening. We undertook multilevel modelling to explore whether there was an association between daily steps and happiness across all participants in our training dataset (we did not examine whether this differed by gender due to a very unequal gender balance). A basic mixed model was constructed in SPSS for the purposes of demonstration (see syntax document in OSF project). This basic model indicated a statistically significant though small fixed effect for steps (unstandardized beta 0.00002 p=0.02) on happiness across participants; this means that for every one step increase, happiness increases by 0.00002 across participants or for each 1,000 steps, happiness increases by 0.02. However, there was some variation in direction and strength of this association between participants, which would be worthy of further investigation, i.e., through random effects analysis.

Applicability and scalability of N-of-1 design

Several challenges exist with N-of-1 design and with data analysis; also, several questions arise about applicability, ecological validity and potential application of person specific design: "*How is it useful and how is it scalable?*". Person specific approaches can employ EMA to gather data regarding cognitions, behaviour predictors and outcomes. EMA can be applied in different forms, e.g., using increasingly less popular pen and paper methods (e.g., in a diary form), using surveys delivered to the device of choice, e.g., mobile phone, tablet, computer, smart watch, hand-hold devices and via text message, text message embedded link, app, email etc. Data can be harvested automatically from the mobile phone,

from wearables (e.g., geo location), from geo-spatial sensors (e.g., via RFID technology) etc. Data can be also captured by the participant with cameras or via voice recordings. Several novel data capture technologies and methods make frequent assessment feasible and scalability of the design is increasing through the means of new technology development.

In the area of sports and exercise psychology investigation using N-of-1 methods has the potential to be applied at scale with the employment of new technologies and sensors, e.g., Fitbit devices allow gathering physical activity data with good long-term compliance (Burg et al., 2017; Kwasnicka et al., 2017). Most mobile phone devices have built-in sensors which allow us capturing longitudinal activity data and geo-location data unobtrusively (Bort-Roig, Gilson, Puig-Ribera, Contreras, & Trost, 2014). Using mobile phone devices in N-of-1 studies to gather outcome data is cost-effective and usually also demonstrates high compliance, although gathering physical activity/steps data using mobile phone sensors has variable accuracy (Case, Burwick, Volpp, & Patel, 2015). Specific sensors (placed on the individual or placed in the environment) allow us to capture data about persons movement – intensity, accuracy, estimates of energy expenditure.

Employing an idiographic approach, we can assess trajectories of change within individuals, for instance instead of assessing groups of athletes, we can use longitudinal assessment to gather data regarding one particular athlete – including his/her performance predictors and outcomes, e.g., speed and accuracy measures. We can then design interventions which are person specific and highly tailored to the athlete based on previously gathered data, e.g., knowing that person trains best when they feel intrinsically motivated, supported by colleagues and happy on that day, we can advise the coach to tap into those variables during training. Other athletes may train best when their confidence is high, when they feel relaxed and rested, then the advice given to these athletes should mainly focus around increasing confidence, improving sleep hygiene and emphasising rest breaks. Using

N-of-1 methods, we can make the most personalised recommendations for each athlete to improve their performance (Guyatt et al., 2000).

N-of-1 methodology also allows testing and comparing different interventions in one participant or in one sports team, over time. Different interventions can be randomly allocated to different time periods and their effectiveness compared in one measurement unit, e.g., one athlete, one team. We can also test behavioural theories within individuals and measurement units, athletes, teams, football clubs etc. rather than in groups of individuals, as conventionally done in observational group studies and RCTs. Employing idiographic methods, we can explore trajectories of change and test theories in one measurement unit to conduct precision studies and to design truly personalised interventions.

N-of-1 methodology also has some clear limitations, such as high intensity measurement, low scalability unless technology is used, difficulty in generalising findings to a larger population than that studied and resource intensive analysis. N-of-1 requires a high number of assessments on the same participants that can often lead to high participant burden or self-selection bias, i.e., only highly motivated individuals take part in N-of-1 studies. Finally, if the research questions being investigated are seeking average relationships in the population assessed then a nomothetic approach is more applicable.

Conclusion

Knowledge of how to employ N-of-1 methods enables researchers to capitalise on recent technology developments to design personalised behavioural studies and interventions. This can help identify patterns of behaviour, inter-person differences in those patterns and provides a tool for identifying potentially important antecedents of behaviour. Using unobtrusive data capture from wearables and smartphone sensors makes it easier to collect longitudinal N-of-1 data, combined with self-report EMA data, makes it possible to design person centred studies and interventions. We are at an opportune time to expand our use of

idiographic designs to better understand health behaviour and to deliver personalised interventions.

References

Ashour, M., Bekiroglu, K., Yang, C.-H., Lagoa, C., Conroy, D., Smyth, J., & Lanza, S. (2016). *On the mathematical modeling of the effect of treatment on human physical activity.* Paper presented at the 2016 IEEE Conference on Control Applications (CCA).

Barbot, B., & Perchec, C. (2015). New Directions for the Study of Within□Individual Variability in Development: The Power of "N= 1". *New directions for child and adolescent development, 2015*(147), 57-67.

Bartholomew, L. K., Parcel, G. S., & Kok, G. (1998). Intervention mapping: a process for developing theory and evidence-based health education programs. *Health Education & Behavior, 25*(5), 545-563.

Bolger, N., Stadler, G., & Laurenceau, J.-P. (2012). Power analysis for intensive longitudinal studies.

Bort-Roig, J., Gilson, N. D., Puig-Ribera, A., Contreras, R. S., & Trost, S. G. (2014). Measuring and influencing physical activity with smartphone technology: a systematic review. *Sports medicine, 44*(5), 671-686.

Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association, 65*(332), 1509-1526.

Burg, M. M., Schwartz, J. E., Kronish, I. M., Diaz, K. M., Alcantara, C., Duer-Hefele, J., & Davidson, K. W. (2017). Does stress result in you exercising less? Or does exercising result in you being less stressed? Or is it both? Testing the bi-directional stress-

exercise association at the group and person (N of 1) level. *Annals of Behavioral Medicine, 51*(6), 799-809.

Buysse, D. J., Reynolds III, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry research, 28*(2), 193-213.

Case, M. A., Burwick, H. A., Volpp, K. G., & Patel, M. S. (2015). Accuracy of smartphone applications and wearable devices for tracking physical activity data. *Jama, 313*(6), 625-626.

Dijkstra, A., & De Vries, H. (1999). The development of computer-generated tailored interventions. *Patient education and counseling, 36*(2), 193-203.

Guyatt, G. H., Meade, M. O., Jaeschke, R. Z., Cook, D. J., & Haynes, R. B. (2000). Practitioners of evidence based care: Not all clinicians need to appraise evidence from scratch but all need some skills: British Medical Journal Publishing Group.

Hobbs, N., Dixon, D., Johnston, M., & Howie, K. (2013). Can the theory of planned behaviour predict the physical activity behaviour of individuals? *Psychology & health, 28*(3), 234-249.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of statistical software, 45*(7), 1-47.

Huitema, B. E., & Mckean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38-58.

Johnston, D. W., & Johnston, M. (2013). Useful theories should apply to individuals. *British journal of health psychology, 18*(3), 469-473.

Kwasnicka, D., Dombrowski, S. U., White, M., & Sniehotta, F. F. (2017). N-of-1 study of weight loss maintenance assessing predictors of physical activity, adherence to weight loss plan and weight change. *Psychology & health, 32*(6), 686-708.

Kwasnicka, D., Inauen, J., Nieuwenboom, W., Nurmi, J., Schneider, A., Short, C. E., . . . Haukkala, A. (2019). Challenges and solutions for N-of-1 design studies in health psychology. *Health psychology review*(just-accepted), 1-36.

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological rehabilitation, 24*(3-4), 445-463.

Lilienfeld, S. O., McKay, D., & Hollon, S. D. (2018). Why randomised controlled trials of psychological treatments are still essential. *The Lancet Psychiatry*.

McDonald, S., Quinn, F., Vieira, R., O'Brien, N., White, M., Johnston, D. W., & Sniehotta, F. F. (2017). The state of the art and future opportunities for using longitudinal n-of-1 methods in health behaviour research: a systematic literature overview. *Health Psychology Review, 11*(4), 307-323.

Naughton, F., Jamison, J., Boase, S., Sloan, M., Gilbert, H., Prevost, A. T., . . . Evans, R. (2014). Randomized controlled trial to assess the short-term effectiveness of tailored web-and text-based facilitation of smoking cessation in primary care (i Q uit in P ractice). *Addiction, 109*(7), 1184-1193.

Naughton, F., & Johnston, D. (2014). A starter kit for undertaking n-of-1 trials. *European Health Psychologist, 16*(5), 196-205.

Nielsen, L., Riddle, M., King, J. W., Aklin, W. M., Chen, W., Clark, D., . . . Ferrer, R. (2018). The NIH Science of Behavior Change Program: Transforming the science through a focus on mechanisms of change. *Behaviour research and therapy, 101*, 3-11.

Nyman, S. R., Goodwin, K., Kwasnicka, D., & Callaway, A. (2016). Increasing walking among older people: A test of behaviour change techniques using factorial randomised N-of-1 trials. *Psychology & health, 31*(3), 313-330.

Phatak, S. S., Freigoun, M. T., Martín, C. A., Rivera, D. E., Korinek, E. V., Adams, M. A., . . . Hekler, E. B. (2018). Modeling individual differences: A case study of the application of system identification for personalizing a physical activity intervention. *Journal of biomedical informatics, 79*, 82-97.

Riley, W. T., Martin, C. A., Rivera, D. E., Hekler, E. B., Adams, M. A., Buman, M. P., . . . King, A. C. (2015). Development of a dynamic computational model of social cognitive theory. *Translational behavioral medicine, 6*(4), 483-495.

Shaffer, J. A., Kronish, I. M., Falzon, L., Cheung, Y. K., & Davidson, K. W. (2018). N-of-1 Randomized Intervention Trials in Health Psychology: A Systematic Review and Methodology Critique. *Annals of Behavioral Medicine*.

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

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of behavioral medicine.*

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., . . . McDonald, S. (2016). The single-case reporting guideline in behavioural interventions (SCRIBE) 2016 statement. *Aphasiology, 30*(7), 862-876.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological rehabilitation, 23*(5), 619-638.

Timms, K. P., Martín, C. A., Rivera, D. E., Hekler, E. B., & Riley, W. (2014). *Leveraging intensive longitudinal data to better understand health behaviors.* Paper presented at

the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Vieira, R., McDonald, S., Araújo-Soares, V., Sniehotta, F. F., & Henderson, R. (2017). Dynamic modelling of n-of-1 data: powerful and flexible data analytics applied to individualised studies. *Health Psychology Review, 11*(3), 222-234.

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Tate, R., . . . Guyatt, G. (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. *bmj, 350*, h1738.

Yang, X., Ram, N., Gest, S. D., Lydon-Staley, D. M., Conroy, D. E., Pincus, A. L., & Molenaar, P. (2018). Socioemotional Dynamics of Emotion Regulation and Depressive Symptoms: A Person-Specific Network Approach. *Complexity, 2018*.

Yardley, L., Morrison, L., Bradbury, K., & Muller, I. (2015). The person-based approach to intervention development: application to digital health-related behavior change interventions. *Journal of medical Internet research, 17*(1).

638 **Highlights**

639 - N-of-1 methods test predictions, outcomes and interventions within individuals;

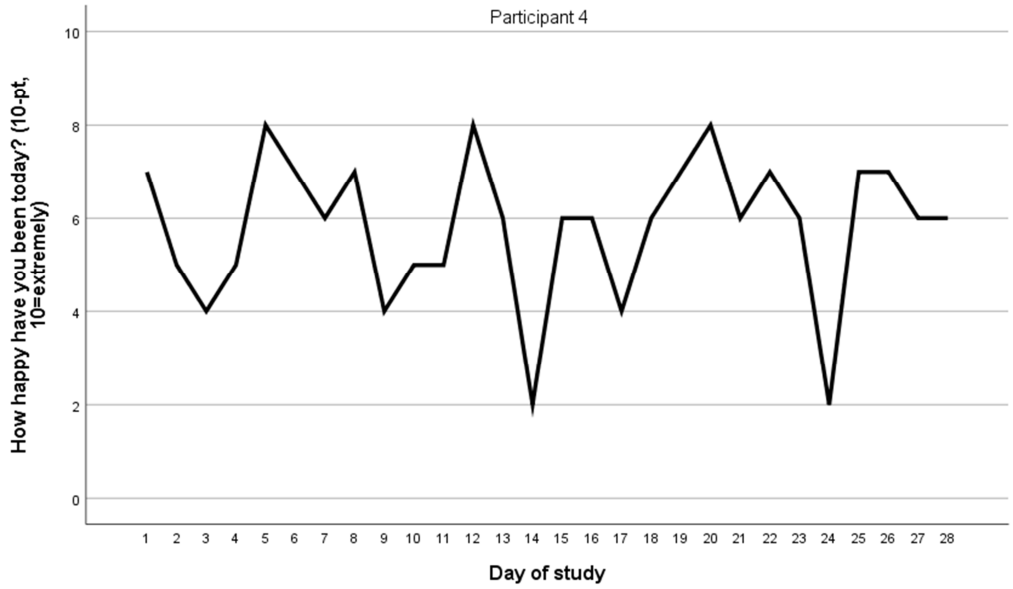
640 - N-of-1 approach has been vastly underutilised in exercise psychology;

641 - This article provides a step by step guide to N-of-1 study design and analysis;

642 - EMA, sensors and wearables can be successfully applied in N-of-1 research;

643 - Recent technology developments make it possible to apply N-of-1 approach at scale.

644



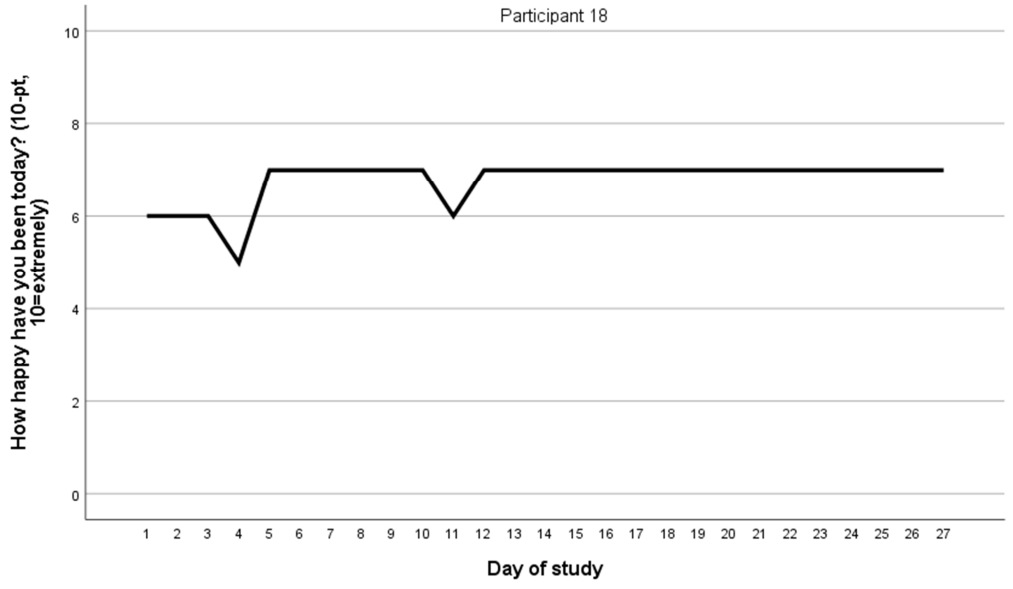


Figure 1: Two plots presenting two participants' happiness rating over the 28-day study period

Table 1: Autocorrelation in 'happiness' variable

Autocorrelations[a]

Series: How happy have you been today? (10-pt, 10=extremely)

Lag	Autocorrelation	Std. Error[b]	Box-Ljung Statistic		
			Value	df	Sig.[c]
1	.379	.179	4.462	1	.035
2	-.184	.176	5.550	2	.062
3	-.218	.173	7.152	3	.067
4	.051	.169	7.242	4	.124
5	.096	.165	7.578	5	.181
6	-.023	.162	7.599	6	.269
7	-.104	.158	8.035	7	.330
8	-.034	.154	8.083	8	.425
9	-.009	.150	8.087	9	.525
10	-.154	.146	9.194	10	.514
11	-.079	.142	9.503	11	.576
12	-.004	.138	9.504	12	.659
13	-.005	.134	9.505	13	.734
14	-.032	.129	9.566	14	.793
15	-.041	.124	9.675	15	.840
16	-.076	.120	10.079	16	.862

Notes. a. Participant ID = 7; b. The underlying process assumed is independence (white noise); c. Based on the asymptotic chi-square approximation.

659

660

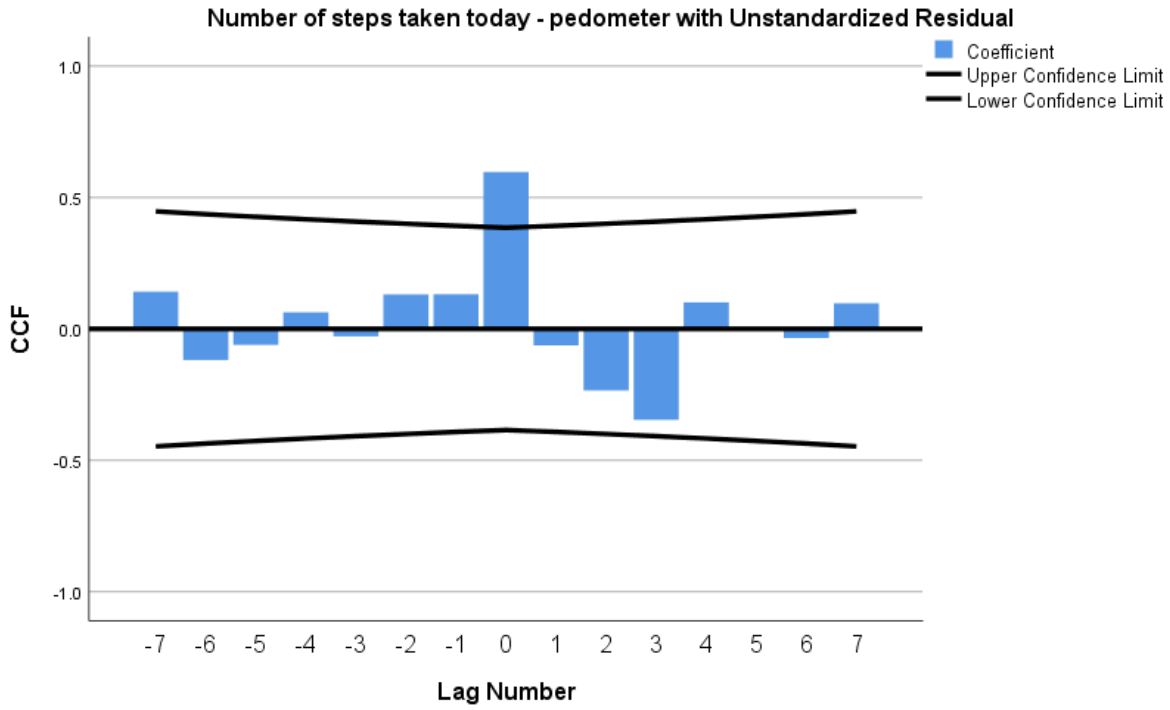
661

662 Figure 2: Autocorrelation charts for participant 7 demonstrating a 1st order autocorrelation of

663 a happiness measure

664

665



666

667

668

669 Figure 3: Cross-correlation plot for participant 7 indicating steps and happiness are only

670 associated cross-sectionally (when assessed on the same day)

Conflict of interest statement for "N-of-1 methods: A practical guide to exploring individual trajectories and antecedents of behaviour"

The authors have no conflict of interest to disclose.

Kind regards,

Dr Dominika Kwasnicka and Dr Felix Naughton