An N-of-1 Evaluation Framework for Behaviour Change Applications

Claire McCallum
John Rooksby
Parvin Asadzadeh
Northumbria University
Newcastle-upon-Tyne, UK
firstname.lastname@northumbria.ac.uk

Cindy M Gray
Matthew Chalmers
University of Glasgow
Glasgow, UK
firstname.lastname@glasgow.ac.uk

ABSTRACT

Mobile behaviour change applications should be evaluated for their effectiveness in promoting the intended behavior changes. In this paper we argue that the "gold standard" form of effectiveness evaluation, the randomised controlled trial, has shortcomings when applied to mobile applications. We propose that N-of-1 (also known as single case design) based approaches have advantages. There is currently a lack of guidance for researchers and developers on how to take this approach. We present a framework encompassing three phases and two related checklists for performing N-of-1 evaluations. We also present our analysis of using this framework in the development and deployment of an app that encourages people to walk more. Our key findings are that there are challenges in designing engaging apps that automate N-of-1 procedures, and that there are challenges in collecting sufficient data of good quality. Further research should address these challenges.

KEYWORDS

Mobile Health; Behaviour Change; Evaluation; Single Case Design; Mobile Application

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312923

We have identified the following limitations of using RCTs for evaluating behaviour change apps:

- Time: RCTs take several years to conduct, yet apps are developed and released rapidly.
- Fixedness: RCTs typically require interventions to remain unchanged yet apps are regularly updated and modified after release.
- Compliance: Study participants are presented with compliance criteria and often financial incentives, but real world engagement with apps is via user choice and nudges via notifications.
- Homogeneity: Involvement of participants with similar characteristics is common in RCTs (to ensure varying characteristics are not responsible for behavior change), yet apps are used by diverse individuals.
- Averaging: RCTs involve determination of whether the app is effective "on average" across individuals. App users have a great deal of choice and may prefer to know what works (or does not work) for them specifically.
- Contact: There are often high levels of researcher-participant contact in an RCT. App users typically have no contact with developers.

The limitations listed here relate to external validity and efficiency.

Sidebar 1: Limitations of RCTs for evaluating behaviour change apps

INTRODUCTION

At the top of Blandford et al's [1] evaluation hierarchy for health behaviour change applications is *effectiveness* –evaluating whether a technology is successful in promoting the desired change. The "gold standard" form of effectiveness evaluation is the randomised controlled trial (RCT). However, as Blandford et al.[1] and McCallum et al. [9] note, RCTs do not fit comfortably with app design. Several problems with using RCTs in this context are summarised in sidebar 1.

There are alternatives to RCTs that fit more comfortably in the design process, including: N-of-1, multioptimisation strategy, CEEBIT, and microrandomized trials. However, in their recent review, McCallum et al. [9] found these alternatives are rarely used. One of the reasons for the lack of uptake, we believe, is the absence of guidance for researchers and developers on how to use such evaluation methods. In this paper we contribute a framework for using one of the alternatives: N-of-1. We will outline the benefits of automated N-of-1 trials for researchers and developers, present a framework and accompanying checklists to support the design of automated N-of-1 trials, and outline challenges and lessons learned from using the framework.

Advantages of N-of-1 Evaluation

N-of-1 (also known as single case design (SCD)), is a family of research methods (see sidebar 2) that involve participants serving as their own "control" condition. This requires a baseline phase for each participant and frequent measurement of outcomes. Repeated data collection from individuals can be laborious unless automated. The rise of smartphones and wearables suits the approach because data can be collected frequently, and remotely, via in-device sensors and logging [2, 4].

A distinguishing feature of N-of-1 is the ability to test the effectiveness of an intervention for a particular individual. Rather than provide a blanket estimation of effectiveness for an "average" individual as in group designs (such as RCTs), N-of-1 can identify who an intervention works/does not work for. They can also be used to assess the effectiveness of individual intervention components. Therefore N-of-1 can be of utility to HCI researchers who can use results to improve the design of their health apps by including components that work, defining target users, and ultimately, tailoring designs to different users [4, 6]. Although the terms "single-case" and "N-of-1" refer to individuals, trials can be aggregated to produce statistically valid inferences about effectiveness for a population.

N-of-1 methods are beginning to be used in HCI and digital health, but mainly for "self experimentation" apps [3] or for use in small scale trials conducted locally. We believe effectiveness evaluations of behaviour change apps can be more useful when conducted in the "destination setting" in which the app will ultimately be made available: app stores. Combining N-of-1s with app store based distribution approaches can (i) improve external validity by capturing real world engagement issues and assessing

Reversal (e.g. ABAB): The most rigorous type of N-of-1. Involves systematic introduction and withdrawal/removal of the intervention or its components. Removing app features may confuse users.

Multiple baseline: Several participants begin a baseline measurement only phase, either at the same time (concurrently) or different times (nonconcurrently). Baseline lengths vary across participants. This means some users would endure long baselines without intervention features.

Changing criterion: Following a baseline phase, intervention goals are implemented in a step-wise manner. Goals become progressively more challenging as they are met. Useful for app store apps incorporating goal setting features.

Alternating Treatments: Two or more interventions or components are rapidly "switched" and compared. App features that continiously change may confuse users.

Mixed (or combined) design: Elements of any type of N-of-1 can be combined. These should be combined in a way that makes sense to users (i.e. provide a convincing/coherent app storyline or flow).

Sidebar 2: The Main Types of N-of-1

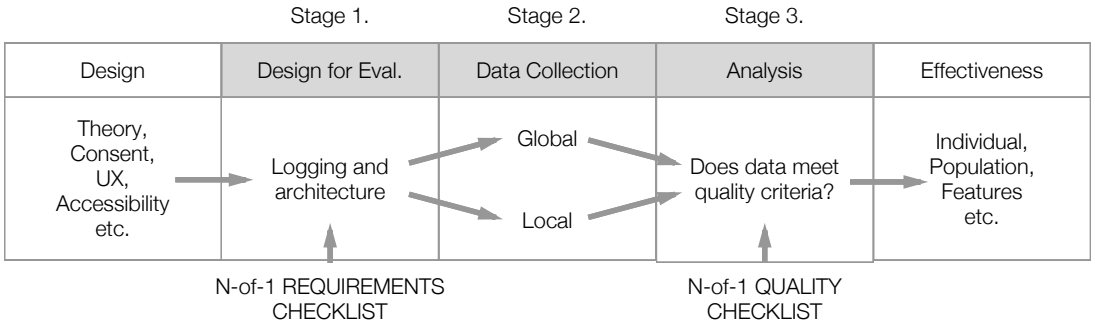

Figure 1: Framework for N-of-1 evaluation of behaviour change applications.

various individuals who download the app in different contexts, and (ii) facilitate remote, on-going data collection that enables evaluations to respond to regular app changes and updates.

A FRAMEWORK FOR N-OF-1 EVALUATIONS ENCOMPASSING APP STORE RELEASE

The framework (figure 1) has three key stages: design, data collection and analysis. The stages are explained in sidebar 3. There are two associated checklists, the N-of-1 Requirements Checklist (table 1) for the design stage, and the N-of-1 Quality Analysis Checklist (sidebar 4) for the analysis stage.

To develop the framework we collated N-of-1 criteria across existing guidelines [2, 5, 7, 8, 11] using thematic analysis. A major decision during framework development was which N-of-1 type would be acceptable in the context of app store distribution. For example, while reversal N-of-1s are the most rigorous (see sidebar 2), withdrawal of the app or its features without human explanation may create confusion and a highly disruptive user experience. Therefore, to optimise rigour and user experience, the framework recommends a mixed multiple baseline changing criterion design (sidebar 2).

In developing the framework we found three N-of-1 quality indicators recommended in the literature were problematic for operationalizing in app-store releases. These were: user authentication; active randomisation of baseline lengths, and blinding.

APPLICATION TO A BEHAVIOUR CHANGE APP

We have applied the framework to the design and deployment of a physical activity app "Quped", released on the Apple App Store. The app uses personalised goal setting and social comparison to encourage users to walk more. Over a six month period 80 users consented in the app to data collection.

Table 1: N-of-1 Requirements Checklist

		Quality indicators (QI)
Dependent Variable (DV)	1.1	DV is described with operational precision and measured with a procedure that generates a quantifiable index
	1.2	DV is repeatedly measured over time, at regular intervals
	1.3	Sufficient number of data points within baseline and intervention phases (minimum of 3/5)
	1.4	Data is collected or referenced on the validity and reliability of dependent variable measurement
Independent Variable (IV)	2.1	IV is described with replicable precision
	2.2	IV is systematically manipulated and under control of experimenter Changing criterion design: IV is continuously implemented over time
	2.3	If multiple treatments or intervention components are examined, components are introduced separately
	2.4	Fidelity (delivery and receipt of intervention) is measured
Baseline	3.1	Includes a baseline phase that provides repeated measurement of the DV
	3.2	Baseline conditions are described with replicable precision.
	3.3	Multiple baseline design: Baseline lengths vary across participants
	3.4	Baseline establishes a pattern of responding that can be used to predict the pattern of future performance, if introduction or manipulation of the independent variable did not occur (i.e. baseline data is stable).
Internal Validity	4.1	The design provides at least three replications of experimental effect at three different points in time. Multiple baselines design: There are at least three participants Changing criterion design: At least three sub-phases containing different criterion levels (i.e. increases in the goal) are implemented.
	4.2	The design facilitates opportunities for verification Multiple baseline: Introduction of the intervention is staggered over time to create verification periods Changing criterion: Verification is facilitated through varying the length of time a participant is exposed to each goal and the amount by which the goal changes
External Validity	5.1	Design supports replication of the experiment across participants and settings
	5.2	Participants and critical features of the setting are described with sufficient detail (e.g., age, gender, health condition, therapeutic setting).
	5.3	The process for selecting participants is described with replicable precision.
	5.4	Procedures for ensuring generalisability of results over time are implemented or described
Social Validity	6.1	The DV is socially important
	6.2	The intervention and experimental procedures are acceptable
	6.3	The IV is implemented in a way that is practical and cost effective, by typical intervention agents, in typical physical and social contexts

Stage 1: Design. Design involves creating or modifying an existing physical activity app to allow it to run an automated n-of-1 trial. The "N-of-1 Requirements Checklist" (table 1) outlines quality indicators in relation to 6 key criteria (dependent variable, independent variable, baseline, internal validity, external validity, social validity). These should be operationalised in the app through its interface and data logging architecture.

Stage 2: Data Collection. Data collection follows the "hybrid" approach proposed by Morrison et al [10] – app store deployment, collecting global log data (as planned in stage 1), and conducting local user interviews. In the framework, interviews are recommended for addressing specific quality indicators relating to "social validity".

Stage 3: Analysis. Analysing the extent to which data collected from Stage 2 meets N-of-1 requirements and can be used to assess app effectiveness. The N-of-1 Quality Analysis Checklist (sidebar 4) outlines key questions that should be asked of the data collected. If the data is not of sufficient quality then it is advisable to return to stage 1.

Sidebar 3: Stages of the Framework

After applying the framework we found that the app may have been effective only for a small number of users over a short period. It is not our intention to discuss effectiveness outcomes here, but to discuss lessons learned in applying the framework and analysing data quality.

Findings

Dependent Variable (DV). The DV is the measure of the target behavior to change. For our app, this was daily step count logged via the device. Log data revealed 53.8% (45/80) users had enough data points (n=3) in all 3 phases. Several users had zero-value step counts (missing data).

Independent Variable (IV). The IV was the phased introduction of the intervention components. The app store setting meant entry to the study was not in our control but dependent on the date users downloaded. However we algorithmically controlled the introduction of the intervention (the goal setting component). Logs revealed that 59% of users used the app long enough to proceed from baseline collection to the intervention phase and receive criterion changes (i.e. weekly step goal changes).

Baseline. Baseline steps are collected before receiving the intervention. In multiple baseline designs, baselines should vary in length across participants. We anticipated that upon download, app store users would expect a fully functioning app with engaging features. To prevent users having to endure a baseline phase and dropping out, we used an iPhone feature that retrieves 1 week of step data prior to when the app was installed. This meant baseline lengths must be varied (randomly) post-hoc, during effectiveness analysis. Log data analysis revealed that for many users, baseline data was unstable (highly variable) and unsuitable for predicting step values had the intervention not been introduced.

Internal Validity. N-of-1 supports internal validity (i.e. certainty that the intervention had an effect or not) by replicating the experiment multiple times, at different points in time. The app store enabled experiments to be replicated across multiple users on different weeks. Only those who used the app for several weeks received multiple criterion changes (i.e. weekly changes in step goals).

External validity. Most users (77.5%, 62/80) entered their gender and age, allowing us to describe these basic characteristics. Of these, 54.8% (34/62) were male and 45.2% (28/62) female. Most were 18-39 years (30%), only 2 (2.5%) were 60-69, and others were 40-59 (22, 27.5%). Time zone data logs showed that most users (62.5%, 50/80) were in the UK (37.5%, 30/80 elsewhere).

Social validity. We conducted 18 interviews. Users generally felt walking was important, and found the app and data collected, acceptable. However, some users were surprised to see that the app had collected historic data from before the date the app was downloaded, and would have preferred more study information explaining this in the in-app information and consent form.

Quality indicator (QI) to test–

Dependent Variable (DV):

- ☐ Are there a sufficient number of data points within baseline and intervention phases? (QI 1.3)

Independent Variable (IV):

- ☐ Was the intervention delivered and received as intended? (QI 2.4)

Baseline:

- ☐ Can baseline data be used to predict patterns of future performance? (QI 3.5)

Internal validity:

- ☐ Did the design facilitate replications in at least three points in time? (QI 4.1)
- ☐ Did the design support verification of effectiveness results? (QI 4.2)

External validity:

- ☐ Was the experiment replicated across different participants and settings? (QI 5.1)
- ☐ Can participants and settings be described? (QI 5.2)

Social Validity:

- ☐ Is the dependent variable socially important? (QI 6.1)
- ☐ Are intervention and study procedures acceptable? (QI 6.2)

Sidebar 4: N-of-1 Quality Analysis Checklist

CONCLUSION

In our efforts to perform N-of-1 effectiveness evaluations that are of high quality and are feasible for real-world contexts, we have developed a framework and reflected on its application to an app. Fundamental to the framework is design: apps must be designed to execute n-of-1 procedures and collect necessary data. We found that a framework can aid but not prescribe the necessary design work. Designing a behaviour change app that lends itself to evaluation but is also acceptable and engaging was a challenge and remains an area for much further exploration. From releasing the app and analysing the data we found that there are also challenges and issues to do with data itself. This includes dealing with missing data and high variability.

ACKNOWLEDGEMENT

McCallum was funded by a University of Glasgow Lord Kelvin Adam Smith scholarship. Rooksby, Asadzadeh and Chalmers were funded by EPSRC (EP/J007617/1).

REFERENCES

[1] Ann Blandford, Jo Gibbs, Nikki Newhouse, Olga Perski, Aneesha Singh, and Elizabeth Murray. 2018. Seven lessons for interdisciplinary research on interactive digital health interventions. *Digital Health* 4 (2018).

[2] Jesse Dallery, Rachel N Cassidy, and Bethany R Raiff. 2013. Single-case experimental designs to evaluate novel technology-based health interventions. *Journal of Medical Internet Research* 15, 2 (2013).

[3] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons learned from two cohorts of personal informatics self-experiments. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. ACM.

[4] Eric Hekler, Pedrag Klasjna, William Riley, Matthew Buman, Jennifer Hubert, Daniel Rivera, and Cesar Martin. 2016. Agile science: creating useful products for behavior change in the real world. *Translational Behavioral Medicine* 6, 2 (2016).

[5] Robert H Horner, Edward G Carr, James Halle, Gail McGee, Samuel Odom, and Mark Wolery. 2005. The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children* 71, 2 (2005), 165–179.

[6] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3063–3072.

[7] Liesa A. Klein, Daniel Houlihan, James L Vincent, and Carlos J Panahon. 2017. Best practices in utilizing the changing criterion design. *Behavior analysis in practice* 10, 1 (2017).

[8] Thomas R Kratochwill, John H Hitchcock, Robert H Horner, Joel R Levin, Samuel L Odom, David M Rindskopf, and William R Shadish. 2013. Single-case intervention research design standards. *Remedial and Special Education* 34, 1 (2013).

[9] Claire McCallum, John Rooksby, and Cindy M Gray. 2018. Evaluating the impact of physical activity apps and wearables: Interdisciplinary review. *JMIR mHealth and uHealth* 6, 3 (2018).

[10] Alistair Morrison, Donald McMillan, Stuart Reeves, Scott Sherwood, and Matthew Chalmers. 2012. A Hybrid Mass Participation Approach to Mobile Software Trials. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1311–1320. https://doi.org/10.1145/2207676.2208588

[11] Robyn L Tate, Michael Perdices, Ulrike Rosenkoetter, Donna Wakim, Kali Godbee, Leanne Togher, and Skye McDonald. 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 (2013), 619–638.