**Title:** Challenges and solutions for N-of-1 design studies in health psychology.

.

Dominika Kwasnicka, Jennifer Inauen, Wim Nieuwenboom, Johanna Nurmi, Annegret Schneider, Camille E Short, Tessa Dekkers, A. Jess Williams, Walter Bierbauer, Ari Haukkala, Federica Picariello, Felix Naughton

Author Note

Dominika Kwasnicka, School of Psychology, Curtin University, Kent Street, Perth, 6102, Western Australia, +61 8 9266 2559, dominika.kwasnicka@curtin.edu.au; Jennifer Inauen, Department of Environmental Social Sciences, Environmental and Health Psychology, Eawag – Swiss Federal Institute of Aquatic Science & Technology, Ueberlandstrasse 133, 8600 Duebendorf, Switzerland, +41 58 765 5780, jennifer.inauen@eawag.ch; Wim Nieuwenboom, University of Applied Sciences Northwestern Switzerland, School of Social Work, Institute for Social Work and Health, Riggenbachstrasse 16, CH-4600 Olten, +41 62 957 21 09, wim.nieuwenboom@fhnw.ch; Johanna Nurmi, (1) Faculty of Social Sciences, University of Helsinki, Finland; (2) Behavioural Science Group, Institute of Public Health, Forvie Site, Cambridge CB2 0SR, University of Cambridge, United Kingdom, +358 504484297, johanna.nurmi@helsinki.fi; Annegret Schneider, Department of Clinical, Educational and Health Psychology, University College London, Gower St, London, WC1E 6BT, United Kingdom, +44 (0) 20 7679 25438, a.schneider@ucl.ac.uk; Camille E Short, The Freemasons Foundation Centre for Men's Health, School of Medicine, University of Adelaide, Adelaide, South Australia, Australia, +61 8 83130532, camille.short@adelaide.edu.au; Tessa Dekkers, Faculty of Industrial Design Engineering, Delft University of Technology, Landbergstraat 15, 2628 CE Delft, the Netherlands, +31 (0) 15 27 84 640, t.dekkers@tudelft.nl; A. Jess Williams, Institute for Mental Health, School of Psychology, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK., AXW847@student.bham.ac.uk; Walter Bierbauer, Department of

Psychology, Applied Social and Health Psychology, University of Zurich, Binzmuehlestrasse 14, 8050 Zurich, Switzerland, +41 44 635 72 58; walter.bierbauer@psychologie.uzh.ch; Ari Haukkala, Faculty of Social Sciences, University of Helsinki, Finland, +358 (0) 504480951, ari.haukkala@helsinki.fi; Federica Picariello, Health Psychology Section, Psychology Department, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, 5th floor Bermondsey Wing, Guy's Campus, London, SE1 9RT, United Kingdom, +44 (0) 20 7188 9324, federica.picariello@kcl.ac.uk; Felix Naughton, School of Health Sciences, University of East Anglia, Norwich NR4 7UL, United Kingdom, + 44 (0) 1603 593459, f.naughton@uea.ac.uk.

This conceptual review was written as a result of an N-of-1 design workshop prior to 31st European Health Psychology Society (EHPS) Conference in Padua, Italy, 2017. We would like to acknowledge the EHPS for the opportunity to meet and discuss N-of-1 design. We would like to thank all meeting participants who contributed to the discussions that instigated this paper. The authors would also like to acknowledge the helpful statistical advice from Dr George Savva and Dr Kathryn Richardson.

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.

**Abstract**

Theories of behaviour change and health behaviour change interventions are most often evaluated in between-person designs. However, behaviour change theories apply to individuals not groups and behavioural interventions ultimately aim to achieve within-person rather than between-group change. Within-person methodology, such as N-of-1 (also known as single case design), can circumvent this issue, though has multiple design-specific challenges. This paper provides a conceptual review of the challenges and potential solutions for undertaking N-of-1 studies in health psychology. Key challenges identified include participant adherence to within-person protocols, carry-over and slow onset effects, suitability of behaviour change techniques for evaluation in N-of-1 experimental studies, optimal allocation sequencing and blinding, calculating power/sample size, and choosing the most suitable analysis approach. Key solutions include involving users in study design, employing recent technologies for unobtrusive data collection and problem solving by design. Within-person designs share common methodological requirements with conventional between-person designs but require specific methodological considerations. N-of-1 evaluation designs are appropriate for many though not all types of interventions. A greater understanding of patterns of behaviours and factors influencing behaviour change at the within-person level is required to progress health psychology into a precision science.

See Supplementary Material 1 for video abstract.

*Keywords:* N-of-1, single case study, within-person design, idiographic design

# Introduction

N-of-1 studies test predictions within a single object of investigation based on repeated measurement of variables over time. N most often refers to an individual but it can also be defined as one cluster or unit, e.g., a family, hospital, organisation. For N-of-1 designs the power of the study is determined by the number of repeated observations. Therefore, it is possible to satisfy study objectives with just one individual or unit of investigation (Duan, Kravitz, & Schmid, 2013).

The N-of-1 design is a recommended method for testing behavioural theory within individuals through repeated measures (Craig et al., 2008) and is the optimal design for examining within-person variability in cognitions and behavioural outcomes (Johnston & Johnston, 2013). N-of-1 studies have been successfully used in various settings to study a range of health behaviours, including treatment adherence, physical activity, drug/alcohol use, sleep, smoking and eating behaviour (McDonald et al., 2017). For example, Smith et al. (2017) found in six N-of-1 investigations that no individuals had the same pattern of associations between Social Cognitive Theory constructs and their physical activity, and the theory derived constructs demonstrated bi-directional relationships with activity. Within-person assessments of cognitions underlying health-related behaviours are increasingly common in behavioural science and due to recent technology development (Versluis et al., 2016) changes in predictor variables and associated outcome variables can be tested in real time or near real time and in the context of daily life. For an extensive summary of examples of how N-of-1 designs have been applied to health psychology research questions, see McDonald et al.'s (2017) review.

N-of-1 approaches allow health psychologists to develop and conduct personalised behaviour change studies. Different conditions or health interventions can be delivered in a fixed or random order for each participant to investigate which intervention is the most efficacious for each participant. The ‘traditional’ between-group randomised control trial (RCT) only ever provides an estimate of effect at the group level, neglecting intra-individual differences. Importantly, effects observed at the intra-individual level can differ from those found at the between-participant level (Inauen, Shrout, Bolger, Stadler, & Scholz, 2016). Using an N-of-1 design overcomes the issue of effects heterogeneity in nomothetic (between-individual) designs. Furthermore, N-of-1 can more easily enable the modelling of temporal changes. In terms of making individual treatment decisions, N-of-1 studies, namely N-of-1 RCTs, are regarded at the top of the evidence hierarchy (Guyatt et al., 2000) over and above systematic reviews of RCTs.

Unlike health psychology, other disciplines have a relatively long tradition of using within-person designs, providing indications of the type of questions this approach could help to answer. In education research, repeated measures on the level of schools, classrooms or individual learners contributed to the development of learning theories and individualised learning support and have been recommended to document evidence-based best practice in teaching, in particular in the field of special education (Horner et al., 2005; Kennedy, 2005; Moeller, Dattilo, & Rusch, 2015). Other areas such as experimental economics, investigating individuals’ responses to diverse choices or their willingness to pay or accept payments in different scenarios, often used within-individual designs as such findings reflect more closely real-world scenarios where people encounter a series of changing conditions over time (Charness, Gneezy, & Kuhn, 2012; Hogarth, 2005).

Similarly, in medical research within-individual designs resemble clinical practice where physicians treat their patients based on within-individual considerations (Davidson, Peacock, Kronish, & Edmondson, 2014; Janosky, 2005). N-of-1 studies have also helped advance other psychology sub-disciplines such as clinical, neuro or educational psychology and broad behaviour change research, investigating how a person changes with changing circumstances and exposure to different interventions (Barlow & Hersen, 1984; Hogarth, 2005; Sidman, 1960; Tate, Perdices, McDonald, Togher, & Rosenkoetter, 2014). Those diverse applications of within-individual research designs illustrate a methodological diversity, requiring further discipline-specific definition of key constructs and methods.

To help expand the use of N-of-1 methods in health psychology, the field would benefit from a common terminology. Bolger and Laurenceau (2013) offer *intensive longitudinal methods* as an umbrella term for methods involving sequences of repeated measurements sufficiently frequent to allow characterising a separate change process for each unit of assessment. N-of-1 studies, which are also known as *single-participant*, *within-person* and *single-case study design*, fall under this umbrella term and include observational and experimental multiple cross-over studies comparing two or more treatments within individuals (Duan et al., 2013). N-of-1 is arguably the most commonly used term in health psychology to describe this type of study (McDonald et al., 2017) and so we use it here, understanding it as interchangeable with the aforementioned terms. While a single N-of-1 study seeks to understand idiographic within person changes, researchers might also be interested in aggregating those processes at the between-person level to reach generalisable conclusions. In this case the term *aggregated N-of-1 study* can be used, or also *cumulative N-of-1s*, referring to the same principle.

In health psychology, there is currently no established tradition of N-of-1 studies (Davidson et al., 2014; McDonald et al., 2017), meaning the design has been underused and is often misunderstood in the field. While this idiographic design offers many advantages over more traditional nomothetic approaches, it comes with its own challenges, some of which are particularly pertinent to health psychology investigations. The purpose of this paper is to review the key challenges for undertaking health psychology related N-of-1 research and provide potential solutions for resolving or minimising these and, in doing so, encourage greater confidence in using this design. The key challenges were identified during an N-of-1 design workshop prior to the 31st European Health Psychology Society (EHPS) Conference in Padua, Italy, 2017.

## General challenges and solutions across N-of-1 designs

There are two main types of N-of-1 studies, observational and experimental. The principles of N-of-1 observational studies are usually more basic than principles of experimental N-of-1 studies. The N-of-1 observational study involves repeated measures of behavioural predictors and outcomes over time within an individual with no manipulation on observed variables. The aim of the observational N-of-1 is to describe the relationship between predictor and outcome over time often examining a temporal pattern, and time lags between predictor and outcome, e.g., an individual reports higher energy levels immediately after coffee than 2 hours later. N-of-1 observational designs can also address questions regarding temporal association between variables (direction of association), e.g., does stress result in you exercising less?; or does exercising result in you being less stressed?; or is it both? Burg et al. (2017) demonstrated that the relationship between stress and exercise can be uni- or bi-directional in either direction and varies from person to person. Observational N-of-1s allow the testing of psychological theory within individuals, often demonstrating that psychological variables predicting behaviours within individuals differ compared to between individuals (Kwasnicka, Dombrowski, White, & Sniehotta, 2017). In addition, such studies often find that the pattern and magnitude of outcome variance accounted for by predictor variables differ between participants. Given the repetitive nature of data collection in N-of-1 studies, a high risk of missing data is a pertinent challenge of this study design.

### Challenge 1: Non-adherence to data collection and missing data

Ecological Momentary Assessment (EMA; Stone & Shiffman, 1994), a form of Ambulatory Assessment (AA; Trull & Ebner-Priemer, 2014), is one of the most common methods of collecting data in N-of-1 studies. The repetition of EMA provides valuable insights into an individual's behaviours and subjective states. However, it also poses a burden on participants and may result in low adherence. In this context, adherence refers to the extent to which a person's behaviour corresponds with the agreed terms of usage or agreed recommendations (Sieverink, Kelders, & van Gemert-Pijnen, 2017). Interestingly, despite increased burden on users, overall, adherence to EMA protocols reported in reviews is relatively high, e.g., in older adults (Cain, Depp, & Jeste, 2009) and in youth (Wen, Schneider, Stone, & Spruijt-Metz, 2017). In older adults only 4 out of 27 studies reported adherence rates under 80% and in youth average adherence was 78% (N = 36). Cain et al. (2009)'s review also found that among clinical populations adherence was greater in studies with higher daily sampling frequencies (6+ times versus 2-3 or 4-5 times). The inverse was true among nonclinical populations, with studies with a low sampling rate (2-3 times per day) demonstrating the highest adherence. Adherence to event-contingent reporting (e.g., completing an EMA every time a participant smokes) was found in one study to be similar to adherence to signal-based (random prompts) reporting (Schüz, Walters, Frandsen, Bower, & Ferguson, 2013). Even with relatively high adherence rates, missing data still poses a challenge for researchers undertaking N-of-1 studies. Two broad approaches to address missing data in N-of-1 studies are imputation to manage missing data or, better still, user-centred study design to reduce or avoid it in the first place.

Regarding the first potential solution – imputation – it is important to determine if non-adherence is problematic by examining the pattern of missing assessments. Similar to other methodologies, missing data in N-of-1 studies can be distinguished in three patterns, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Rubin, 1976). Generally, we assume sporadic missing data to be MAR. This means that the probability of a missing response is independent of both observed and unobserved variables. MAR data can be imputed without having an impact on the causal inference. Therefore (multiple) data imputation has been advocated as a suitable approach to missing data in general, although this is only sometimes practiced in the context of N-of-1 studies (McDonald et al., 2017).

While multiple data imputation is common practice in cross-sectional and non-intensive longitudinal designs, there are reasons to question this approach in N-of-1 studies. Namely, since data in N-of-1 studies is repeatedly collected from individuals and potentially auto-correlated, the assumption of independence between variables is likely violated. In fact, longitudinal within-person studies often produce missing data that is a mixture of MAR and MNAR (Feng, Cong, & Silverstein, 2012; Graham, 2009). For example, data may be missing due to longitudinal attrition (e.g., lower response rates near the end of longitudinal data collection) and increased respondent burden (Deeg, van Tilburg, Smit, & de Leeuw, 2002; Twisk & de Vente, 2002); in these cases, multiple data imputation is unlikely to be a suitable solution. The approaches that are designed specifically to be used to deal with missing data in N-of-1 studies include using Amelia II software (www.gking.harvard.edu/amelia; Honaker & King, 2010) that imputes missing data in a single cross-section from a time series or from a time-series-cross-sectional data set implementing a bootstrapping-based algorithm. For a full comparison of missing data methods and software to fit incomplete data regression models see Horton and Kleinman (2007).

To mitigate non-adherence in N-of-1 studies User-Centred Design (UCD), also known as participatory design or co-design, can be used.  UCD describes design processes in which users influence how a design takes shape (Abras, Maloney-Krichmar, & Preece, 2004). This includes inviting users to participate in feasibility and usability studies early and often (Stappers & Giaccardi, 2017) and can also include users nominating their own predictor variables of measurement based on their experiences. This design is often advocated as a method to increase acceptance and uptake of the final study/intervention, which results from the engaging users involved in a project design (e.g., Kent & Bush, 2018). A further feasible option is using objectively measured data from unobtrusive measurement using technologies such as smartphones and wearables. While measuring an individual's physical activity using accelerometers is well established and practiced, new approaches, such as gesture recognition to identify instances of smoking (Skinner, Stone, Doughty, & Munafò, 2018) offer opportunities for high rates of behavioural measurement.

**Challenge 2: Calculating power/sample size**

Power analysis has been argued to be the most important statistical procedure when planning a study (Bolger & Laurenceau, 2013). For N-of-1 studies, conducting power analyses can be complex for several reasons. First, power has to be estimated for all levels of analysis, i.e., within participants, and, where of interest, between participants (and potentially further levels, e.g., schools). Second, conducting a power analysis for an N-of-1 study requires in-depth knowledge of the statistical procedures to analyse N-of-1 data. Fortunately, there are resources available that provide hands-on explanations of how to conduct power analyses for N-of-1 studies. Bolger, Stadler, and Laurenceau (2012) and Bolger and Laurenceau (2013), for example, offer a step-by-step approach to conduct power analysis using simulations in Mplus. Third, and perhaps most critically, N-of-1 power analyses require the assumptions about many more parameters than simpler observational and experimental studies, such as effect heterogeneity. Information about these parameters is often not available in the literature as N-of-1 studies in health psychology are still rare, and sometimes researchers fail to report all model parameters. Informed guesses about estimates are therefore often necessary. Some authors like Chen and Chen (2014) conclude from the results of their simulation study that individual design of N-of-1 studies should not be considered unless the effect size is sufficiently large. If using analysis techniques such as Autoregressive

Integrated Moving Average (ARIMA) models, some have recommended that at least 50 observations are required (Yaffee, 2012). However, such rules of thumb are highly dependent on multiple factors such as effect size, anticipated variance in measures etc and so should be considered with caution. In sum, and as with between-participant studies, those planning an N-of-1 study should consider conducting a power analysis. It prepares the researcher for the later analyses, and sensitises for the importance of detailed reporting of all model parameters in later publications.

**Challenge 3: Autocorrelation**

A distinct feature of time series data, produced in N-of-1 studies, is that of autocorrelation or serial dependency. This is where sequential data points for a given measure, particularly when there is a short time interval between them, may be associated with each other. For example, a person's stress levels today may be associated with their stress yesterday, which would be an example of a 1$^{st}$ order autocorrelative relationship. Another pattern sometimes observed when examining autocorrelation is day of the week, e.g., lower stress on a Sundays, which would be a 7$^{th}$ order autocorrelative relationship (for a graphical example of the autocorrelation of stress and a definition of autocorrelation see Naughton and Johnston, 2014). Autocorrelation can provide valuable insight into the influence of the past on present and future measurements and therefore modelling, rather than eliminating, autocorrelation is preferable from an analysis perspective (Borckardt, Nash, & Balliet, 2011). Not adjusting for autocorrelation can lead to inaccurate estimates of statistical significance; a positive autocorrelation can increase the risk of a type I error (false positive) and a negative autocorrelation, though less common, can increase the risk of a type II error (false negative) (Vieira, McDonald, Araújo-Soares, Sniehotta, & Henderson, 2017).

Two main ways to statistically manage autocorrelation are an autoregressive model or a dynamic model (Kravitz et al., 2014). Most methods using either of these broad approaches enable adjustment for autocorrelation, e.g., ARIMA modelling, dynamic regression-etc. Autocorrelation can also be accounted for when aggregating N-of-1s in a combined analysis, such as through multi-level modelling. While historically the majority of N-of-1 studies have not used appropriate statistical techniques that account for autocorrelation (McDonald et al., 2017), there are multiple techniques such as those listed above that can account for autocorrelation, though innovative statistical approaches are still needed (Davidson & Cheung, 2017). Naughton and Johnston (2014) provide a guide to a simple method for transforming an outcome or predictor variable that takes autocorrelation into account referred to as 'prewhitening,' for use when analysing N-of-1s separately. Though care should be taken with this approach if complex autocorrelation is expected beyond a simple 1$^{st}$ and/or 7$^{th}$ autocorrelative relationship in case some of the effect being investigated is removed through the transformation process (Vieira et al., 2017). However, in cases where there is insufficient information to identify and accommodate autocorrelation patterns, often a comprehensive descriptive analysis is preferable to using simpler statistical methods that cannot take autocorrelation into account. While the challenges and solutions presented so far can apply to all N-of-1 studies, there are a number of challenges specific to experimental N-of-1 studies.

## Challenges and solutions for experimental N-of-1 designs

Experimental N-of-1 studies (also sometimes referred to as within-person experiments and micro-randomised trials) are most often cross-over trials conducted with one participant acting as their own control. As a result, most individual-level confounders are held constant across different treatment periods; thus, controlling for their potential influence on the outcome of interest. Although treatment periods or blocks can be ordered according to a non-random schedule, it is preferable and more common for blocks to be randomly allocated. Therefore, from here on we will refer to the N-of-1 RCT when referring to experimental N-of-1 designs. N-of-1 RCTs are designed to include a sufficient number of treatment cross-over points in order to minimise the influence of confounding and provide enough data to establish the impact of a given treatment on the outcome of interest. In this section, we will highlight advantages and disadvantages of these designs over between-participant designs and describe how advantages can be optimised and disadvantages mitigated.

**Challenge 4: When is an N-of-1 RCT preferable to a traditional between-person RCT**

Currently, intervention evaluation design for health psychology-related research is dominated by between-person designs. In most cases, we would estimate that alternative designs, such as within-person methodology, are not considered at the design stage. Understanding why and when an N-of-1 design might be preferable is an important step. Firstly, an N-of-1 RCT could be advantageous when it is assumed that intra-individual effects might differ from those found in between-participant studies. In other words, when individuals have different change trajectories or different types or responses to an intervention. For example, Brannon et al. (2017) demonstrate how a simple text message intervention providing goal attainment feedback for a physical activity-based goal only increased activity among three out of ten adolescents, with different responses depending upon the source of the feedback (parent, peer or behavioural specialist). Furthermore, with N-of-1, we may study whether and how the amount of exposure influences each participant: is one dose enough, does efficacy increase gradually, or is there a saturation point when there has been enough intervention exposure for a change to happen? If participants need different exposure, group comparisons may conceal the effect. N-of-1 RCTs can also be used to identify mechanisms of effect of interventions through assessing the temporal relationships along a mediation pathway, e.g., does a momentary change in self-efficacy precede a change in behaviour as a result of an intervention or is it a change in behaviour that precedes a change in subsequent self-efficacy.

From a practical perspective, there are two clear advantages that increase the feasibility of N-of-1 RCTs over group based RCTs, namely time and costs. In fast developing fields, such as mHealth, the implementation of a full-scale RCT may be too slow for practical purposes and technological solutions may be outdated before study results are published. In addition, N-of-1 studies are a feasible platform for tailoring intervention delivery and data collection, possibly increasing engagement and adherence with the intervention elements (Yoon et al., 2018). For instance, participants may be interviewed to identify their preferred physical activity, and the intervention and data collection can be tailored to that specific activity. N-of-1 analyses can then examine which specific components in interventions (elements of intervention, modes of delivery) were suitable and effective for each participant. This type of tailored approach is encouraged in personalised medicine, which is driving tailored health care solutions for individuals (Hood & Friend, 2011).

In some cases, for instance in the case of rare diseases, the participant number is limited and there may not be enough statistical power to run a full-scale RCT (e.g., the example of Xeroderma Pigmentosum, systematic intervention development in rare and unstudied skin condition: Sainsbury, Walburn, Araujo-Soares, & Weinman, 2017). Moreover, strict inclusion criteria in group-delivered RCTs aiming for high internal validity could limit the intake of participants so much that external validity of the results would be diminished. An N-of-1 design could provide more flexibility in these situations. The limited number of participants and the high number of observations in N-of-1 studies may also be an advantage for a mixed methods approach to exploring intervention effects, such as explaining quantitative outcomes with patient interviews as done by Daughters, Magidson, Schuster, and Safren (2010). N-of-1 RCTs are usually preferable to traditional RCTs when the intervention can be delivered in a way that avoids carry-over effects and conditions can be randomised within a person.

**Challenge 5: Carry-over and slow onset effects**

The impact of many types of health psychology-relevant interventions often does not end abruptly after withdrawal of the treatment. For example, changing someone's attitude or enhancing a person's self-efficacy might not be easily reversible, at least not in the short-term. In fact, long-lasting effects are usually the goal of health psychology interventions. In the context of an N-of-1 trial, the treatment effect may, therefore, carry-over into a period when the treatment is removed and may influence participants' responses during this period (Elbourne et al., 2002). In N-of-1 RCTs, health psychologists need to carefully consider carry-over effects of the behaviour change techniques (BCTs) of interest.

Where carry-over effects are moderate, additional periods in-between intervention and control periods should be considered to "washout" treatment effects. Washout by design (i.e., purposefully built into the design of the evaluation) is the ideal approach for dealing with carry-over effects, though it is possible to use analytical washout, where, in its simplest form, observations immediately after a treatment has been stopped are excluded. More advanced analytical approaches to address carry-over effects are described elsewhere (e.g., Senn, 2002). Where there are likely to be enduring carry-over effects from specific interventions or BCTs, these interventions are unlikely to be appropriate for N-of-1 RCT trial evaluations. Instead, more basic cross-over N-of-1 designs could be considered, e.g., AB designs with long baseline and post-intervention data collection periods or 'traditional' nomothetic between group approaches.

If such long-term carry-over effects are present, in typical health psychology N-of-1 studies this would almost always result in attenuation of any observed treatment effect and so have an overall conservative effect on the study outcomes through inflation of the type II error rate. Less well understood and recognised slow-onset effects refer to a situation in which the full effect of an intervention may not occur immediately  (Duan et al., 2013). For example, the effect of daily self-monitoring may be more powerful after several days compared to just the first day. Designing to account for "slow onset effects" can be done in a similar way to designing for carry-over effects.

In preparation for N-of-1 trials, health psychologists therefore need to consult the literature or directly investigate (e.g., pilot studies) which interventions or BCTs exhibit carry-over and slow onset effects and the extent of these, and which do not. For this purpose, experimental N-of-1s investigating the temporal dynamics of the effects of behaviour change interventions are particularly valuable. Inauen et al. (2017), for example, investigated the temporal effectiveness of smartphone-based support groups on healthy eating and found that the support effects ended immediately after the end of the support groups. Hence, for this particular behaviour change intervention, assignment or alternation of treatment sequences would be acceptable.

A simple approach for identifying short-term carry-over effects is to compare all observations in the period after any active treatments have stopped to those same post-treatment periods but where the first or first few observations after active treatments have stopped are omitted (Duan et al., 2013). If there are no differences then there are unlikely to be short-term carry-over effects, although longer lasting carry-over effects are unlikely to be identifiable using this approach. The same approach can be used to identify slow onset effects, though where the active treatment periods rather than the periods after the treatment period are compared with and without the first observation(s). While statistical techniques can be used to identify carry-over effects, the most common approaches are controversial and are not generally recommended (Duan et al., 2013; Senn, 2002). Therefore, designing, rather than modelling, is the optimal way of managing such effects.

**Challenge 6: Identifying appropriate interventions for N-of-1 RCTs**

In light of the issue of carry-over effects identified above, only some behavioural interventions or BCTs are likely to be appropriate or feasible for evaluation using N-of-1 RCTs, though this is likely to vary according to the behaviour or phenomenon under investigation. As part of this review, we undertook a scoping assessment to explore which BCTs (Michie et al., 2013) are most/least likely to generate carry-over effects. BCTs that are time-specific were identified as being particularly suited to N-of-1 RCTs due to lesser likelihood of carry-over effects. These included BCT domains such as *scheduled consequences* (e.g., 14.1 *behaviour cost*) and BCTs such as 12.4 *distraction* and 7.1 *prompts/cues* that have temporal boundaries i.e. are most prominent 'in the moment'. Other BCTs, such as many of those within the *goals* and *planning* domain, could have potentially only small carry-over effects when they were time restricted, e.g., *goal setting* for each day separately rather than for a longer period of time. Goal setting has been successfully applied in factorial N-of-1 RCTs testing and separating effects of goal setting and self-monitoring used to increase physical activity in the general population (Sniehotta, Presseau, Hobbs, & Araújo-Soares, 2012) and older individuals (Nyman, Goodwin, Kwasnicka, & Callaway, 2016).

Less appropriate BCTs include those associated with *learning*, as learning has a high (intended) likelihood of carry-over effects. BCTs which influenced *identity changes* (e.g., 13.1 *identification of self as a role-model*), *attitudes* (e.g., 5.1 *information about health consequences*) or *knowledge/skills* of how to conduct a behaviour (e.g., 4.1 *instruction of how to perform behaviour*) were also identified as being less appropriate for N-of-1 evaluation. Once this type of information is delivered and processed, it is relatively unlikely to be 'reversible', i.e., unlearned, though this is dependent on the information, the individual and the delivery. BCTs including gradual temporal progression, such as 8.6 *graded task*, are also unlikely to be suitable due to the discrepancy between the graduation of tasks and the abrupt end of an intervention between treatment blocks required in the typical N-of-1 RCT design. Therefore, the most suitable BCTs are the ones that have time specific boundaries and the least suitable BCTs are the ones that have intended long-term effects.

**Challenge 7: Optimal allocation sequencing and blinding**

A complexity experienced when undertaking N-of-1 RCTs, not encountered with most between-person RCTs, is the need to allocate one or more cross-over periods (treatments) to the same individual, as individuals act as their own control. Several authors proposed randomisation of treatment sequences as the gold standard for N-of-1 trials (e.g., Edgington, 1996; Guyatt et al., 1990; Guyatt et al., 1988; Sackett, 1997). When the effects of a treatment cease immediately after treatment withdrawal, randomisation should lead to the strongest conclusions about the causal nature of the treatment (Tate et al., 2014). The simplest approach for creating an allocation sequence for the simplest design (a single-factorial study) is to randomly assign an equal number of treatment and control blocks during the experimental period (e.g., 7 experimental and 7 control days for a 14-day period). Urn randomisation (e.g., removing intervention and control 'balls' one at a time from an urn), which is part of adaptive biased-coin designs, is ideal for an N-of-1 RCTs as it enables a specified (usually equal) number of blocks for each treatment and increases the chances of cross-over periods compared to standard permuted-block randomisation (Naughton & Johnston, 2014). The latter is achieved using urn randomisation as the probability of assignment to each block is influenced by the blocks already allocated – each time a control 'ball' is taken from the urn, the chances of an intervention 'ball' being selected next increases. Furthermore, when a random sequence is generated, when aggregating N-of-1s, then as each participant will receive a different sequence, any ordering effects should be cancelled out.

To help with understanding carry-over effects, particularly when this is not well understood for an intervention of interest, block length can be adapted. For example, different blocks might be created with 1, 2, or 3 control days after an intervention day, to observe the length of possible carry-over effects during these 'washout' days. Alternatively, different blocks with 1, 2, or 3 consecutive treatment days may be created to investigate whether multiple treatment days increase treatment effectiveness compared to single days. Blocks with varying lengths can again be randomly allocated to create the full allocation sequence. Blocks with varying numbers of treatment or non-treatment days can also reduce the chances of predicting the change from a treatment to non-treatment period or vice versa when participants are blinded to allocation.

When undertaking a blinded N-of-1 RCT, it can be as important to blind participants from cross-over times as it is for treatment allocation for each day. An example of an allocation sequence generated using urn randomisation with different block lengths for a double blind N-of-1 RCT can be found in Naughton and Johnston (2014). When the number of participants is high (i.e., aggregated N-of-1s), different allocation sequences can be assigned to participants to test the effect of the number of treatment days on study outcomes. The same principles of allocation apply if testing multiple interventions simultaneously, e.g., with a 2x2 factorial design. An allocation sequence can be generated that ensures two or more independent treatments are switched on and off in a random fashion such that the effect of each can be efficiently assessed.

**Challenge 8: Type I error risk in N-of-1 designs**

A Type I error occurs if the null hypothesis is falsely rejected, i.e., when it is true. Several factors may contribute to a Type I error in the context of an N-of-1 design. These factors can be specific to N-of-1 designs or common to quantitative research designs in general and just encountered in N-of-1 research. In terms of N-of-1-specific sources, one factor that increases the risk of making a Type I error common to many N-of-1 studies is through a failure to account for a positive autocorrelation, as described in the autocorrelation section. A further factor is that multiple testing can be an issue when undertaking N-of-1 studies in at least two scenarios. Firstly, if analysing multiple N-of-1 datasets and evidence of a significant association between variables of interest for any one of them would be considered evidence of an association generalisable to a larger population, then this could be considered a form of multiple testing. Though when undertaking N-of-1 research, inferences about the findings are commonly considered to be limited to those individuals participating, particularly in terms of informing treatment decisions (Guyatt et al., 2000), and so multiple testing would not be relevant. The second scenario is if a series of N-of-1 datasets are analysed individually and subsequently aggregated together in further analyses (e.g., multi-level modelling) where the same general hypothesis is tested, perhaps in multiple subgroups. In principle, this could be considered multiple testing, although a single aggregated analysis would represent only one additional hypothesis test and could be considered only a minor infringement of multiple testing conventions. For both scenarios, if deemed necessary, multiple testing could be managed by a Bonferonni correction or, for a less conservative approach, the false discovery rate could be controlled for using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). Contrary to some erroneous views, analysing repeated measures collected from an individual does not constitute a form of multiple testing as each observation is collected at a separate time point.

Other factors inflating a Type I error risk that are not specific to N-of-1 but can be encountered include studies where multiple dependent variables are used ~~measured~~ or where dependent measures are broken down into sub-measures without a priori planning. As with multiple testing, in principle the Type I error risk can be mitigated if required by using correction approaches as described above, although some of these approaches can inflate

Type II error risk. A priori specification, such as in an openly accessible or published protocol and statistical analysis plan, can avoid the need for corrections including p value adjustment, which can be overly conservative.

## Discussion

Despite a long tradition of regard for between-person RCTs as the design that generates the best evidence, increasingly N-of-1 research is becoming highly valued to promote individualised behavioural medicine and behaviour change interventions (Gabler, Duan, Vohra, & Kravitz, 2011; Lillie et al., 2011). One factor that is likely to be inhibiting the greater use of within-participant methods in health psychology is a lack of training and expertise in these methods. As a consequence, poor management of the challenges identified in this paper can either lead to poorly conducted research or discourage people from using N-of-1 methods altogether, even when these methods would better address the research questions. Challenges relevant to N-of-1 designs include participants' non-adherence to data collection protocols and consequently missing data that is often inevitable with high frequency assessments, resulting in high participant burden. Calculating power in an N-of-1 study is also often more of a challenge than in between-person studies due to a greater number of potential parameters, such as the frequency of measurement, the variability in measures over time within and between individuals and the feasible frequency for providing and withdrawing an intervention (for N-of-1 RCTs).

The challenges that are specifically relevant to N-of-1 RCTs include assessing when within-person designs are preferable to a traditional between-person RCT – deciding if a high intensity within person assessment is necessary to evaluate predictors, outcomes and alternating interventions. A key question is whether an intervention of interest is likely to demonstrate carry-over or slow onset effects such that it cannot be mitigated through design or statistical counter-measures. Many health psychology interventions aim to rapidly create long lasting effects that are not easily reversible – such interventions are generally not well suited to typical N-of-1 RCT studies with regular (e.g., daily) measurement. The researchers designing N-of-1 RCTs also face a challenge of not achieving optimal allocation sequencing and blinding and increased probability of type I error (as compared to conventional RCT) due to multiple testing on the same participant. As well as identifying some of the key challenges, in this conceptual review we present specific solutions for each of the challenges (summarised in Table 1) and suggest future directions for overcoming them.

Please insert Table 1 here

**Future directions: overcoming N-of-1 study challenges**

As N-of-1 studies focus on single or a relatively small number of individuals assessed with high frequency, they are an ideal method for not only detecting but also explaining the detected behavioural patterns. Collecting and analysing multisource, multimethod data has become easier with modern technology, e.g., N-of-1 RCTs have been conducted with pedometers (Nyman et al., 2016; Sniehotta et al., 2012) and activity bracelets (Nurmi et al., 2015) to study daily steps. Multidisciplinary collaboration is needed to create advanced data management tools that enable integrating and interpreting, e.g., smartphone usage data, i.e., information from the calendar, social media, and location, to model behaviour more accurately. Adjusting for automatically tracked life events helps to control factors that may bias intervention outcomes. With consent from participating individuals, future studies could utilise the huge pool of personal data collected actively and passively by social networks, internet searches, digital calendars and smartphones, possibly through collaboration with organisations that own the services (Onnela & Rauch, 2016). Another option for data collection are keyboard apps that track all the text typed with a smartphone; the text input could be then analysed with automatic text analysis methods, e.g., Linguistic Inquiry and Word Count (Pennebaker, Francis, & Booth, 2001).

Combining multiple sources of data could give a multi-angle picture of the determinants of behaviour (Munafò & Smith, 2018). One possibility is to let participants explore their intervention data using think-aloud methods, helping them to provide possible reasons for their actions (Kwasnicka, Dombrowski, White, & Sniehotta, 2015). Novel technologies also enable self-adapting interventions that learn features that precede desired outcomes and uses those more often in the future (Dallery, Kurti, & Erb, 2015). This type of feature could further enhance smartphone apps that learn about antecedents of health behaviours and then delivers behavioural support tailored to those antecedents when real time sensor monitoring indicates a need or opportunity (Naughton et al., 2016). Evaluation of these novel technologies could draw from engineering models, such as sequential multiphase optimisation treatment (SMART) in which participant responses or characteristics influence the interventions (Collins, Murphy, & Strecher, 2007).

N-of-1 research should be participatory as it focusses on building evidence for individuals. Providing insight in data that carries such personal relevance may encourage participation and provide individuals with an opportunity to be actively involved in their health (Dunton, 2018). Furthermore, an invitation to participate in the development of the study may help participants to be more adherent to data collection as explored in the clinical health domain (e.g., Garcia, 2014). Participatory approaches of this kind may include inviting participants to develop the research protocol and to decide on the number and timing of measurements (Orlowski et al., 2015). However, there can be resource implications for this, primarily researcher time. While a participatory approach to the development of N-of-1 research may improve adherence, it does introduce new difficulties. A key factor is that it requires researchers to be flexible and open-minded towards participant-initiated changes to research protocols. For example, participants may not all complete the same measures, although a core set of measures can still be specified. Planning an N-of-1 requires taking into account the aforementioned challenges; however, when appropriate, N-of-1 design can provide higher accuracy data on temporal relationships and better intervention adherence than conventional RCT design.

**Conceptual review strengths and limitations**

N-of-1 design is widely used in other disciplines including behavioural economics and medicine. In health psychology, this design is underutilised due to several design-specific challenges. The key strength of this conceptual review is explanation of the most topical challenges that may often prevent health psychologists and behavioural scientists from undertaking N-of-1 studies. We explored and explained each challenge and provided an actionable solution for how to best overcome it in order to design systematic and high-quality N-of-1 studies. However, the review of N-of-1 challenges is not exhaustive and there are additional practical and methodological challenges that can be explored. Through this review, we would like to encourage a conversation among health psychologists and behavioural scientists about any additional prominent challenges and solutions in order to support and promote the use of N-of-1 design.

## Conclusions

The mismatch between the idiographic basis of most theories and intervention causal models and the nomothetic approach of the methodologies typically used to evaluate them is fundamentally problematic. Recent technological and statistical advances facilitate the use of high measurement methodologies such as N-of-1 to test theoretical predictions and response to intervention at the intra-individual level and sidestep this scientific incompatibility. However, there are a mixture of challenges in carrying out N-of-1 studies that are likely to be inhibiting its use in our field. Some of these challenges are unique to within-person methodology, such as statistically addressing autocorrelation and identifying which types of interventions are suitable to be evaluated using experimental N-of-1 design. Other challenges are essentially the same as those encountered with between-person approaches, such as the need to undertake power calculations, managing missing data and multiple testing issues. However, the solutions to these challenges are largely different for within compared to between-person designs. Looking ahead, due to recent technology development we are now able to design behavioural studies and interventions which can be tailored to each individual; this includes through unobtrusive data capture such as wearables and smartphone sensors. These data combined with self-report EMA data can be used to create individual models of behaviour using N-of-1 approaches to develop a truly personalised intervention for each individual. We are therefore at an opportune time to expand our use of within-person designs to better understand health behaviour and to deliver precision behaviour change interventions.

**References**

Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centered design. In W. Bainbridge (Ed.), *Berkshire encyclopedia of Human-Computer Interaction* (pp. 763-768). Great Barrington, MA: Berkshire.

Barlow, D. H. D. H., & Hersen, M. (1984). *Single case experimental designs; strategies for studying behavior change*. Oxford: Pergamon Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bolger, N., & Laurenceau, J. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York: Guilford Press.

Bolger, N., Stadler, G., & Laurenceau, J. P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285-301). New York: Guilford Press.

Borckardt, J. J., Nash, M. R., Balliet W. (2011). Statistical Analysis of Single-Case Data. In: G.J. Madden, T. Hackenberg, K.A. Lattal (Eds.). *APA Handbook of Behavior Analysis.* Washington, DC: APA.

Brannon, E. E., Cushing, C. C., Walters, R. W., Crick, C., Noser, A. E., & Mullins, L. L. (2018). Goal feedback from whom? A physical activity intervention using an N-of-1 RCT. *Psychology & Health*, *33*(6), 701-712. doi:10.1080/08870446.2017.1385783

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. doi:10.1007/s12160-017-9902-4

Cain, A. E., Depp, C. A., & Jeste, D. V. (2009). Ecological momentary assessment in aging research: a critical review. *Journal of Psychiatric Research, 43*(11), 987-996. doi:10.1016/j.jpsychires.2009.01.014

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1-8. doi:10.1016/j.jebo.2011.08.009

Chen, X., & Chen, P. (2014). A comparison of four methods for the analysis of N-of-1 trials. *PLoS ONE, 9*(2), e87752. doi:10.1371/journal.pone.0087752

Collins, L. M., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *American Journal of Preventive Medicine, 32*(5), S112-S118. doi:10.1016/j.amepre.2007.01.022

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ, 337*, a1655. doi:10.1136/bmj.a1655

Dallery, J., Kurti, A., & Erb, P. (2015). A new frontier: Integrating behavioral and digital technology to promote health behavior. *The Behavior Analyst, 38*(1), 19-49. doi:10.1007/s40614-014-0017-y

Daughters, S. B., Magidson, J. F., Schuster, R. M., & Safren, S. A. (2010). ACT HEALTHY: A combined cognitive-behavioral depression and medication adherence treatment for HIV-infected substance users. *Cognitive and Behavioral Practice, 17*(3), 309-321. doi:10.1016/j.cbpra.2009.12.003

Davidson, K. W., & Cheung, Y. K. (2017). Envisioning a future for precision health psychology: innovative applied statistical approaches to N-of-1 studies. *Health Psychology Review*, *11*(3), 292-294. doi: 10.1080/17437199.2017.1347514

Davidson, K. W., Peacock, J., Kronish, I. M., & Edmondson, D. (2014). Personalizing behavioral interventions through single-patient (N-of-1) trials. *Social and Personality Psychology Compass, 8*(8), 408-421. doi: 10.1111/spc3.12121

Deeg, D. J., van Tilburg, T., Smit, J. H., & de Leeuw, E. D. (2002). Attrition in the Longitudinal Aging Study Amsterdam: the effect of differential inclusion in side studies. *Journal of Clinical Epidemiology, 55*(4), 319-328. doi:10.1016/S0895-4356(01)00475-9

Duan, N., Kravitz, R. L., & Schmid, C. H. (2013). Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology, 66*(8), S21-S28. doi: 10.1016/j.jclinepi.2013.04.006

Dunton, G. F. (2018). Sustaining health-protective behaviors such as physical activity and healthy eating. *JAMA*. doi:10.1001/jama.2018.6621

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy, 34*(7), 567-574. doi:10.1016/0005-7967(96)00012-5

Elbourne, D. R., Altman, D. G., Higgins, J. P., Curtin, F., Worthington, H. V., & Vail, A. (2002). Meta-analyses involving cross-over trials: methodological issues. *International Journal of Epidemiology, 31*(1), 140-149. doi:10.1093/ije/31.1.140

Feng, D., Cong, Z., & Silverstein, M. (2012). Missing data and attrition. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences* (71-96). New York: Routledge.

Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). N-of-1 trials in the medical literature: a systematic review. *Medical Care, 49*(8), 761-768.

Garcia, J. C. (2014). *Beyond the numbers: A user-centered design approach for personal reflective healthcare technologies* (Doctoral dissertation). Retrieved from https://research.utwente.nl/files/6088735/Thesis-JuanJimenez.pdf

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. doi:10.1146/annurev.psych.58.110405.085530

Guyatt, G., Keller, J. L., Jaeschke, R., Rosenbloom, D., Adachi, J. D., & Newhouse, M. T. (1990). The n-of-1 randomized controlled trial: clinical usefulness. Our three-year experience. *Annals of Internal Medicine, 112*(4), 293-299. doi: 10.7326/0003-4819-112-4-293

Guyatt, G., McGinn, T. G., Wyer, P. C., Naylor, C. D., Stiell, I. G., Richardson, W. S., & Group, E.-B. M. W. (2000). Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA, 284*(1), 79-84. doi:10.1001/jama.284.1.79

Guyatt, G., Sackett, D., Adachi, J., Roberts, R., Chong, J., Rosenbloom, D., & Keller, J. (1988). A clinician's guide for conducting randomized trials in individual patients. *Canadian Medical Association Journal, 139*(6), 497.

Hogarth, R. M. (2005). The challenge of representative design in psychology and economics. *Journal of Economic Methodology, 12*(2), 253-263. doi:10.1080/13501780500086172

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science, 54*(2), 561-581. doi:10.1111/j.1540-5907.2010.00447.x

Hood, L., & Friend, S. H. (2011). Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nature Reviews Clinical Oncology, 8*(3), 184. doi: 10.1038/nrclinonc.2010.227

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. doi:10.1177/001440290507100203

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61*(1), 79-90. doi:10.1198/000313007X172556

Inauen, J., Bolger, N., Shrout, P. E., Stadler, G., Amrein, M., Rackow, P., & Scholz, U. (2017). Using smartphone-based support groups to promote healthy eating in daily life: A randomised trial. *Applied Psychology: Health and Well-Being, 9*(3), 303-323. doi:10.1111/aphw.12093

Inauen, J., Shrout, P. E., Bolger, N., Stadler, G., & Scholz, U. (2016). Mind the gap? An intensive longitudinal study of between-person and within-person intention-behavior relations. *Annals of Behavioral Medicine, 50*(4), 516-522. doi:10.1007/s12160-016-9776-x

Janosky, J. (2005). Use of the single subject design for practice based primary care research. *Postgraduate Medical Journal, 81*(959), 549-551. doi:10.1136/pgmj.2004.031005

Johnston, D. W., & Johnston, M. (2013). Useful theories should apply to individuals. *British Journal of Health Psychology, 18*(3), 469-473. doi:10.1111/bjhp.12049

Kennedy, C. H. (2005). *Single-case designs for educational research.* London: Pearson.

Kent, A., & Bush, P. (2018). Co-designing for care: Craft and wearable wellbeing. In R. Lynch & C. Farrington (Eds.), *Quantified lives and vital data: Exploring health and technology through personal medical devices* (pp. 251-280). Basingstoke, UK: Palgrave Macmillan.

Kravitz, R., Duan, N., Duan, N., Eslick, I., Gabler, N., Kaplan, H., . . . Schmid, C. (2014). Design and implementation of N-of-1 trials: a user's guide. *Agency for healthcare research and quality.* Rockville, MD: Agency for Healthcare Research and Quality.

Kwasnicka, D., Dombrowski, S. U., White, M., & Sniehotta, F. F. (2015). Data-prompted interviews: Using individual ecological data to stimulate narratives and explore meanings. *Health Psychology, 34*(12), 1191-1194. doi:10.1037/hea0000234

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. doi:10.1080/08870446.2017.1293057

Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized Medicine, 8*(2), 161-173. doi:10.2217/pme.11.7

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. doi:10.1080/17437199.2017.1316672

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., . . . Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine, 46*(1), 81-95. doi:10.1007/s12160-013-9486-6

Moeller, J. D., Dattilo, J., & Rusch, F. (2015). Applying quality indicators to single-case research designs used in special education: A systematic review. *Psychology in the Schools, 52*(2), 139-153. doi:10.1002/pits.21801

Munafò, M. R., & Smith, G. D. (2018). Robust research needs many lines of evidence. *Nature, 553*, 399-401. doi: 10.1038/d41586-018-01023-3

Naughton, F., Hopewell, S., Lathia, N., Schalbroeck, R., Brown, C., Mascolo, C., . . . Sutton, S. (2016). A context-sensing mobile phone app (Q sense) for smoking cessation: a mixed-methods study. *JMIR mHealth and uHealth, 4*(3). doi: 10.2196/mhealth.5787

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

Nurmi, J., Knittle, K., Helf, C., Zwickl, P., Palacios, P. L., Tejedor, C. C., . . . Haukkala, A. (2015). A personalised, sensor-based smart phone intervention for physical activity and Diet–PRECIOUS N-of-1 trial. *Frontiers in Public Health, 4*(1), 1-1. doi: 10.3389/conf.FPUBH.2016.01.00098

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. doi:10.1080/08870446.2015.1088014

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

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology, 41*(7), 1691. doi:10.1038/npp.2016.7

Orlowski, S. K., Lawn, S., Venning, A., Winsall, M., Jones, G. M., Wyld, K., . . . Smith, D. (2015). Participatory research as one piece of the puzzle: a systematic review of consumer involvement in design of technology-based youth mental health and well-being interventions. *JMIR human factors, 2*(2). doi:10.2196/humanfactors.4361

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates, 71*(2001), 2001.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. doi:10.2307/2335739

Sackett, D. L. (1997). *Evidence-based medicine how to practice and teach EBM*. London: Churchill Livingstone.

Sainsbury, K., Walburn, J., Araujo-Soares, V., & Weinman, J. (2017). Challenges and proposed solutions for formative research to inform systematic intervention development in rare and unstudied conditions: The case example of Xeroderma Pigmentosum. *British Journal of Health Psychology, 23*(2), 229-237. doi:10.1111/bjhp.12287

Schüz, N., Walters, J. A., Frandsen, M., Bower, J., & Ferguson, S. G. (2013). Compliance with an EMA monitoring protocol and its relationship with participant and smoking characteristics. *Nicotine & Tobacco Research, 16*(Suppl_2), S88-S92. doi:10.1093/ntr/ntt142

Senn, S. (2002). *Cross-over trials in clinical research* (Vol. 5): John Wiley & Sons.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.

Sieverink, F., Kelders, S. M., & van Gemert-Pijnen, J. E. (2017). Clarifying the concept of adherence to eHealth technology: Systematic review on when usage becomes adherence. *Journal of Medical Internet Research, 19*(12). doi:10.2196/jmir.8578

Skinner, A. L., Stone, C. J., Doughty, H., & Munafò, M. R. (2018). StopWatch: The preliminary evaluation of a smartwatch-based system for passive detection of cigarette smoking. *Nicotine & Tobacco Research, 0*(0), 1-5. doi:10.1093/ntr/nty008

Smith, G., Williams, L., O'Donnell, C., & McKechnie, J. (2018). A series of n-of-1 studies examining the interrelationships between social cognitive theory constructs and physical activity behaviour within individuals. *Psychology & Health*, 1-16. doi: 10.1080/08870446.2018.1500576

Sniehotta, F. F., Presseau, J., Hobbs, N., & Araújo-Soares, V. (2012). Testing self-regulation interventions to increase walking using factorial randomized N-of-1 trials. *Health Psychology, 31*(6), 733-737. doi: 10.1037/a0027337

Stappers, P., & Giaccardi, E. (2017). Research through Design. In M. Soegaard & R. Friis-Dam (Eds.) The encyclopedia of human-computer interaction, (pp. 43). Aarhus N, Demark: The Interaction Design Foundation.

Tate, R., Perdices, M., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The design, conduct and report of single-case research: Resources to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation, 24*(3-4), 315-331. doi:10.1080/09602010802009201

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science, 23*(6), 466-470. doi:10.1177/0963721414550706

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: how to deal with missing data. *Journal of Clinical Epidemiology, 55*(4), 329-337. doi:10.1016/S0895-4356(01)00476-0

Versluis, A., Verkuil, B., Spinhoven, P., van der Ploeg, M. M., & Brosschot, J. F. (2016). Changing mental health and positive psychological well-being using ecological momentary interventions: a systematic review and meta-analysis. *Journal of Medical Internet Research*, *18*(6), e152. doi: 10.2196/jmir.5642

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. doi:10.1080/17437199.2017.1343680

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis. *Journal of Medical Internet Research, 19*(4). doi:10.2196/jmir.6641

Yaffee, R. (2012). *An Introduction to Forecasting Time Series with Stata.* New York: Taylor & Francis.

Yoon, S., Schwartz, J. E., Burg, M. M., Kronish, I. M., Alcantara, C., Julian, J., . . . Diaz, K. M. (2018). Using behavioral analytics to increase exercise: A randomized N-of-1 study. *American Journal of Preventive Medicine, 54*(4), 559-567. doi:10.1016/j.amepre.2017.12.011

Table 1

*Challenges and solutions of N-of-1 studies in Health Psychology*

| **Challenge type/name** | **Challenge defined** | **Proposed solution** |
|---|---|---|
| *Challenge 1: Non-adherence to data collection and missing data* | N-of-1 study requires repetitive measurements performed on the same person, non-adherence to study protocol often leads to missing data; patterns of missing data vary. | Imputation under certain circumstances; employing user centred design; using non-obtrusive data collection techniques (e.g., sensors, GPS data) and novel technologies to collect data on the participant. |
| *Challenge 2: Calculating power/sample size* | Power of an N-of-1 study is relevant to the number of observations (not to the number of participants). As compared to conventional RCT designs, additional parameters need to be considered when calculating power/sample size. | Employing resources available to conduct power analyses for N-of-1 studies, for example a step-by-step approach using simulations in Mplus by Bolger, Stadler, and Laurenceau (2012) and Bolger and Laurenceau (2013). Sample size heuristics (e.g. a minimum of 50 observations to run an ARIMA model; Yaffee, 2012) are suggested by some, though these should be used with caution. |
| *Challenge 3: Autocorrelation* | Sequential data points, particularly when there is a short time interval between them, may be associated with each other. Not taking autocorrelation into account can lead to inaccurate estimates of statistical significance. | Using one of the many statistical techniques that enables the modelling or adjustment of autocorrelation, e.g., prewhitening, ARIMA modelling, dynamic regression, multi-level modelling. |
| *Challenge 4: When is an N-of-1 RCT preferable to a traditional between-person RCT?* | Understanding the circumstances when an N-of-1 RCT would be preferable to a traditional between-person RCT. | N-of-1 RCTs are preferable when: intra-individual effects differ from those found in between-participant studies; to study whether and how the amount of exposure influences each participant; to explore causal temporal relationships within the participant; to save time, cost, and to tailor |

| | | |
|---|---|---|
| | | interventions effectively; to investigate rare conditions and behavioural treatments for rare conditions. |
| *Challenge 5: Carry-over and slow onset effects* | The impact of health psychology-relevant interventions sometimes does not start abruptly after a treatment is initiated (slow onset effect) and often does not end abruptly after withdrawal of the treatment (carry-over effect). | Using wash-out periods (e.g., no intervention present), or analytical wash-out, i.e., omitting certain intervention periods in analysis. Piloting interventions to explore if the effects do carry over time or have slow onset. Using single cross-over designs (e.g. AB designs) with long data collection periods. |
| *Challenge 6: Identifying appropriate interventions for N-of-1 RCTs?* | Not all interventions or Behaviour Change Techniques (BCTs) are suitable to be used in N-of-1 experimental studies due mainly to excessive carry over effects. | Only using N-of-1 methodology when testing appropriate interventions or BCTs, i.e., those that are time specific and not designed to cause enduring changes in participants. BCTs likely to be least suitable for N-of-1 RCTs are those that include learning, changing identity, and gaining knowledge or skills. |
| *Challenge 7: Optimal allocation sequencing and blinding* | The need to maximise the number of cross-over periods and, where appropriate, minimise the ability of participants to predict cross-over points in N-of-1 experimental studies. | Randomisation of treatment sequences has been proposed as the gold standard for N-of-1 trials, i.e., to randomly assign an equal number of treatment and control blocks during the experimental period. Urn randomisation is ideal as it enables a specified (usually equal) number of blocks for each treatment and increases the chances of cross-over periods compared to standard permuted-block randomisation. Differential (e.g., random) treatment block size can reduce predictability. |
| *Challenge 8: Type I error risk in N-of-1 designs* | It is possible that when using N-of-1 designs in certain ways the risk of a Type I error is inflated, such as when generalising from the findings of any identified effects across N-of-1 investigations, when analysing the same data more than once (e.g., as separate cases and in aggregated | Multiple testing could be managed by a Bonferonni correction or, for a less conservative approach, the false discovery rate could be controlled for using the Benjamini-Hochberg procedure (1995). A priori specification, such as in an openly accessible or published protocol and statistical analysis plan, can avoid the need for adjustment. |

| | |
|---|---|
| | analyses) and when more than one dependent variable is analysed. |