Evaluating a financial incentives scheme intervention to promote healthy eating and physical activity: a randomised control trial

Protocol for a mixed-method evaluation

Version 9

11th Aug 2023

Trial title		Evaluating a financial incentives scheme intervention to promote healthy eating and physical activity
Intervention name *(if applicable)*		Financial incentives
Trial registration	*Trial identifier*	43198
	Registry name	ISCTN
	Items from WHO Trial Registration Data Set	-
Protocol date and version		V8 24/07/2023
Funding sources *(specify type of support: financial, material, or other)*		Financial support from DHSC
Duration of funding		19 months
Trial sponsor	*Name*	Hannah Behrendt
	Contact information (address, email)	Hannah.Behrendt@bi.team
Main contact	*Name*	Hazel Wright
	Contact information (address, email)	hazel.wright@bi.team

Version control		
Version	**Date**	**Changes**
V8	21/07/23	Major • Include/exclude analysis in light of new contractual arrangements between HUL and BIT • Change in the focus groups in the qualitative IPE, as per DHSC's request. • Removing the interim analysis, as per DHSC's request Minor • Expand the list of contributors in light of changes within the BIT team • New mitigation strategies to high attrition added • Correction to table 9 - Data collection method for dietary outcomes • Minor change in one of the IPE's metrics definition for additional clarity • Fixing an imprecision in some of the regression equations - they should not include household ID fixed effects.
V9	11/08/23	Minor • Update the list of contributors in light of changes within the BIT team • Re-instate some subgroup analysis in the IPE section that had been removed by mistake in v8 (Tables 16-17-18).

Roles and responsibilities of protocol contributors		
Name	**Affiliation**	**Role in trial**
Hazel Wright	Behavioural Insights Team	Project lead Chief Investigator and Principal Investigator
Dr Filippo Bianchi	Behavioural Insights Team	Policy and IPE lead
Dr Giulia Tagliaferri	Behavioural Insights Team	Impact evaluation lead
Dr Lev Tankelevitch	Behavioural Insights Team	IPE and impact evaluation support
Dr Yihan Xu	Behavioural Insights Team	Impact evaluation support
Tim Hardy	Behavioural Insights Team	Analytical strategy support
Dr Zara Goozee	Behavioural Insights Team	Implementation and incentive design support
Rizwaan Malik	Behavioural Insights Team	IPE support
Jemuwem Eno-Amooquaye	Behavioural Insights Team	IPE support

Dr Alex Sutherland	Behavioural Insights Team	Quality assurance to the Trial Protocol
Bram Reitsma	Behavioural Insights Team	Impact evaluation support
Niall Daly	Behavioural Insights Team	IPE support
Dr Helen Brown	Behavioural Insights Team	IPE lead

Table of contents

Executive Summary 5
1. Introduction 7
2. Impact Evaluation 9
2.1 Objectives and hypotheses 9
2.2 Procedure 10
2.2.1 Study design, interventions and comparators 10
2.2.2 Setting and recruitment 12
2.2.3 Baseline phase 16
2.2.4 Randomisation 19
2.3 Outcomes 23
2.3.1 Primary outcomes 24
2.3.2 Secondary outcomes 27
2.3.3 Exploratory outcomes 29
2.4 Sample size and minimum detectable size 31
2.4.1 Summary 31
2.5 Assignment of interventions 37
2.5.1 Allocation 37
2.5.2 Blinding (masking) 38
2.6 Data collection, management, and analysis 38
2.6.2 Data quality management 44
2.6.3 Statistical methods 44
2.7 Monitoring 55
2.7.1 Data monitoring 55
2.7.2 Harms 59
2.7.3 Auditing 59
3. Ethics and dissemination 59
4. Implementation and Process Evaluation 61
4.1 IPE Design 61
4.1.1 Research questions 61
4.1.2 Unintended consequences 64
4.2 Quantitative methods 65
4.2.1 Methods overview 65
4.2.2. Reach, acquisition, and usage analysis 66
4.2.3. Engagement with the intervention and mechanisms of action 68
4.2.4. Analysis of ‘gaming’ and data errors 72
4.3 Qualitative Methods 73
4.3.1 Methods overview 73
4.3.2 Data collection with intervention recipients and intervention non-recipients 76

4.3.3 Data collection with the delivery partner (HUL) 81
4.3.4 Data collection with the reward partners 82
4.3.5 Data collection with City of Wolverhampton Council 84
4.3.6 Data collection - timelines 85
4.3.7 Risks 86
4.4 Qualitative Data Analysis 88
6. Limitations and generalisability 89
7. Appendices 90
Appendix A - Alternatives to Intake24 90
Appendix B - Additional information on power calculations 91
Appendix C - Recommended dietary intake according to the Recommended dietary intake according to the Government Dietary Recommendations 94
Appendix D - Possible additional analysis subject to contractual agreement 94

Executive Summary

Promoting healthy eating and more physical activity in England can help to improve population health. To promote these health behaviours the Department for Health and Social Care (DHSC) commissioned HeadUp Labs (HUL) to develop and implement an app-based financial incentive scheme to provide rewards to people contingent on performing behaviours related to healthy diets and physical activity. The Behavioural Insights Team (BIT) will design and conduct an independent mixed-methods evaluation of the financial incentive scheme including an impact, and an implementation and process evaluation. This document summarises BIT's plans for the impact evaluation.

This document includes more detail than what would normally be covered in an academic protocol for an impact evaluation, as it also captures considerations around the implementation, mitigation strategies, or the rationale for specific design choices.

Summary of the impact evaluation

To test the effectiveness of the financial incentives, BIT will run a randomised controlled trial (RCT). The main research question this study aims at answering is **whether the incentive scheme improves physical activity (PA) and diet healthfulness.** This will be a clustered randomised trial, in which randomisation happens at the household level.

Participants will be adults, recruited from the community in Wolverhampton. All participants will receive access to the HeadUp app and a wearable tracker if they do not already own one. Participants will complete a baseline period during which they will familiarise themselves with the app and wearable and provide baseline data, after which they will be randomised to one of four conditions:

- **A.** Control group: Access to the HeadUp app and the wearable tracker
- **B.** Intervention groups: On top of the HeadUp app and the wearable tracker participants will receive a financial incentive intervention for the entire duration of the trial with:
    - **B.1** low value incentives, i.e. low £ per point ratio
    - **B.2** medium value incentives, i.e. medium £ per point ratio
    - **B.3** high value incentives, i.e. high £ per point ratio

A detailed description of the intervention and baselining logic is provided separately by HUL.

The primary outcomes consist of **2 physical activity outcomes** and **4 dietary outcomes**. The primary physical activity outcomes will be (i) moderate and vigorous physical activity (MVPA) in minutes per day and (ii) daily steps measured objectively through a wearable device. The primary dietary outcomes will be daily intake of (i) fruit and vegetables (g/day), (ii) fibre (g/day), (iii) saturated fat (% of food energy intake) and (iv) free sugars (% of food energy intake) measured through 24 hour dietary recalls.

Secondary outcomes will be (i) daily energy expenditure as measured by the wearable, (ii) daily energy intake as measured by 24 hour dietary recalls (Intake24 surveys), (iii) a healthy eating score based on consumption of key food groups, macro- and micronutrients, and (iv) self-reported weight.

Exploratory outcomes will include (i) motivation to change PA and dietary intake as well as (ii) potential unintended health consequences (i.e. mental health and sleep duration). The motivational and mental health outcomes will be measured by in-app mini-surveys, whereas sleep duration will be measured by the wearable.

All primary and secondary outcome data will be collected at the baseline (i.e. pre-randomisation), at one month, three months, and at five months after the randomisation. The exploratory outcomes will be collected at the baseline and at five months after the randomisation.

The pre-planned primary analysis will compare the 2 physical activity outcomes and 4 dietary outcomes at five months post-randomisation between the control group and the three pooled financial incentives groups (i.e. being offered any level of financial incentive).

The pre-planned secondary analyses (see Table 11 for details) will focus on investigating (i) energy intake, weight, energy expenditure and a healthy eating score; ; (ii) shorter-term impact of the intervention (i.e. the impact at one month and three months after randomisation) on the primary outcomes; (iii) the impact of the intervention intensity (i.e. low, medium, high incentive level) on primary outcomes five months after randomisation.

The pre-planned exploratory analysis will investigate (i) the intervention impact on the primary outcomes at five months for various subgroups of interest (i.e. deprivation, sex, age, ethnicity, baseline PA level, and baseline dietary intake[1], see Table 12 for details), (ii) the intervention impact on participants' motivation to increase PA and improve dietary intake at five months (subject to contractual agreements, see **Section** 2.3.3.1 for details**)**, and (iii) whether the interventions had any unintended health consequences on sleep quality and mental well-being at five months.

Outcomes will be presented as the absolute difference between the control and intervention group.

Summary of the implementation and process evaluation

A mixed method implementation and process evaluation (IPE) will be conducted alongside the impact evaluation. Whilst the impact evaluation will test the effectiveness of the financial incentive scheme, the IPE will aim to identify why and how the intervention achieves - or fails to achieve - the expected outcomes in relation to the Theory of Change (ToC). The IPE will also explore potential desirable and undesirable unintended consequences. The IPE will focus particularly on

[1] Subgroup analysis of baseline PA level and baseline dietary intake are subject to contractual agreement.

understanding issues pertaining to (i) reach of the intervention, (ii) engagement with the intervention, (iii) mechanisms of action, (iv) and implementation and feasibility.

The methods used for the IPE will be (i) rooted in the details of the Theory of Change and the user journey, (ii) mindful of the needs of the research participants, especially those from more deprived communities, and (iii) form part of an integrated plan with the impact evaluation, so that the analysis from the IPE can help to explain any significant or null effects observed. Based on these three principles, the mixed-methods IPE will triangulate findings from qualitative interviews and focus groups and quantitative analyses of in-app data. Comparing and contrasting between these multiple sources of data will allow us to (i) gain insights across a large number of individuals whilst also (2) developing an in-depth understanding of individual experiences.

1. Introduction

Behavioural risk factors represent the largest opportunity to reduce health burdens across the population, making up more than 50% of the preventable Disability Adjusted Life Years (DALYs) as estimated by the Global Burden of Disease study.[2] Furthermore, behavioural risk factors have a steep social gradient, and are therefore a key contributor to health inequities.

Unhealthy diets and low levels of physical activity are associated with a wide range of chronic conditions, including excess weight, cardiovascular disease, type 2 diabetes, and some forms of cancer.[3] [4] Despite the importance of eating healthily and being physically active, it is estimated that nearly 40% of adults do not reach the recommended 150 minutes of physical activity per week[5], 72% of adults consume less than five portions of fruit and vegetables per day[6], and on average adults consume more energy, saturated fat, and sugar than recommended[7].

In the UK, the Eatwell guide and the Chief Medical Officers (CMO) guidelines for physical activity provide recommendations for how adults can achieve a healthy diet and healthy physical activity levels. To support adults in England to translate these recommendations into practice, the DHSC has decided to pilot an app-based financial incentive scheme to incentivise adults to eat healthier diets and be more physically active. The financial incentive scheme will be piloted in Wolverhampton and its potential effectiveness will be evaluated to inform decisions about whether and how to further scale the intervention.

---

[2] Global Burden of Disease Data (link)
[3] ibid.
[4] Scarborough P, Bhatnagar P, Wickramasinghe KK, Allender S, Foster C, Rayner M. The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006–07 NHS costs. Journal of public health. 2011 Dec 1;33(4):527-35.
[5] Sport England Active Lives Adult Survey November 2019/20 Report (link)
[6] Health Survey for England 2018 (link)
[7] National Diet and Nutrition Survey 2014/15 to 2015/16 (link)

DHSC commissioned HeadUp Labs (HUL) to develop the financial incentive scheme and deliver it through their digital healthy-lifestyle app. The Behavioural Insights Team (BIT) will act as an independent evaluator of the financial incentive scheme to understand whether, and to what extent, financial incentives can motivate behaviour change.

BIT will also conduct an implementation and process evaluation of the scheme to gain an in-depth qualitative understanding of how users viewed and interacted with the intervention, the mechanisms through which the intervention worked (or the barriers for why it did not), and to identify opportunities to further improve the intervention. This will be done using both qualitative (interviews and focus groups) and quantitative methods (analysis of app-based metrics).

This document sets out the methodology for how BIT intends to evaluate the financial incentive scheme. This document should be read in conjunction with the *Intervention design plan by HUL.*

2. Impact Evaluation

2.1 Objectives and hypotheses

The primary objective of the impact evaluation study is to assess the effectiveness of the financial incentive scheme at

- **increasing moderate-vigorous physical activity (MVPA minutes/day and steps/day)** among adults recruited from the general public in Wolverhampton.

  and

- **improving the healthfulness of the diet (fruit and vegetables in g/d, fibre in g/d, free sugars in % of food energy/day, and saturated fat in % of food energy/day)** among adults recruited from the general public in Wolverhampton.

BIT's impact evaluation will be designed to test the hypothesis that financial incentive schemes significantly increase physical activity (PA) levels **(throughout, this is defined through our two primary outcomes for physical activity)** and improve the healthfulness of recipients' dietary intake **(throughout, this is defined through our four primary outcomes for diet)**, by comparing the behaviour of users allocated to a control group (no financial incentives offered) to the behaviour of users allocated to the pooled treatment groups (low, medium or high level of financial incentive offered).

The secondary research questions include:

- **Broader effects on PA and diet:** Does offering financial incentives significantly affect participants' energy expenditure, energy intake, their score on a healthy eating score based on consumption of key food groups, macro- and micronutrients, and weight) five months after randomisation?
- **Shorter-term effects:** Does offering financial incentives effectively improve participants' dietary intake and PA in the shorter term (one and three months after randomisation), comparing the behaviour of users allocated to a control group (no financial incentives offered) to the behaviour of users allocated to the pooled treatment groups (low, medium or high level of incentive offered)?
- **Optimal incentive value:** Do different levels of financial incentives achieve different effect sizes on participants' dietary intake and PA level five months after randomisation, comparing the behaviour of users allocated to a control group (no financial incentives offered) to the behaviour of users allocated to

each treatment group (low, medium and high level of incentive offered)?

The exploratory research questions include:

- **Subgroup analyses:** Does the incentive scheme also work among specific population subgroups five months after randomisation in terms of improving dietary intake and PA (focusing on primary outcomes), comparing the behaviour of users allocated to a control group (no financial incentives offered) to the behaviour of users allocated to the pooled treatment groups (low, medium or high level of incentive offered)?
- **Longer-term effects on motivation to change:** Does offering financial incentives effectively increase participants' motivation to improve dietary intake and to increase PA level five months after randomisation, comparing the motivation level of users allocated to a control group (no financial incentives offered) to that of users allocated to the pooled treatment groups (low, medium or high level of incentive offered)?
- **Unintended consequences:** Does the incentive scheme have any significant impact on participants' sleep or mental health five months after randomisation, comparing the behaviour of users allocated to a control group (no financial incentives offered) to the behaviour of users allocated to the pooled treatment groups (low, medium or high level of incentive offered)?

2.2 Procedure

2.2.1 Study design, interventions and comparators

To provide the most robust causal inferences while minimising bias, we will conduct a randomised controlled trial (as recommended by the Magenta Book) for impact evaluation. In the RCT, we will compare various outcomes of interest across different time points between the aforementioned financial incentive intervention groups with the control (no financial incentives) group.

All participants, regardless of the treatment conditions, will get access to the "Better Health: Rewards" App developed by HUL and receive a wearable tracker if needed. The App is a mobile application that provides users with personalised health tracking services, offering them health-promoting feedback based on real-time monitoring (via the wearable) of their physical activities, sleep, and psychophysical indicators. The app can be paired up with mainstream wearable devices, including Apple Watch, Fitbit, Garmin, Google Fit, and HeadUp's own wearable, branded "Better Health: Rewards fitness tracker".

Our proposed trial is designed so that participants assigned to the *control group* will have an experience that is identical to that of users in the intervention groups, *except for the financial incentives themselves that are* contingent on their diet and PA

behaviours. This means that our comparison group is an ‘active control group’ because having access to wearables and the app may induce behaviour change.

Participants in the intervention group will receive one of three versions of the financial incentive intervention outlined below - which will only differ in the size of the rewards. Participants assigned to the *three intervention groups* will similarly receive a wearable and a version of the pilot app which enables measuring of physical activity and diet-related behaviours, together with redeeming of low-, medium-, and high-value in-app incentives, respectively (see Table 1).

Table 1. Summary of features available to users in each trial arm

Trial arms	Access to PA measurement / surveys via the app	Access to app content such as nudges and goal-setting	Access to wearable device if needed	Access to financial incentives contingent to behavioural change
Control group	Yes	Yes	Yes	No
Treatment group 1 - Low value incentive	Yes	Yes	Yes	Yes - Low value
Treatment group 2 - Medium value incentive	Yes	Yes	Yes	Yes - Medium value
Treatment group 3 - High value incentive	Yes	Yes	Yes	Yes - High value

The choice to provide control group participants with an experience that is identical to that of users in the intervention groups, except for the financial incentives themselves, has three core advantages:

- **Isolating the effect of the incentives:** the core question of the programme is whether *financial incentives in the context of the app* can improve health behaviours (on top of the app on its own), not whether financial incentives combined with a wider digital behavioural intervention can promote health behaviours. Offering access to the app and the wearable device to participants in the control group will enable us to better isolate the impact of financial incentives.
- **Evaluation feasibility:** Offering an engaging app experience and a wearable device worth £39 to the control group minimises the risk of a high dropout rate in the control group, which would otherwise make the evaluation challenging from a feasibility perspective.

- **Marketing and engagement:** Offering an engaging app experience and a wearable device to the control group would help to engage people in the scheme and make it easy for the marketing campaign to position the scheme as an appealing health promotion offer across all study arms: at a minimum all participants will receive a free wearable device worth £39 and an engaging app experience.

2.2.2 Setting and recruitment

2.2.2.1 Recruitment target and timeline

This study will be conducted in the local authority (LA) of Wolverhampton. Participants will be adults recruited from the community in Wolverhampton through an engagement and marketing campaign led by HUL.

The user recruitment period will last 6 to 8 weeks depending on the pace of recruitment, starting in February 2023. During week 6 of recruitment, uptake will be evaluated and a decision will be made whether to extend the recruitment window by 2 weeks.

During the recruitment period, HUL aims at having about 25,800 users starting the onboarding phase (see the Section 2.4 power calculation for more details). During the onboarding phase, a user downloads the app, signs up, and provides consent to participate in the study. By the end of week 6, or week 8 in case of slow recruitment, the recruitment window closes, meaning anyone who has not completed the onboarding process by then cannot take part in the study.

Since the marketing and acquisition campaigns will predominantly be deployed through online marketing, participants are likely to have clicked a link from their phone to the app store listing. From there, they can download the app for free.

For offline campaigns, a QR code will be used to direct people to the app store listing directly to maximise conversion. Prospective participants can also search the App Store / Google Play for the app.

2.2.2.2 Providing informed consent

After an initial set of app-orientation screens explaining what the app is about, participants provide informed consent to participate in the study. As part of the consent process, interested users will be asked to read a participant information sheet and agree (or not agree) to:

- Take part in a study in which they will be randomly allocated to a control or a financial incentive intervention.

- Being contacted for research purposes (e.g. interviews being conducted as part of the implementation and process evaluation).
- Have their data shared with BIT to allow for data analysis.

Participants not providing informed consent will be excluded from the trial. Consent forms will be provided as part of the documentation for the ethical approval process. Consented users are subsequently asked to confirm they are over 18 years old and to the privacy notice and Terms of Service relating to the app. They may also optionally subscribe to ‘reminder’ emails from the app about their progress and other features. Users may unsubscribe from these communications at any time in the app.

2.2.2.3 Registration process

Participants will then be prompted to register their details on the app using one of the NHSX-approved authentication methods (Apple ID, Facebook, Google or email / password). To mitigate fraud risk, the app also requires participants to confirm their phone number via a One-Time Password (OTP).

At the second stage of registration, participants will be asked to enter mandatory information, including full name, date of birth, self-reported height, self-reported weight, gender, ethnicity, and postcode (full address). Once registered and logged-in to the app, they will also be asked to enter non-mandatory information, including education, motivations to change, and disability status.

Participants will be considered eligible if they meet the following criteria:

(i) geographic criteria (resident in Wolverhampton)

(ii) age criteria (18+ years old)

Those that are non-eligible will not be granted further access to the app, instead they will be signposted to relevant services where applicable. Please see **HUL's intervention design plan** for more detail. The app will show participants information on eligibility criteria and ask them to self-assess whether they are eligible to participate. Participants will be informed that the app provides general health information to encourage a healthy lifestyle.

During the sign-up process, participants will be asked to confirm they have read and understood the eligibility information and whether the pilot is suitable for them. The app will advise that prospective participants should contact a health professional if they have any concerns about using the app. The app may not be suitable for people who have (or have had) eating disorders or any other condition which may affect someone's ability to change their diet or physical activity behaviours. Users should only sign up to the app if they, and their health professional if appropriate, agree that the app is suitable for their use to ensure that the app will be used safely.

The eligibility criteria that participants need to self-assess against a clear list of eligibility criteria to use the app (and therefore be enrolled in the study). The eligibility criteria is set out in **Appendix I Eligibility_Screening criteria_V6.1**.

If participants' self-assessment indicates that they are eligible, they will proceed to the second stage of screening, otherwise their sign-up will end and they will not be able to use the app further, nor be included in the trial.

Once a participant has provided informed consent, registered, and confirmed they meet the eligibility criteria, their onboarding phase ends and their baseline phase starts, which is elaborated in the next section.

2.2.3 Baseline phase

The baseline period is specific to each user. It starts the day the user has been positively assessed for eligibility (passes eligibility criteria). This can be any day within the recruitment period. It ends when the recruitment window closes:

- Participants cannot complete baselining if they do not respond to at least one 24-hour dietary recall using Intake24[8]. The second intake24 is optional; users are encouraged to complete the second one within 3 days – this 3-day window is intended to streamline the onboarding experience for participants and provide a sufficient time period for participants to complete a second Intake24, while reducing the risk that the participants' baselining period is unduly protracted (which may add to abandonment-risk).
- Participants cannot complete baselining if they do not connect a wearable device; the (physical activity) challenges require at least 5 days of data[9] during the baseline period. A day of data constitutes a day in which at least one step is registered. To calculate the baseline physical activity (PA), the first day is excluded as it is assumed that the data is partial only on the first day.
- Participants answer a food frequency questionnaire which will be used to create personal diet challenges (please see HUL's intervention design plan for more detail).

Their baseline period ends when this happens, and they enter randomisation

During this time:

- Participants indicate whether they need a wearable device and they receive it via mail;
- Participants may complete optional survey instruments (mental health, how they heard about the programme, education survey, motivations to change, disability survey);
- Participants provide the baseline data that HUL requires to fine-tune and personalise the incentive scheme (see the design protocol developed by HUL for more details);
- Participants provide the baseline outcome data that BIT requires for evaluation:
    - At least one Intake24 recall, to measure baseline dietary intakes;

[8] Intake24 is a web-based self-completed 24-hour dietary recall system and has been widely used in national food and nutrition surveys.
[9] Users exercise minutes and heart rate (HR) data are not required during the baseline period.

- Physical activity data for 5 days by the wearable device[10];
- Sleep quality data and mental health *(note: providing these data is not a prerequisite for entering randomisation)*
- Motivation to increase PA and to improve dietary intake

Collecting outcome measures before exposing participants to the intervention will allow us to control for baseline average dietary intake and physical activity in the analysis (increasing precision of the treatment effect estimates through reducing pre-intervention between-group variance). See the data transfer supplement for the full list of data to be collected and transferred between HUL and BIT.

A significant risk for the success of the evaluation is if not enough participants successfully enter the trial (measured as entering randomisation), either because the target recruitment number is not met, or because attrition during the baseline phase is higher than estimated by HUL.[11] To address this, BIT and HUL have decided on the following mitigation strategies (more details are provided by HUL in their acquisition and mitigation plan):

Table 2. Mitigation strategies

Decision - When	Focus	What	Team responsible
Before the trial	Recruitment and baseline phase	BIT adopted a **conservative approach to power calculations** to determine the required sample size (pre-post correlation of outcome measures, distribution of household sizes that are expected to sign up, proportion of people that are expected to sign up within each household).	BIT
Before the trial	Recruitment	**Providing a free wearable device** to all participants who do not have one (including those in the control group), which might incentivise people to enrol into the study. This will help to recruit participants within the short timelines available.[12]	HUL

[10] Note that a minimum of 4 valid days of physical activity data are required for inclusion in the evaluation. The first of the 5 days of data collected at baseline is therefore excluded from the evaluation as it is assumed to be a partial day. A valid day of data is one for which there is at least 6 hours of device wear time. This is explained in section 2.6 Data collection, management, and analysis

[11] BIT and HUL acknowledge that the risk of attracting too many users is also present. In this case, HUL will collect the same data as currently proposed during onboarding /registration (i.e. email, name, DOB, BMI, consent, gender etc) from the 'extra' users who will be invited to take part in the pilot. The app will limit the total number of registered users to 31,500. Once this number has been reached, users will see a message before they sign up, alerting them that the pilot is full and no more registrations are permitted, nor data captured.

[12] For the aim of the evaluation we do not consider the wearable a financial incentive (as part of the scheme), as the wearables will be provided to both the intervention and control group primarily to measure physical activity. Providing the wearable to both the intervention and control group will allow us to isolate the effects of providing financial rewards that are contingent on behaviour change. However, we acknowledge that users' behaviour will be affected by the presence of a wearable device.

Before the trial	Recruitment	**Testing and fine-tuning of recruitment materials** via UX research and online RCTs.	HUL, BIT supported with one online RCT
Before the trial	Baseline phase	**Testing and fine-tuning the onboarding phase** via UX and acceptability user testing research. This includes **minimising the number of questions** users are asked during the onboarding phase (for the evaluation and the fine-tuning of the scheme).	HUL
Before the trial	Baseline phase	**Make data collection easy and attractive:** (i) Adopt reminders to encourage data collection; (ii) Display the Intake24 survey link prominently on the 'to-do' list dashboard of the app when users open it so they'll see it every time they open the app regardless of the prompt; (iii) Encourage two recalls of Intake24 and wearing the device with a small payment (£5 as points); (iv) Provide feedback after each Intake24 recall	HUL
Before the trial *(triggered)*	Baseline phase	**Extend the baseline phase** to up to 2 weeks if users are slow to provide their data	HUL
Before the trial	Onboarding phase	**Adopt a flexible onboarding period for 'slow onboarding users' if needed** - ideally a user completes screening in 2 or 3 weeks (nudged to do so), but if they don't, they enter the trial (randomisation) when they have met the baseline criteria, rather than after a predetermined window of time. The latest calendar date that a user can complete onboarding (pass screening) is the last day of the recruitment window.	HUL
During the trial	Recruitment	**Extend the recruitment period beyond 6 weeks if necessary**	HUL
During the trial (triggered at M5)	Trial period	**Make data collection attractive:** Consider encouraging wearing and syncing the wearable device with the app in the crucial data collection weeks at m3 and m5 with a small payment (£10), if attrition is deemed relatively too high[13] at the data collection point at m1 and m3.	HUL
During the trial (triggered at M3 and M5)	Trial period	Make data collection attractive: Consider encouraging filling in Intake24 at m3 and m5 with a larger payment (£10), if attrition is	HUL

[13] We consider the attrition too high if the number of users not providing their data 10% smaller than expected at the start of the 1 month mark (see Table 6 for the expected attrition at different trial stages).

		deemed relatively too high[14] at the data collection point at m1. At m5 - for PA, users will receive £5 for syncing their fitness tracker (any amount of data > 0) during weeks 21 and 22 - up to 2 times so they can earn up to an extra £10.	
During the trial (triggered)	Trial period	**Extend the data collection window** from 2 to 3 weeks (21 days) at m3 and m5 for all measures (i.e. Intake24, steps, WHO5, sleep and weight)	HUL
During the trial	Trial period	**Pull data collection forward.** Change M5 data collection window to start earlier i.e. during the 20th week (day 134 to day 155).	HUL
After the trial	Recruitment and baseline phase	**Shift the focus of the evaluation to a shorter follow-up** (e.g. after 3 months of treatment instead of 5 months). If the sample size at 5m is less than 80% of the total sample size (n = 4,200) that the trial requires in order to detect a minimum effect size for primary outcomes (see Table 7 for details), we will use data from the 3m mark instead of the 5m mark in the analysis.[15]	BIT

2.2.4 Randomisation

Randomisation will be conducted at the household level. This approach has two advantages over individual-level randomisation: (i) avoids the risk of contamination within the same household (but opens up the possibility of additive effects of multiple household members being in treatment together); (ii) reduces the risk that we alienate and / or confuse users in the control group whenever they cohabit with someone assigned to a treatment arm.

Users will enter randomisation only if/when they have completed the baseline phase. This implies that the day of randomisation is different across users even within the same households: it always happens at the end of their baseline phase (if successful), but the exact calendar day depends on when they have downloaded the app, when they have completed the onboarding, and how long it took them to complete the baseline phase (see Figure 1 for an example from a user's perspective).

The randomisation algorithm developed by HUL would then:

1. Check if another user from the same household has already been randomised

[14] We consider the attrition too high if the number of users not providing their data is 10% smaller than expected at the start of the 1 month mark (see Table 6 for the expected attrition at different trial stages).

[15] Results for the 5m mark will still be reported; however, we will rely on results from the 3m mark to provide an assessment of the impact of offering financial incentives.

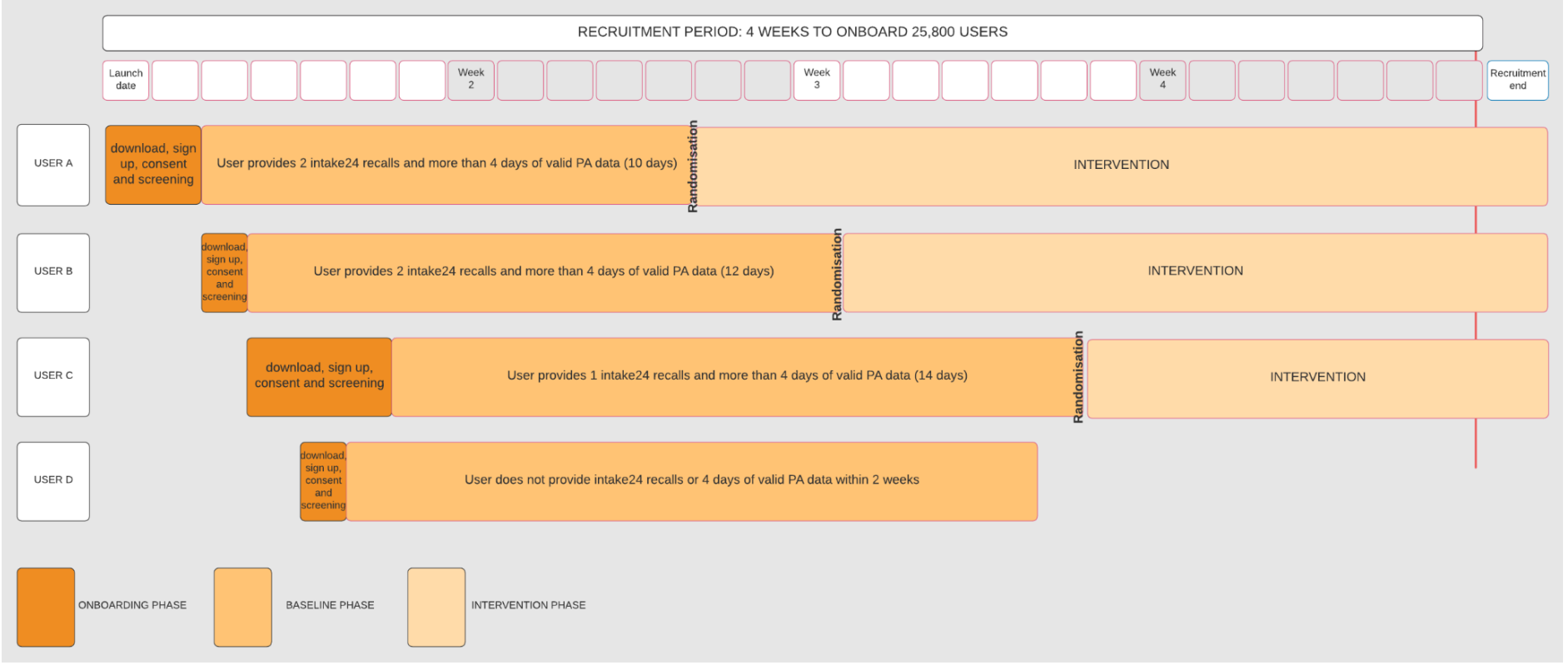
based on their full addresses (including postcode).

2. If YES, the algorithm assigns this user to the same trial arm
3. If NO, the algorithm generates a random number from 1-100 (inclusive) for the user.
4. Based on the rule described in Figure 2, users are assigned to one of the 4 trial arms. Note that the control arm will be larger than the other arms. This is because at the analysis stage (5m mark), we aim at having 3 users in the control group for each user in the treatment arms. However, we expect attrition to differ across control and treatment groups, as shown in Table 6 in Section 2.4.2.4.. Based on these attrition estimates, at randomisation 12% of users will be assigned to the high incentive arm, 12% to the medium incentive arm, 15% to the low incentive arm, 61% to the control arm.

BIT's in-house IT expert will quality assure the code of the randomisation algorithm developed by HUL before the trial launch.

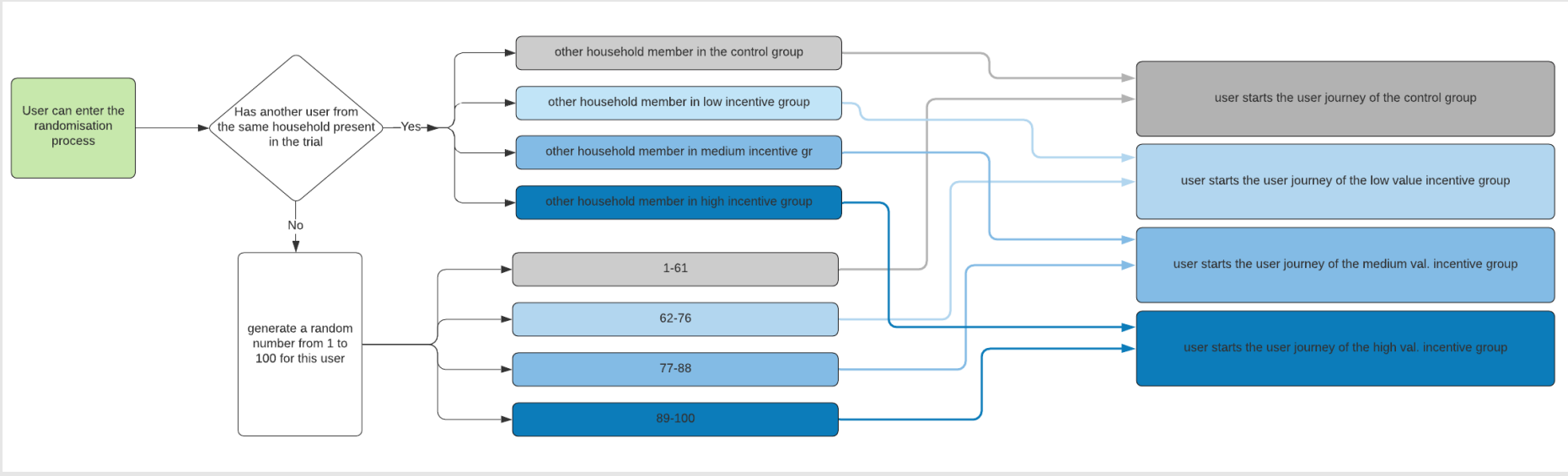
Figure 1. Onboarding and baseline phase from a user's point of view[16].

User A *downloads the app on trial launch date, their onboarding phase lasts 2 days and their baseline phase lasts 10 days (as they provide the required data within 10 days), hence they enter randomisation on day 12 after trial launch day.* ***User B*** *downloads the app three days on day 3 after the trial launch day, their onboarding phase lasts 1 day and their baseline phase lasts 12 days (as they provide the required data within 12 days), hence they enter randomisation on day 15 after trial launch day.* ***User C*** *downloads the app on day 4, their onboarding phase lasts 3 days and their baseline phase lasts 14 days (as in 2 weeks they provide 1 dietary recall instead of 2 and at least 5 days of PA), hence they enter randomisation on day 20.* ***User D*** *downloads the app on day 5, their onboarding phase lasts for 1 day and their baseline phase lasts14 days, at the end of which they have not provided the required data, hence they cannot enter randomisation and their user journey ends.*



[16] In Figure 1, the recruitment period lasts for 4 weeks, which is the best-case scenario. This may be extended to 8 weeks if needed to achieve the target sample size.

Figure 2. Randomisation process



2.3 Outcomes

We propose three types of outcomes for our impact evaluation: primary, and exploratory (see Table 3).

The primary outcomes consist of **2 physical activity outcomes** and **4 dietary outcomes**.

- The primary physical activity outcomes include: (i) moderate and vigorous physical activity (MVPA) in minutes per day and (ii) daily steps measured objectively through a wearable device.

- The primary dietary outcomes include (i) fruit and vegetables (g/day) and (ii) fibre (g/day), (iii) saturated fat (% of energy from food), and (iv) free sugars (% of energy from food) measured through two Intake24 questionnaires per measurement time point.

The secondary outcomes include (i) daily energy expenditure as measured by the wearable, (ii) daily energy intake as measured by 24-hour dietary recalls (Intake24 surveys), (iii) a healthy eating score based on consumption of key food groups, macro- and micronutrients, and (iv) self-reported weight.

The exploratory outcomes include unintended health impact of the interventions, such as (i) sleep quality and (ii) mental well-being. It will also include motivation to change (iii) PA and (iv) diet.

Table 3. Outcome measures

Outcomes	Category	Metric	Collection method	Data collection point	Analysed
Primary	PA	• MVPA (min/day) • Daily steps	Wearable device	Throughout the full duration of the pilot	Effect at 1m, 3m, 5m marks
	Diet	• Fruit and vegetables (g/day) • Fibre (g/day) • Free sugars (% daily food energy) • Saturated Fat (% daily food energy)	24h recall survey (Intake24)	Baseline, 1m, 3m, 5m marks	Effect at 1m, 3m, 5m marks
Secondary	PA	• Energy expenditure (kcal/day)	Wearable device	Throughout the full duration of the pilot	Effect at 5m mark
	Diet	• Energy intake (kcal/day)	24h recall survey (Intake24)	Baseline, 1m, 3m, 5m marks	Effect at 5m mark

		• A healthy eating score based on consumption of key food groups, macro- and micronutrients, (see Section 2.3.1.2 for details)			
	Weight	Weight (kg)	Self-reported survey	Baseline, 1m, 3m, 5m marks	Effect at 5m mark
Exploratory	Motivation to change	Motivation to change diet	In-App survey	Baseline, 5m marks	Effect at 5m mark
		Motivation to change PA	In-App survey	Baseline, 5m marks	Effect at 5m mark
	Unintended impact on mental and physical well-being	Self-reported mental well-being measured by WHO-5 Index (0~100, 0 = worst, 100 =best)	In-App survey	Baseline, 1m, 3m, 5m marks	Effect at 5m mark
		Sleep quality (hours/day)	Wearable device	Throughout the full duration of the pilot	Effect at 5m mark

2.3.1 Primary outcomes

We propose two categories of primary outcomes for our impact evaluation - outcomes that relate to PA and outcomes that relate to diet. The overarching rationale for selecting these outcomes is focusing on behaviours that are directly encouraged by the in-app challenges - please see HUL's intervention design plan and the complementary Theory of Change note for further detail.

2.3.1.1 Primary outcome - physical activity

The proposed primary outcomes for PA include: (i) moderate and vigorous physical activity (MVPA) in minutes per day and (ii) daily steps. We chose those two outcomes chiefly because they have shown to provide clinical health benefits.

MVPA. We focus on daily MVPA minutes as one of the PA primary outcomes because cumulative MVPA mins have shown to provide various clinical health benefits (e.g. reducing BMI, cholesterol level, and blood pressure)[17]. Additionally, a meta-analysis of interventions to reduce sedentary time showed that both light-intensity physical activity (e.g. walking) and MVPA increased energy expenditure

[17] Hajna, S., Ross, N. A., & Dasgupta, K. (2018). Steps, moderate-to-vigorous physical activity, and cardiometabolic profiles. *Preventive Medicine*, *107*, 69–74. doi: 10.1016/j.ypmed.2017.11.007

in adults but interventions increasing MVPA had a larger effect on energy expenditure than those focused on light-intensity activities [18]. Currently, the CMO's Physical Activity Guidelines recommends at least 150 minutes of MVPA per week for adults for good health.

Steps. We include daily steps as another primary outcome for three reasons. First, daily steps, like MVPA, have been widely used in previous similar studies and including it would increase the comparability of our results. Second, daily steps have also shown health benefits: a recent study[19] has found that a greater number of steps per day was significantly associated with lower all-cause mortality among U.S. adults. Third, daily steps are a good proxy for the degree to which one has led a sedentary lifestyle, a risk that has been highlighted by the CMO's Physical Activity Guidelines.

These two outcomes are also directly incentivised by the app. Some incentives will be tied to a 'Let's get moving' (Do ((minutes goal)) minutes ((days goal)) times each week) or 'Step it up' challenge ((Do at least (goal steps) per day) (see the *Theory of Change* note for details), making these two outcomes good candidates to measure the direct effects of the financial interventions.

The physical activity data is passively collected. All devices, including the device provided by HUL, can detect PA metrics (heart rates, step counts, MVPA, calories and sleep) automatically without the user's input. However, for HUL to capture this data, the device needs to be synced with the app on a regular basis.

Daily steps are directly measured by the wearable. MVPA is calculated as the sum of vigorous and moderate activities. All devices define a PA to be "vigorous" if the user's heart rate falls within the cardio or peak heart rate zones. Most of the devices auto-classify a "moderate" PA if two conditions are met: (1) the heart rate is within fat-burning heart rate zones; (2) sufficient movements detected by accelerometers built in the wearables.

2.3.1.2 Primary outcome - diet

As dietary change evokes a spectrum of changes, the study will use four primary outcomes for diet.[20]

---

[18] Biswas A, Oh PI, Faulkner GE, Bonsignore A, Pakosh MT, Alter DA. The energy expenditure benefits of reallocating sedentary time with physical activity: a systematic review and meta-analysis. Journal of Public Health. 2018 Jun 1;40(2):295-303.

[19] Saint-Maurice, P. F., Troiano, R. P., Bassett, D. R., Graubard, B. I., Carlson, S. A., Shiroma, E. J., … Matthews, C. E. (2020). Association of Daily Step Count and Step Intensity With Mortality Among US Adults. *JAMA*, *323*(12), 1151–1160. doi: 10.1001/JAMA.2020.1382

[20] In previous versions of the evaluation protocol, calorie intake had been selected as the primary outcome for diet. The DOPA/OHID expressed concerns around use of kcal as the primary outcome measure for diet, in particular around i) underreporting, and ii) the fact that the pilot is not a weight loss programme and so there was the concern of not being able to detect a change in kcals even if the pilot did improve participants' diet. Following further conversations with the DOPA/OHID, BIT and DHSC agreed on the list of primary outcomes included in the current evaluation protocol, that reflect behaviours that are directly incentivised by the programme.

These include:

- Fruit and vegetables (g/day)
- Fibre (g/day)
- Saturated fat (% daily food energy[21] intake)
- Free sugars (% daily food energy intake)

Evidence suggests that people typically consume insufficient amounts of fruit and vegetables and fibre whilst overconsuming saturated fat and free sugars.[22] (Also see Appendix C for an overview of the recommended dietary intake **according to UK dietary recommendations**).

Tackling the primary outcomes above, might benefit health:

- **Fruit and vegetable consumption**: The health benefits of increased fruit and vegetable consumption include reduced risk of cardiovascular disease, cancer, and all-cause mortality.[23] [24] A recent meta-analysis also provided moderate certainty evidence that consuming fresh fruit promotes weight maintenance or modest weight loss over periods of 3–24 weeks.[25] This is in line with other evidence suggesting increased fruit and vegetable consumption can support maintenance of a healthy weight.[26] [27]
- **Fibre consumption:** Consumption of whole-grain and dietary fibre has been associated with lower mortality, and lower risk of CVD, obesity, and diabetes.[28] [29]
- **Consumption of free sugars and saturated fat.** Saturated fat and free sugars contribute to a range of chronic conditions such as cardiovascular disease, Type 2 diabetes, and obesity.[30] [31]

---

[21] Food energy intake is defined as total energy intake minus energy intake from alcohol consumption.
[22] Scheelbeek P, Green R, Papier K, Knuppel A, Alae-Carew C, Balkwill A, Key TJ, Beral V, Dangour AD. Health impacts and environmental footprints of diets that meet the Eatwell Guide recommendations: analyses of multiple UK studies. BMJ open. 2020 Aug 1;10(8):e037554.
[23] Aune et al., 2017. Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality—a systematic review and dose-response meta-analysis of prospective studies, International Journal of Epidemiology, 46 (3): 1029–1056
[24] Wang et al., 2014. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies BMJ; 349 :g4490
[25] Guyenet SJ. Impact of Whole, Fresh Fruit Consumption on Energy Intake and Adiposity: A Systematic Review. Front Nutr. 2019;6:66
[26] Ledoux TA, Hingle MD, Baranowski T: Relationship of fruit and vegetable intake with adiposity: a systematic review. Obes Rev. 2011, 12: e143-e150.
[27] Mytto et al., 2014. Systematic review and meta-analysis of the effect of increased vegetable and fruit consumption on body weight and energy intake. BMC Public Health 14, 886
[28] Smith, C., & Tucker, K. (2011). Health benefits of cereal fibre: A review of clinical trials. Nutrition Research Reviews, 24(1), 118-131.
[29] Barber TM, Kabisch S, Pfeiffer AFH, Weickert MO, 2020. The Health Benefits of Dietary Fibre. Nutrients. 2020;12(10):3209.
[30] Hooper L, Martin N, Jimoh OF, Kirk C, Foster E, Abdelhamid AS. Reduction in saturated fat intake for cardiovascular disease. Cochrane database of systematic reviews. 2020(8).
[31] Scientific Advisory Committee on Nutrition. Carbohydrates and Health. 2015. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/445503/SACN_Carbohydrates_and_Health.pdf

2.3.2 Secondary outcomes

The overarching rationale for selecting these secondary outcomes is capturing more holistic, broader effects of offering financial incentives on users' behaviours. There are three categories of secondary outcomes: PA, diet, and weight.

2.3.2.1 Secondary outcome - physical activity

We will focus on one secondary PA outcome to supplement the primary PA outcome, and it will also be collected via the wearable device:

Energy expenditure. Energy expenditure is defined as daily calories (kcal/day) burnt from physical activity. Since MVPA doesn't take into account lighter forms of physical activity (e.g. walking) and steps are only one specific form of activity, we included energy expenditure as a secondary outcome as a holistic measure of physical activity.

2.3.2.2 Secondary outcome - diet

We will supplement the four primary dietary outcomes two secondary outcomes:

Energy intake (kcal/day). Given the high prevalence of overweight and obesity and the role of energy intake in the aetiology of obesity, we will assess the impact of the scheme on participants' energy intake. [32]

Healthy eating score

We will derive a healthy eating score based on consumption of key food groups, macro- and micronutrients following the methodology by Scheelbeek et al. 2020[33] The rationale to include this score is twofold:

- First, this holistic measure enables us to better capture the combined effects of the 10 health challenges, each of which targets different food groups.

- Second, it allows us to capture the intervention impact on diet aspects that are not covered by the primary outcomes (e.g. salt) without unduly increasing the number of individual outcomes assessed by the study.

- The scheme allows us to holistically capture intervention effects on metrics that:
    - are closely embedded in the ToC we proposed for dietary intake
    - are tied with at least one incentive
    - have been used in studies with similar contexts, which increases comparability of results and enables easier synthesis of evidence

[32] Health Survey for England 2019. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2019#summary

[33] Scheelbeek P, Green R, Papier K, Knuppel A, Alae-Carew C, Balkwill A, Key TJ, Beral V, Dangour AD. Health impacts and environmental footprints of diets that meet the Eatwell Guide recommendations: analyses of multiple UK studies. BMJ open. 2020 Aug 1;10(8):e037554.

- are linked with a clear government dietary recommendations for healthy eating

Following Scheelbeek et al. 2020[34] we will score participants' diets measured with Intake24 using a dichotomous system to assess whether the diet meets each of the following criteria:

Table 4. Intake24 Criteria

Metric	**Criteria for point allocation**	**Following methodology by**
Fruit and vegetables (g/day)	≥ 400 g	Scheelbeek et al. 2020
Red & processed meat (g/day)	≤ 70g	Scheelbeek et al. 2020
Free sugars (g/day)	≤ 30 g	Scheelbeek et al. 2020
Saturated fat (g/day)	For males ≤ 30 g For females ≤ 20 g	Scheelbeek et al. 2020
Fibre (g/day)	≥ 30 g	Scheelbeek et al. 2020
Total fat (g/day)	For males ≤ 97g, ≤ 91g, ≤ 89 g, respectively for age groups 18-64, 65-74, 75+ For females ≤ 78g, ≤ 74g, ≤ 72g, respectively for age groups 18-64, 65-74, 75+	Scheelbeek et al. 2020
Salt (g/day)	≤ 6 g	Scheelbeek et al. 2020

For each metric, participants will score 1 if their consumption meets the corresponding criteria for point allocation, and 0 if they don't. Since we will measure diets with 24 hours dietary recalls and recommendations for fish are expressed on a weekly basis, we decided to drop the metrics pertaining to (i) oily fish and (ii) other fish consumption, which were used in the original index by Scheelbeek et al. 2020.[35] Since there are 7 metrics in total, the value range of the score varies from 0 to 7 (0 = least healthy, 7 = most healthy).

2.3.2.3 Secondary outcome - weight

Although the financial incentive scheme is not a weight loss app, measuring weight is important to help assess whether this healthy eating and physical activity intervention could also help to reduce the prevalence of obesity or indeed to rule out potential unintended consequences on weight.

[34] Scheelbeek P, Green R, Papier K, Knuppel A, Alae-Carew C, Balkwill A, Key TJ, Beral V, Dangour AD. Health impacts and environmental footprints of diets that meet the Eatwell Guide recommendations: analyses of multiple UK studies. BMJ open. 2020 Aug 1;10(8):e037554.

[35] Scheelbeek P, Green R, Papier K, Knuppel A, Alae-Carew C, Balkwill A, Key TJ, Beral V, Dangour AD. Health impacts and environmental footprints of diets that meet the Eatwell Guide recommendations: analyses of multiple UK studies. BMJ open. 2020 Aug 1;10(8):e037554.

Participants will be asked to self-report their current weight at the baseline and each measurement timepoint. The app will encourage participants to weigh themselves when entering weight data.

2.3.3 Exploratory outcomes

The primary outcomes focus on participants' behaviours. To understand whether the financial incentive scheme affects participants' motivation to change behaviours as well as potential unintended consequences, we propose two types of exploratory outcomes.

2.3.3.1 Exploratory outcomes - Motivation to change

To understand whether the financial incentive scheme affects participants' motivation to change behaviours in PA and diet, we propose two exploratory outcomes to capture participants' motivation to change before and after the intervention. We will use a 5-point Likert scale (1=not at all confident/willing to 5=very confident/willing with 3 being neutral) for each question, and average across the two items:

- Motivation to change physical activity:
    - How willing are you to make changes in your physical activity habits in order to be healthier in the next 6 months?
    - How confident are you in making these changes to your physical activity habits in the next 6 months?

- Motivation to change diet:
    - How willing are you to make changes in your eating habits in order to be healthier in the next 6 months?
    - How confident are you in making these changes to your eating habits in the next 6 months?

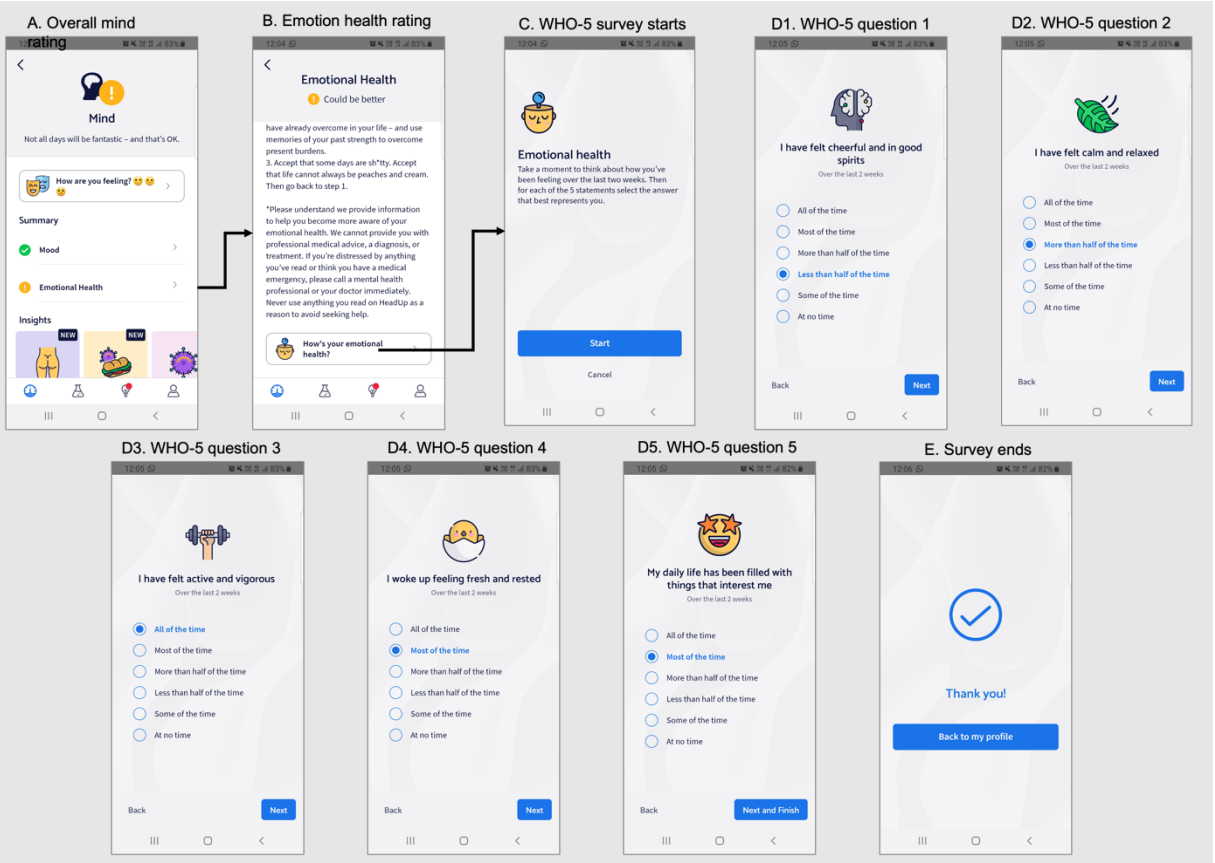
We will collect these measures at baseline and m5, and examine the impact of offering incentives compared to not offering any incentives.

2.3.3.2 Exploratory outcomes - unintended consequences

We propose two exploratory outcomes to investigate the unintended effects of the interventions: **mental health, and sleep duration**.

First, as suggested by a recent meta-analysis, increased levels of physical activity and healthier diet might improve people's **mental health**. We propose to measure participants' well-being as an exploratory outcome to capture the potential effects of the interventions beyond physical health (positive or negative). Mental wellbeing will be collected at baseline and key collection points (m1, m3, m5) by prompting users via push notifications and in-app prompts to assess their mental health using the WHO-5 WHO-5 Well-being questionnaire to measure people's mental well-being.

The WHO-5 is a simple and validated index consisting of five statements, which respondents rate (in relation to the past two weeks) according to the screenshots below.



For each statement, participants choose one of six frequencies and their answers will be coded into a numeric value (all of the time = 5, at no time = 0). The total raw score, ranging from 0 to 25, is multiplied by 4 to give the final score, with 0 representing the worst well-being and 100, the best.

Second, we will analyse effects on **sleep duration** (as measured by the duration of sleep (hour/day) by the wearable).[36] We use sleep duration as potential unintended consequence because physical activity could improve sleep quality in general[37], but it also could impair sleep quality if participants do vigorous physical activity within 1 hour before bedtime[38].

[36] HUL will prompt users to wear the device at nights, but as a secondary call-to action (i.e. prompt once in the app at key collection points (baseline, m1, m3, m5).
[37] Banno M, Harada Y, Taniguchi M, et al. Exercise can improve sleep quality: a systematic review and meta-analysis. *PeerJ*. 2018;6:e5172. Published 2018 Jul 11. doi:10.7717/peerj.5172
[38] Stutz J, Eiholzer R, Spengler CM. Effects of Evening Exercise on Sleep in Healthy Participants: A Systematic Review and Meta-Analysis. Sports Med. 2019 Feb;49(2):269-287. doi: 10.1007/s40279-018-1015-0. PMID: 30374942.

2.4 Sample size and minimum detectable size

2.4.1 Summary

The trial will adopt **a household-level randomisation,** with a target sample size of **around 4,200 participants** (2,100 in the control arm, 700 per intervention arm) at the final data collection point **(5 months after the randomisation)** to be powered to detect effect sizes deemed substantive from previous literature (see Section 2.4.2.2 Table 5)**.** We expect users to come from ~2,300 households based on the demographic profile of Wolverhampton LA (see Appendix B).

To achieve a post-attrition sample of 4,200 participants, HUL expects to recruit a sample of around **25,800 participants** at the beginning of the baseline stage, given the attrition rate at various trial stages estimated by HUL (see Section 2.4.2.3 Table 6).

BIT has conducted power calculations for the primary research question *"Does the incentive scheme improve physical activity (PA) and diet at the five month mark?".* The power calculations focus on the primary outcomes. The primary analysis pools all the treatment arms together and assesses their overall difference against the control.

Given the post-attrition sample size of 4,200 participants, the minimum effect sizes we are powered (at 80% level) to detect at the 5-month's mark in the primary analysis are:

Physical activity primary outcomes:

- Daily MVPA (min/day): 2.4 min/day
- Daily steps: 371 steps/day

Dietary primary outcomes:

- Fruit and vegetables (g/day): 18.7 g/day
- Fibre (g/day): 0.74 g/day
- Free sugars (% food energy): 0.63%
- Saturated Fat (% food energy): 0.31%

The following subsections provide details on the key assumptions and decisions made concerning the power calculation. First, we provide the minimum required sample size at the analysis stage (post-attrition) given the target effect sizes for the primary outcomes (see Table 5). Then, we show the minimum required sample size at the recruitment stage (pre-attrition) given the predicted attrition rate at various stages (see Table 6) based on HUL's estimations. Last, we calculate the minimum detectable effect size for primary outcomes (see Table 7) based on the target sample size confirmed as feasible for this project by HUL and DHSC.

For further information on specific assumptions behind household-level randomisation, multiple-comparisons-adjustment, inclusion of covariates, and allocation ratio among trial arms, please see Appendix B.

2.4.2.1 Target effect sizes we aim at detecting

In this subsection, we select one PA outcome (MVPA min/day) and one dietary outcome (fruit and vegetable intake g/day) to illustrate how we estimate the target effect sizes from the existing literature.

For MVPA minutes per day, we use an estimate of the standard deviation (SD) of MVPA minutes based on HeadUp's existing UK users who signed up in 2019, similar in demographic and socioeconomic status to those of the target groups (SD = 26 min/day, n = 19,898 users). A systematic review[39] of RCTs testing the impact of financial incentive intervention on PA suggested a wide range of effect sizes (from 0.02 to 0.4 SD units). We consider the midpoint of this range to be a plausible estimate of expected effect size, therefore aim to have a **minimum detectable effect size (MDES) of 0.2 SD units, or a difference of 5.2 MVPA min/day (36 MVPA min/week)** between intervention (any incentive) and control (no incentive) groups.

For daily fruit and vegetable intake (g/day), we obtained an SD estimate from the National Diet and Nutrition Survey data (NDNS, 2020)[40], corresponding to adults from England, 19-64 years old (SD = 190g /day, n = 475). Previous meta-analysis[41] suggests that the effect size of interventions designed to increase adults' consumption of fruit and vegetables is in the 0.1 ~ 1.2 serving / day range (corresponding to 8 ~ 112 g/day). **We assumed a more modest target effect size of 0.11 SD units, or a difference of 21 g/day.**

[39] Mitchell et al. (2020)
[40] Public Health England. (2021). National Diet and Nutrition Survey: Diet, nutrition and physical activity in 2020 - a follow-up study during COVID-19. Retrieved from https://www.gov.uk/government/statistics/ndns-diet-and-physical-activity-a-follow-up-study-during-covid-19
[41] Pomerleau, J., Lock, K., Knai, C., & McKee, M. (2005). Interventions Designed to Increase Adult Fruit and Vegetable Intake Can Be Effective: A Systematic Review of the Literature. *The Journal of Nutrition*, *135*(10), 2486–2495. doi: 10.1093/jn/135.10.2486

Table 5. Minimum sample size required at the analysis stage (post-attrition) given the target effect sizes from literature review.

Summary	**Physical activity**		**Diet**			
Primary outcome measure	MVPA minutes per day	Daily steps	Daily fruit and vegetable intake	Daily fibre intake	Free sugars (% food energy)	Saturated Fat (% food energy)
Standard deviation data source	n = 19,898 UK users, whose demographic and obesity profiles are most similar to the target population of this trial[42] *(HeadUp Labs, 2019)*		n = 475 UK adults (19-64 years old) [43] *(National Diet and Nutrition Survey 2020)*			
Standard deviation estimate	26 min/day	4,096/day	190 g/day	7.5 g/day	6.3 %	3.2 %
Effect size range from prior literature	0.02 to 0.4 SD units[44]		0.04 to 0.58 SD units[45]			
Target minimum detectable effect size in SD (MDES)	0.2 SD units		0.11 SD unit			
Target MDES in absolute terms *(note: these are all larger than the effect size we estimate to be powered for)*	5.2 MVPA min/day	819 steps/day	21 g/day	0.83 g/day	0.69 %	0.35 %
Assumed attrition rate	See Table 7 (Section 2.4.2.3)					

[42] We only included user whose age profile matches that of the target population, i.e. 18-65 years old, and 69% of the included users were overweight or obese, highly comparable to the proportion in Wolverhampton, which is 70% according to the UK's latest obesity statistics report.
[43] Public Health England. (2021). National Diet and Nutrition Survey: Diet, nutrition and physical activity in 2020 - a follow-up study during COVID-19. Retrieved from https://www.gov.uk/government/statistics/ndns-diet-and-physical-activity-a-follow-up-study-during-covid-19
[44] Mitchell et al. (2020). Financial incentives for physical activity in adults: systematic review and meta-analysis. *British Journal of Sports Medicine*, 54(21), 1259-1268.
[45] Pomerleau, J., Lock, K., Knai, C., & McKee, M. (2005). Interventions Designed to Increase Adult Fruit and Vegetable Intake Can Be Effective: A Systematic Review of the Literature. *The Journal of Nutrition*, *135*(10), 2486–2495. doi: 10.1093/jn/135.10.2486

Adjustment for baseline covariates	Corr. coef = 0.25-0.75 (0.5 as most realistic)
Target power	80%
Adjusting for multiple comparisons	Bonferroni adjustment[46] for primary outcomes by category (one comparison each of all incentive arms [pooled] vs. control; the number of adjustments is 2 for PA activities and 4 for dietary outcomes)
Allocation ratio across arms post-attrition	3 participants in the control arm for every 1 participant in each intervention arm
Cluster size assumptions for household-level randomisation	Sample will comprise: 39% 1-adult HHs, 41% 2-adult HHs, and 20% 3-adult HHs, with all adults in a HH signing up
Assumed ICC (more details in Appendix B)	0.2-0.5 (0.2-0.4 as more realistic)
Number of **overall users** required for smaller dietary outcome target MDES at **the 5m mark analysis stage**	**4,200** (2,100 in control; 700 per intervention arm)

[46] The Benjamini-Hochberg false discovery adjustment will be applied during analysis of primary outcomes, but as it is not possible to apply this before having the data, the more conservative Bonferroni adjustment provides an upper bound on power here.

2.4.2.2 Minimum required sample size at the analysis stage

Given the target effect size for PA and dietary outcomes, and under the assumptions of household ICC of 0.3, a baseline covariate correlation coefficient of 0.5, we estimate that achieving the targeted MDES requires a post-attrition sample size of **4,200 users** (**2,100** in the control arm, **700** per intervention arm) at the final data collection point, i.e. 5 months after the randomisation (see Table 6 for details).

2.4.2.3 Minimum required sample size at the participant recruitment stage

The minimum required sample size at the analysis stage is not necessarily the sample we need at the participant recruitment stage. An important consideration is the proportion of participants who are expected to drop out of the study post-randomisation and for whom we will not have any outcome data (i.e. *the attrition rate*). To have a well-powered sample for analysis requires setting recruitment targets that are substantially higher to account for this attrition during the trial. In their acquisition plan, HUL estimated the attrition rates by treatment conditions (see Table 6 for details).

Table 6. Estimated sample size required at various trial stages given the corresponding estimated attrition rate

Trial arm	Estimates								
	N at the beginning of the baseline period	N at the end of the baseline period (at randomisation)	Attrition rate at 1m mark	**N needed at 1m mark**	Attrition rate at 3m mark	**N needed at 3m mark**	Attrition rate at 5m mark	**N needed at 5m mark**	Attrition from the beginning of the baseline period to the 5m mark
High incentive	**~3,100**	~1,900	18%	1,550	44%	1,050	62%	**700**	77%
Medium incentive	**~3,200**	~1,950	18%	1,600	45%	1,100	63%	**700**	78%
Low incentive	**~3,800**	~2,250	21%	1,800	50%	1,150	69%	**700**	81%
Control group	**~15,700**	~9,400	26%	6,950	60%	3,800	78%	**2,100**	87%
Total	***~25,800***	*~15,500*	-	***~11,900***	-	***~7,100***	-	***4,200***	-
Note: Attrition rate between baseline and randomisation: 40% for all groups									

This is consistent with systematic reviews of digital health interventions showing attrition rates of up to 80% in some RCTs (*average* attrition is 40-50%), and a common indication of differential attrition between trial arms.[47]

Based on the estimated attrition rate, HUL and DHSC agreed on a target sample size of **25,800** (3,100 in the high incentive group; 3,200 in the medium incentive group; 3,800 in the low incentive group; 15,700 in the control group) **at the beginning of the baseline phase.**

[47] Meyerowitz-Katz et al. (2020); Howarth et al. (2018); Beleigoli et al. (2019).

2.4.2.4 Minimum detectable effect size given the sample size at analysis stage

With the sample size at the final data collection point (i.e. post-attrition, n = 4,200), we present the minimum effects we are powered (at 80% level) to detect for six primary outcomes among the trial participants in Table 7, which also provides corresponding benchmarks to interpret those effect sizes.

Table 7. Minimum detectable effect size for primary outcomes given a post-attrition sample size of 4,200

Primary outcomes		MDES	MDES as % of min. (or max.) recommended intake
PA	Daily MVPA (min/day)	2.4 min/day	~ 10% (150 min/week, according to the UK CMO's Physical Activity Guidelines)
	Daily steps	371 steps/day	~ 4% (8,000 steps/day, according to the UK CMO's Physical Activity Guidelines)
Diet	Fruit and vegetables (g/day)	18.7 g/day	~ 5% (~ a quarter of a F&V portion)
	Fibre (g/day)	0.74 g/day	~ 2.5% (~ a third of an apple)
	Free sugars (% food energy)	0.63%	~ 13% of the upper limit of % food energy intake (≤5% food energy)
	Saturated Fat (% food energy)	0.31%	~ 2.8% of the upper limit of % food energy intake (≤11% food energy)

2.5 Assignment of interventions

2.5.1 Allocation

To collect comparable baseline data between trial arms, all participants will be able to order a free wearable device if they do not own one, install the pilot app, and initially have access to only the control group version of the app. After baseline data is collected, users will be randomly assigned to one of the four trial arms. Those in the intervention arms would gain access to the incentives and other app features; those in the control arm would have access to the app features but not be rewarded with incentives. All individuals in the same household will be allocated to the same trial arm.[48]

At the analysis stage (end of the trial), we aim at having 3 users in the control group for each user in the treatment arms. However, we expect attrition to differ across control and treatment groups, as shown in Table 6.

[48] Please note that this does not mean that we are aiming at recruiting a minimum number of households, or that some households will be excluded from the trial, and it won't be a requirement that all household members sign up. This simply means that whenever more than one user belonging to the same household sign up, they will all be allocated to the same trial arm.

Based on these attrition estimates, Table 8 shows the optimal allocation of users to trial arms at randomisation.

Table 8. Optimal allocation of users to trial arms at randomisation

Group	Estimates	
	N at the end of the baseline period (at randomisation)	Allocation ratio
1. High incentive	~1,900	12%
2. Medium incentive	~1,950	12%
3. Low incentive	~2,250	15%
4. Control group	~9,400	61%

2.5.2 Blinding (masking)

Participants will be aware they are participating in a randomised controlled study, as they will be asked to provide consent. However, we expect that users will not be able to figure out to which treatment arm they are assigned to (low, medium or high level of incentive) based on the information available on the app and their user journey, whilst it is likely that users in the control group will be able to correctly identify that they have been assigned to the control trial arm (no incentive).

While is not possible to deliver a fully blind design for trial analysts given the differences that will be identifiable in the data at the point of analysis, datasets for each arm will be labelled neutrally (e.g. A, B, C and D instead of treatment, control).

2.6 Data collection, management, and analysis

Attrition from the trial (users dropping off from data collection) is the largest threat to the success of the evaluation and knowing whether a financial incentive scheme is effective within this context.

The next section elaborates in details the data collection methods, including:

- How to collect data and time points of data collection;
- How we plan to adjust for differential attrition across trial arms;
- How to mitigate the risk of differential attrition.

2.6.1 Data collection methods

2.6.1.1 Assessment and collection of data

Physical activity outcomes. The PA measures will be automatically recorded through the wearable device for the full duration of the pilot, but it requires users to sync their device with the app for HUL to collect this data. Users will be reminded to wear and sync the device during crucial data collection weeks/moments. Participants will be nudged (via push notifications) to wear and synchronise the wearable device during the trial period to improve accuracy of data collection.

Dietary outcomes[49]. Diet will be measured using Intake24, a retrospective 24 hour dietary recall questionnaire.[50] Completing this questionnaire can take up to 20 minutes. We follow the approach taken by the The Scottish Health Survey and encourage two recalls at each data collection point.[51] Users may provide a recall for a weekday or a weekend day, based on when they complete the questionnaire.[52]

Measurement time points across the pilot. Figure 3 also provides a high-level user journey for participants, from an evaluation point of view, to illustrate when users are asked to provide data throughout the trial. Broadly speaking, as illustrated in Table 9**,** we envisage 4 data collection points during the study:

- the baseline period;
- the 5th week after randomisation for the 1m mark;
- the 13th week after randomisation for the 3m mark;
- the 21st week after randomisation for the 5m mark.

Additional measurement time point at the 12m mark. In addition to the 4 data collection points, we might add another time point at the 12m mark if the pilot programme is extended by DHSC. If the extension takes place, an updated evaluation plan and ethics approval will be sought before data collection. This extension would be subject to contractual agreement.

---

[49] In Appendix B we set out the other options for measuring diets that we considered and the rationale that led us to select Intake24 over other alternatives.

[50] In using Intake24 questionnaire, we will strike a balance between maximising the validity and reliability of the method (Intake24 is used for the National Diet and Nutrition Survey) whilst minimising the burden on participants (more robust measures of dietary intakes, such as food diaries would be much more time-consuming). Based on previous studies, the user experience of completing Intake24 is generally positive. Feedback from participants in the field test of Intake24 suggests that the majority of people found the system user-friendly, enjoyable to use, and felt it accurately captured their diet.

[51] It is our assessment that two recalls is a reasonable mid-point between the single administration recommended by the National Cancer Institute and the four recalls used by PHE's NDNS and in the original validation study. Four recalls would force us to extend the sign-up onboarding period of at least one week, with significant risks to the retention of users in the programme as it would involve asking participants to complete 80 minutes of questionnaire at each measurement time point.

[52] The Intake24 team at Cambridge reassured BIT that it is not necessary that each user provides a recall for a weekday and one for the weekend for obtaining robust results. This is because it is only necessary that across the whole sample some users will provide some recalls for weekdays and some for weekends. As different users are nudged to provide a recall on different days, based on when they begin the onboarding phase, it is reasonable to expect that overall some recalls will reflect weekday dietary behaviour and some recalls weekend dietary behaviour.

Table 9. Data collection points and methods

Data collection point	Data collection method for dietary outcomes	Data collection method for PA outcomes
Baseline period	**Day 1 of baseline**: users are invited to provide their first recall. Users are informed that the first two recalls they will provide (to be provided on two separate days, including information about weight) will be remunerated with a payment of £5 each in points. The second intake24 is optional; users are encouraged to complete the second one within 3 days. The survey will be prominently displayed on the dashboard of the app when users open it for the full duration of the baseline period. *In the analysis, we will use all dietary recalls submitted during the baseline period.*	If a user owns a wearable device, they are encouraged to sync their own device with the app and provide their PA data from the first day of the baseline period. If a user orders a wearable device, they are encouraged to sync it with the app and provide their PA data as soon as they receive the device. Users will be nudged for the full duration of their baseline period to wear the device and periodically sync it with the app. *In the analysis, we will use all valid daily PA data provided during the baseline period (valid = device worn for at least 6h)*
1 month mark (starting the 5th week after randomisation)	**Day 1:** users are invited to provide a recall. Users are informed that the first two recalls they provide in the following two weeks will be remunerated with a payment of £5 each in points. Reminders, feedback and survey prominence are repeated as per the baseline phase. *In the analysis, we will use all recalls submitted during the 5th and 6th week after randomisation for the 1m mark.*	In the 5th week after randomisation, users will be nudged to wear the device as long as possible for 7 days and sync the device with the app. *In the analysis, we will use all valid daily PA data provided during the 5th week after randomisation for the 1m mark. If there are missing values in that week, we will impute them with valid values (if available) on the same day of the week within 2 adjacent weeks (i.e. weeks 3-7).*
3 months mark (starting the 13th week after randomisation)	This will work as at the 1 month mark. *In the analysis, we will use all recalls submitted during the 13th and 14th week after randomisation for the 3m mark.* *If the 'delay' and 'higher incentive' mitigation strategies are triggered: Day 1: users are invited to provide a recall. Users are informed that*	In the 13th week after randomisation, users will be nudged to wear the device as long as possible for 7 days and sync the device with the app. *In the analysis, we will use all valid daily PA data provided during the 13th week after randomisation for the 3m mark. If there are missing values in that week, we'll replace them*

	the first two recalls they provide in the following three weeks will be remunerated with a payment of £10 each. Reminders, feedback and survey prominence are repeated as per the baseline phase. In the analysis, we will use all recalls submitted during the 13th, 14th and 15th week after randomisation for the 3m mark (days 90 to 111 post randomisation).	*with valid values (if available) on the same day of the week within 2 adjacent weeks (i.e. weeks 11-15).* *If the mitigation strategies are triggered:* *In the analysis, we will use all valid daily PA data provided during the 13th, 14th and 15th weeks after randomisation for the 3m mark. If there are missing values in that week, we'll replace them with valid values (if available) on the same day of the week within 2 adjacent weeks (i.e. weeks 11-17).*
5 month mark (starting the 20th week after randomisation)	This will work as at the 1 month mark. *In the analysis, we will use all recalls submitted during the 21st and 22nd week after randomisation for the 5m mark.* *If the 'delay' and 'higher incentive' mitigation strategies are triggered: This will work as at the 3 month mark. In the analysis, we will use all recalls submitted during the 20th, 21st and 22nd weeks after randomisation for the 5m mark.*	In the 21st week after randomisation, users will be nudged to wear the device for as long as possible for 7 days and sync the device with the app. *In the analysis, we will use all valid daily PA data provided during the 21st week after randomisation for the 5m mark. If there are missing values in that week, we'll replace them with valid values (if available) on the same day of the week within 2 adjacent weeks (i.e. weeks 19-23).* *If the mitigation strategies are triggered:* *In the analysis, we will use all valid daily PA data provided during the 20th,21st and 22nd weeks after randomisation for the 3m mark. If there are missing values in that week, we'll replace them with valid values (if available) on the same day of the week within 2 adjacent weeks (i.e. weeks 18-24).*

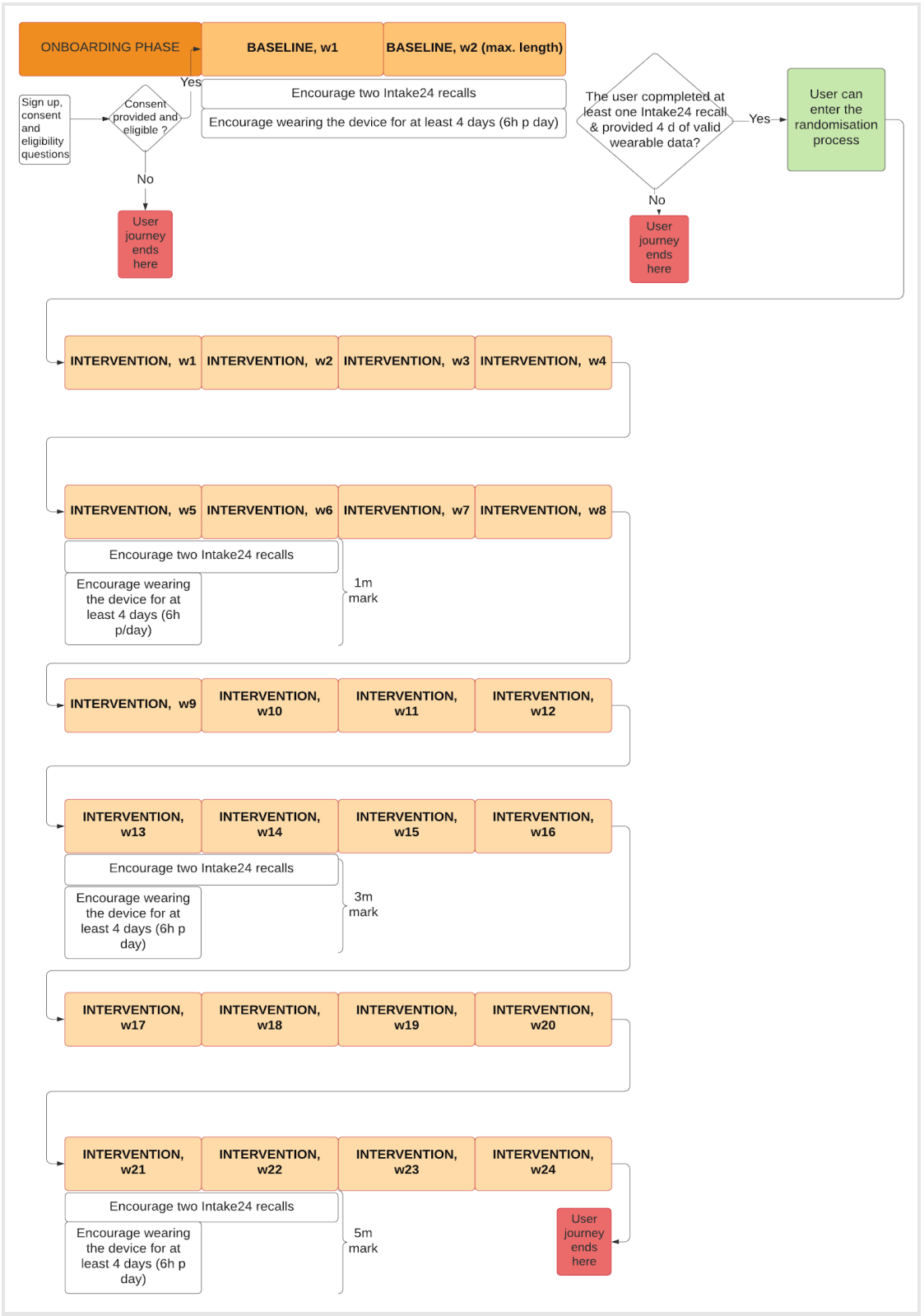
Data fields to be collected. The data collection will be executed by HUL and then transferred securely and anonymously to BIT. Table 10 illustrates the structure of the data fields to be collected by HUL and to be transferred to BIT. See **BIT data list for impact evaluation - V5** for the complete list of data fields.

Table 10. Illustrative structure of data fields to be collected by HUL and to be transferred to BIT

General area	Specific area	Data Field	Mandatory	Usage	Unit and reasonable range
Meta data, including inclusion & exclusion criteria	System	Anonymised GUID	Yes	For joining various datasets	Multi digits
	Household	Household ID	Yes	For analysis	N/A
	Eligibility	Download and install app	No (auto collected)	To select eligible participants	Yes; No
	Eligibility	Consent to participate	Yes	To select eligible participants	Yes; No
	Eligibility	Consent date	Yes	To select eligible participants	Within the recruitment window. If a user does not consent, the consent date will be encoded as 0000-00-00 and the user cannot proceed to health condition and address pages.
	Eligibility	Health condition	Yes	To select eligible participants	Suitable; unsuitable
	Eligibility	Adult status	No (derived from age)	To select eligible participants	Adult; Minor
	Eligibility	Full postcode	Yes	To select eligible participants, and to merge postcode-level deprivation data with GUIDs	N/A
	Eligibility	Resident of Wolverhampton	Yes	To select eligible participants	Yes; No
Uptake and Engagement	Wearable use	Daily wear time	Yes	For imputation of missing values	0~24 hours
	App use	Whether a GUID opens the app by day	Yes	To infer active user status and attrition	Used to identify different types of users: **Very active users**: app installed and opened at least once over the course of 7 days **Active users** (Users who have opened the app at least once in the last 30 days). **Inactive users** (Users who have not opened the app in 30 days)

					Churn outs (Users who have not open the app for more than 30 days)
		App open date	Yes	To examine wear time / app opens for the same GUID over time	N/A

Figure 3. High-level user journey for participants, from an evaluation point of view.



2.6.2 Data quality management

We will seek to obtain high-quality **physical activity data** with two strategies:

- **During data collection.** Participants will be nudged to wear and synchronise the device throughout the trial period to maximise the data fidelity of MVPA mins and steps count and to minimise missingness.
- **Before analysis.** Based on univariate analysis, we will search for and exclude or replace outlier values. We are aware that due to some technical glitches, participants' physical activity data may exceed plausible boundaries (e.g. more than 1440 minutes of MVPA per day, or more than 3 million steps per day). To mitigate the influence of those outliers, we will exclude the implausible records and apply further boundary rules (excluding observations that are below the 2.5th or above the 97.5th percentiles) to ensure data fidelity. BIT will also inspect PA sample data after trial launch and refine the above criteria to handle outliers if necessary (see section 2.6.3.2 for details).

Likewise, the quality of **diet data** will also be managed at two stages: during data collection and before data analysis:

- **During data collection.** Intake24 includes automatic checks and nudges to rule out implausible inputs (e.g. portion size, time gap between meals).
- **Before analysis.**
    - (i) Firstly, we will exclude administrations satisfying any of the following conditions (these conditions were used in the National Diet and Nutrition Survey (NDNS)):
        - Fewer than 10 food items in a recall
        - 3 or fewer eating occasions in a recall (this includes occasions when a participant reports consuming only a drink without food)
        - Completion time of under 3 minutes
        - Total energy intake less than 400kcal or more than 4,000kcal and the individual had not stated that they consumed "less than usual", "more than usual", or that they were on a weight gain or weight loss diet
    - (ii) After doing this, we will inspect the distributions of the dietary outcomes and compare that against the mean level and standard deviation of UK adults' food and macro-nutrients consumption. We will then exclude values below the 1st or above the 99th percentile, by gender.

2.6.3 Statistical methods

2.6.3.1 Attrition management

Based on the previous literature and on HUL's experience, we expect high levels of attrition (See Section 2.4.2.3).

Attrition (users not providing the wearable device data or not completing Intake24) might happen for multiple reasons (e.g. lack of engagement with data collection, lack of engagement with the app, etc.) at any time. For the purpose of the impact evaluation, our core concern is attrition at the specific measurement points:

- 1 month follow-up;
- 3 month follow-up;
- 5 month follow-up.

We anticipate that differential attrition might occur as participants in the control group might be less incentivised to stay enrolled in the study compared to those in the intervention groups. If the ratio of attrition between the control arm and the pooled intervention arm is greater than or equal to 1.1 (the average relative attrition rate according to a recent meta-analysis on differential attrition of behavioural change interventions[53]), we propose to adjust for the differential attrition using inverse probability weighting (IPW)[54], a method proven effective in reducing selection bias for longitudinal studies. The IPW works by modelling the probability of successful retention at 1 month, 3 months and 5 months using baseline observables and then re-weighting those that were retained, so that the reweighted data would be balanced in terms of baseline observables across different arms.

In practice, this means running a logistic regression to calculate the probability of successful retention as a function of baseline variables, getting the fitted probabilities, and then using the inverse of these fitted probabilities as weights in the regression model.

The degree of attrition might vary with outcome measures. As physical activity outcomes are passively captured by the wearable and require no extra burden from the users other than wearing the device and syncing it with the app, we expect the degree of attrition to be lower than for the dietary outcomes measured by Intake24, which require users to spend up to 20 minutes to complete each questionnaire. We will test for differing attrition rates separately for each type of primary outcome (we will use MVPA minutes for physical activity and fruit and vegetables for diet). If we find a ratio of attrition greater than 1.1 for a given primary outcome type, we will apply IPW for all corresponding primary and exploratory outcomes related to physical activity / diet as well.

2.6.3.2 Missing data management

Missing data might occur when a user stops providing data temporarily or permanently. We suggest two different procedures to handle missing data depending on whether participants stop providing data temporarily or permanently:

---

[53] Crutzen, R., Viechtbauer, W., Spigt, M., & Kotz, D. (2015). Differential attrition in health behaviour change trials: A systematic review and meta-analysis. *Psychology & Health*, *30*(1), 122–134. doi: 10.1080/08870446.2014.953526

[54] Schmidt, S. C. E., & Woll, A. (2017). Longitudinal drop-out and weighting against its bias. *BMC Medical Research Methodology 2017 17:1*, *17*(1), 1–11. doi: 10.1186/S12874-017-0446-X

- User stops providing data **temporarily**. For example, PA data may not be available on a given date, or a user provides only one Intake24 recall at a given data collection point. In the sections below, we outline how we would approach such circumstances.
- User stops providing data **permanently** (attrition from measurement). Our general principle is that we will not attempt to replace missing values when missing outcome measures are due to attrition (i.e. a user does not provide any data after a given data collection point in the trial). We define this type of missing data as 'attrition', and we detailed our approach to deal with differential attrition in the previous section.

Physical activity outcomes

For physical activity outcomes (MVPA, steps, energy expenditure), we will get daily reads from the date of onboarding to the end of the trial. It is possible that a user does not provide data on a given day for the following reasons:

- Forgetting to wear the wearable;
- Dead battery.

Individuals may also wear the wearable device on a given day but not long enough for the reading to be sufficiently accurate. We define a 'valid' read as occurring when the individual wears their wearable device for at least 6 hours in the day. Users will not be nudged to meet this criteria, but will be nudged to wear their device as long as possible each day. In the main specification for each physical activity outcome, we use 'valid' reads for all days of the 21st week (i.e. month 5) of the trial as outcomes, and also control for 'valid' reads taken in the last 7 days of the baseline period if they exist.

Additionally, it is possible for reads based on at least 6 hours of wear-time to be erroneous due to issues with the wearable devices (point 1 below). We will replace 'invalid' reads or erroneous daily reads in the last 7 days of the baseline period and 5-month measurement week as follows:

1. We will exclude reads that are below the 2.5th or above the 97.5th percentile[55] within each combination of period (baseline, 1 month, 3 months, 5 months), treatment arm (no / low / medium / high incentives) and day of the week, for reads based on at least 6 hours of wear-time. This is intended to remove erroneous extreme daily reads.
2. We will replace invalid daily reads in the 5-month measurement week with other reads taken by the individual on the same day of the week (e.g. Monday) and within 2 weeks before or after the evaluation week as follows:

[55] We will examine these cut-offs after inspecting the trial data and adjust them if they are inappropriate, e.g. if there is a higher proportion of extremely inactive or active people than expected, we may adjust the cut-off points to make the distribution of included observations less skewed.

- If an individual has a valid read on the same day of the week for the week before or after, we will replace the invalid read with it. If an individual has valid reads for both the week before and after, we will average these two reads.
- Otherwise, if an individual has a valid read on the same day of the week for 2 weeks before or after, we will replace the invalid read with it. If an individual has valid reads for both 2 weeks before and 2 weeks after, we will average these two reads.
- Otherwise, we will not impute.[56]

3. We will impute reads in the last week of the baseline period (i.e. (up to) the last 7 days of the individual's baseline period) in the same way, except that we will only look at earlier reads from the baseline period as potential replacements. We will only do this if a user has at least 4 days of valid baseline data. If they do not, they will be excluded from the analysis.

When analysing the effects of incentives at 1 and 3 months (i.e. the 5th and 13th weeks respectively), we will perform a similar process.

This imputation process will increase the sample size of person-day observations for the 5-month measurement week, which increases the precision of our estimates. It will also remove (i) reads based on low wear-time which may be underestimating physical activity, as well as (ii) reads that are erroneous due to malfunctioning wearable devices. This imputation process may not fully mitigate the issue that users will be more likely to wear the device when doing more PA (likely to be affected by the treatment assignment). However, we do not think this relationship will be too strong near the 6-hour cut-off point, since historical data on HUL Delta[57] users suggests that few users (<3%) wear the wearable for less than 6 hours per day on average. Alternative imputation methods that we have considered such as multiple imputation, which takes advantage of the within-subject auto-correlation over time does not fully mitigate against this issue either.

Dietary outcomes

Dietary recalls are provided by users only at key data collection points (baseline, 1 month, 3 months, 5 months). At these data collection points, users are encouraged to provide two recalls in the same week. We do not impute dietary outcomes. If a user provides one dietary recall at a given data collection point, we will not impute the other. If a user provides zero dietary recalls at a given data collection point, we will not impute either of the recalls since this requires using information that is at least 2 months away.

---

[56] The following example provides further clarification. Say that an individual has an invalid read from Monday 16th May, which is at month 5, i.e. 21st week into the trial after randomisation.
- If they have a valid read from Monday 9th May or Monday 23rd May, we will take this as the value for Monday 16th May. If they have valid reads for both dates, we will average them.
- If they do not have a valid read from Monday 9th May or Monday 23rd May but do have a valid read from Monday 2nd May or Monday 30th May, we will average the available reads from these dates.
- If they do not have a valid read from 2nd, 9th, 23rd or 30th May, we will not impute.

[57] HUL Delta is the name of HUL's own branded wearable device.

For both the primary physical activity and dietary outcomes, we examine the sensitivity of our main findings to different assumptions about missing data (see section 2.6.3.9).

2.6.3.3 Balance checks

We will perform **randomisation balance checks (descriptive statistics)** on the following covariates:

- Baseline MVPA min per day;
- Baseline daily steps;
- Baseline fruit and vegetables (g/day);
- Baseline fibre (g/day);
- Baseline free sugars (% daily food energy);
- Baseline saturated fat (% daily food energy);
- Age;
- Sex;
- Ethnicity;
- Brand of wearable device.

2.6.3.4 Overall analytical strategy

The entire analysis for this trial is based on an **intention to treat (ITT).** This means that we will assess the impact of being offered any financial incentive (compared to being offered no financial incentives).

In this trial, the unit of analysis is the individual's diet and PA outcomes repeatedly collected over multiple windows of evaluation, and the unit of randomisation is the household. To account for the clustering of observations within individuals and individuals within households, we will use linear mixed-effects models to analyse the impact of being offered any incentive (compared to being offered no incentive) for primary outcomes for both physical activity and diet.

The following table summarises the pre-specified analyses detailed in this evaluation protocol.

Table 11. Pre-specified analyses

Analyses	Research question	Outcome	Analysed
Primary	Effect of offering financial incentive on PA	MVPA (min/day) Daily steps	Effect at 5m mark; Report absolute change
	Effect of offering financial incentive on dietary behaviour	Fruit and vegetables (g/day) Fibre (g/day) Free sugars (% daily food energy) Saturated fat (% daily food energy)	
Secondary	Broader effect of offering financial incentive on holistic constructs	Energy expenditure (kcal/day) Energy intake (kcal/day) Weight (kg) A healthy eating score (1-7)	Effect at 5m mark; Report absolute change
	Short-term effects on primary outcomes (PA and diet)	MVPA (min/day) Daily steps Fruit and vegetables (g/day) Fibre (g/day) Free sugars (% daily food energy) Saturated Fat (% daily food energy)	Effect at 1m mark; Report absolute change
	Medium-term effects on primary outcomes (PA and diet)	MVPA (min/day) Daily steps Fruit and vegetables (g/day) Fibre (g/day) Free sugars (% daily food energy) Saturated Fat (% daily food energy)	Effect at 3m mark; Report absolute change
	Impact of different incentive levels	MVPA (min/day) Daily steps Fruit and vegetables (g/day) Fibre (g/day) Free sugars (% daily food energy) Saturated Fat (% daily food energy)	Effect at 5m mark; Report absolute change
Expolorator y	Motivation to change behaviours	Motivation to change PA Motivation to change diet	Effect at 5m mark; Report absolute change
	Subgroup analyses. Analysis by: ● deprivation level ● ethnic group	MVPA (min/day) Daily steps Fruit and vegetables (g/day) Fibre (g/day)	Effect at 5m mark; Report absolute

	• age • sex • baseline diet • baseline PA	Free sugars (% daily food energy) Saturated Fat (% daily food energy)	change
	Unintended consequences	Self-reported mental well-being measured by WHO-5 Index (0~100, 0 = worst, 100 = best)	Effect at 5m mark; Report absolute change
		Sleep quality (hours/day)	
Sensitivity	Robustness to imputation method (subject to contractual agreement)	MVPA (min/day) Daily steps	Effect at 5m mark; Report absolute change

2.6.3.5 Analytical strategy for the primary analysis

The primary outcomes for physical activity are **MVPA minutes per day and daily steps at 5 months into the trial (week 21)**.

We will test for differences between the pooled treatment arms and the control arm five months after randomisation using an intention-to-treat approach and a **linear mixed model (panel regression with random effects)** appropriate for clustered data with repeated measurements, in line with methodology adopted by studies with similar designs[58,59]. As part of the analytical strategy, we will control for the individual characteristics of age, sex, ethnicity, education, BMI at baseline, the brand of wearable.

The general equation for estimating the model is as follows:

$$(1)\quad y_{ijd} = \alpha + \beta\, Treatment_j + \gamma_d + \gamma_W week_{ijd} + \gamma_B baseline_{ijd} + \gamma_M missing.baseline_{ijd} +$$

$$\ldots \gamma_X X_{ij} + \delta_{Cj} + \delta_{Pij} + u_{ijd}$$

$$\delta_{Cj} \sim N(0, \sigma^2_C);\ \delta_{Pij} \sim N(0, \sigma^2_P)\ for\ all\ i, j;\ u_{ijt} \sim N(0, \sigma^2_u)\ for\ all\ i, j, d$$

where:

- y_{ijd} is the daily MVPA minutes for participant i within household j on day d;

[58] Harkins, K. A., Kullgren, J. T., Bellamy, S. L., Karlawish, J., & Glanz, K. (2017). A Trial of Financial and Social Incentives to Increase Older Adults' Walking. *American Journal of Preventive Medicine*, *52*(5), e123–e130. doi: 10.1016/j.amepre.2016.11.011

[59] Finkelstein, E. A., Haaland, B. A., Bilger, M., Sahasranaman, A., Sloan, R. A., Nang, E. E. K., & Evenson, K. R. (2016). Effectiveness of activity trackers with and without incentives to increase physical activity (TRIPPA): a randomised controlled trial. *The Lancet Diabetes & Endocrinology*, *4*(12), 983–995. doi: 10.1016/S2213-8587(16)30284-4

- $Treatment_j$ is a dummy variable taking value 0 for households allocated to the control group, and 1 for households allocated to *any* treatment group;
- γ_d is a fixed effect for the day of the week;
- $week_{ijd}$ is a categorical variable for the (actual) calendar week that the read is recorded;
- $baseline_{ijd}$ is the value of the outcome on the same day of the last week of the baseline period, approximately five months earlier. Missing values for $baseline_{ijd}$ will be coded as -99 to ensure that we include as many participants as possible;
- $missing.baseline_{ijd}$ is a dummy variable that equals 1 if $baseline_{ijd}$ is missing and 0 otherwise and will capture whether MVPA is systematically different for individuals who are missing baseline data for the same day;
- X_{ij} are individual level (including age, sex, ethnicity, education, BMI at baseline, and brand of wearable) and family level covariates - we will take missing values for these covariates as separate categories rather than imputing;
- δ_{Cj} are the cluster (i.e. household) level random effects, which have mean zero and follow a normal distribution;
- For each family *j*, δ_{Pij} are the individual (person) level random effects, which have mean zero within each household and follow a normal distribution;
- u_{ijd} is an idiosyncratic error term. We will use a compound symmetry covariance structure (default for a model with random intercepts).

The primary outcomes for diet are:

- **Fruit and vegetables (g/day);**
- **Fibre (g/day);**
- **Free sugars (% daily food energy);**
- **Saturated fat (% daily food energy);**

as self-reported in Intake24 recalls at the 5-month follow-up point. We will use a similar model to analyse this outcome, as shown below:

$$(2)\quad y_{ij} = \alpha + \beta\, Treatment_j + \gamma_D weekday_{ij} + \gamma_W week_{ij} + \gamma_N num.admin_{ij} +$$

$$\ldots \gamma_B baseline_{ij} + \gamma_A weekday.baseline_{ij} + \gamma_X X_{ij} + \delta_{Cj} + u_{ij}$$

$$\delta_{Cj} \sim N(0, \sigma^2_C) \; for \; all \; j; \; u_{ij} \sim N(0, \sigma^2_u) \; for \; all \; i, j$$

where:

- y_{ij} is person i's energy intake according to Intake24 surveys at 5 months. If the person took two surveys, we average their reports;
- $weekday_{ij}$ is the proportion of administrations at 5 months that occurred on a weekday (rather than on the weekend). For example, if an individual had one administration on Monday and one on Saturday, $weekday.ind_{ij}$ equals 0.5. This is intended to account for the difference in dietary patterns between weekdays and the weekend;
- $num.admin_{ij}$ is the number of administrations at 5 months (1 or 2);
- $baseline_{ij}$ is calculated in an analogous way to y_{ij}, but for the administration(s) at baseline;
- $weekday.baseline_{ij}$ is calculated in an analogous way to $weekday_{ij}$, but for the administration(s) at the baseline;
- All other variables are defined in an analogous way to equation (1). Note that we do not use person-level random effects because there is only one observation per person in the sample.

We are using two primary outcomes, which raises the probability of a false discovery being made on at least one of them. To control for the false discovery rate over primary outcomes, we perform the Benjamini-Hochberg step-up procedure.

2.6.3.6 Analytical strategy for the secondary analysis: Broader effect of offering financial incentive on holistic constructs

The impact of offering incentives (compared to not offering any incentives) on secondary outcomes at the 5 months mark will be assessed using similar methods as the ones for the primary analysis.

2.6.3.7 Analytical strategy for the secondary analysis: Short- and medium-term impacts on primary outcomes

We will estimate the same models as at 5 months (i.e. as in section 2.6.3.5) for the primary outcomes using outcome data from 1 month and 3 months to analyse the average impact of incentives over those shorter time periods. As described above, the three-month outcomes will be the focus for the study should attrition at five months prohibit meaningful analysis at five months.

2.6.3.8 Analytical strategy for the secondary analysis: Impact of different incentive levels

The impact of each incentive level on the primary outcomes will be assessed using the same approach as the one for the primary analysis but we will categorise the treatment indicator differently. In these regressions, $Treatment_j$ will be a series of three dummy variables, each of them taking value 0 for households (or users for the individual level randomisation) allocated to the control group, and 1 for households (users) allocated to the specific treatment group for low, medium or high levels of incentives.

This approach will allow us to statistically compare the effect of each incentive with the control group.

2.6.3.9 Analytical strategy for the exploratory analysis: subgroup analysis

To help us understand how the intervention might work differently for different subgroups, we will repeat the same analysis as specified in the primary analysis above for primary outcomes, for each of the following subgroups separately:

Table 12. Subgroups

Subgroups of interest	Cut-off criteria	Expected proportion of the sample
People who live in the most deprived areas	Participants will be categorised as living in the most deprived areas if their post code corresponds to Index of Multiple Deprivation (IMD) ≤ 2 (top 20% most **deprived**)	51.3% (source: Wolverhampton deprivation status)
Ethnicity	5 GSS macro groups [White / Mixed or multiple ethnic groups / Asian or Asian British / Black or African or Caribbean or Black British / Other ethnic group]	White: 68%; Asian: 18%; Black: 7%; Mixed: 5% Others: 2% (source:Wolverhampton's census statistics)
Sex	Male; female; prefer not to say	~ 50% males (source: Wolverhampton's census statistics)
Age	Age ≤ median age of the sample; Age > median age of the sample	~ 50% vs. 50%
Baseline diet	Self-reported daily consumption of fruit and veggie ≥ 3 portions self-reported daily consumption of fruit and veggie < 3 portions	~ 62% (≥ 3 portions) vs. 38% (< 3 portions)[60]

[60] According to the Health Survey for England, on average, adults consume 3.7 portions per day, and given that the standard deviation is 2.4 portions per day, 3 portions/day is about 0.3 SD below the average, corresponding to 38% of the population assuming the distribution of fruit and vegetable consumption is normal.

Baseline MVPA	MVPA ≥ 30 mins / week (fair active or active) MVPA < 30 mins / week (inactive)	~ 73% (active) vs. 27% (inactive), according to the Sport England Active Lives Survey

2.6.3.10 Analytical strategy for the exploratory analysis: Motivation to change and unintended consequences

The impact of offering incentives (compared to not offering any incentives) on exploratory outcomes at the 5 months mark will be assessed using similar methods as the ones for the primary analysis. We will not impute values at baseline or at 5 months. If there are missing values for an exploratory outcome at baseline, we will add an indicator variable for missingness to the set of covariates so that all individuals with non-missing values for the outcome at 5 months can be included.

2.6.3.11 Sensitivity analysis

We will conduct the following sensitivity analysis subject to contractual agreement. We will test the sensitivity of the results of the physical activity and diet primary outcomes to alternative methods of missing data management (multiple imputation and delta adjustment).

Simple imputation:

- **Physical activity**: Within each combination of data collection point (baseline, 1 month, 3 months, 5 months), treatment arm (no / low / medium / high incentives) and day of the week, for reads based on at least 6 hours of wear-time, we will identify the 2.5th and 97.5th percentiles for each physical activity outcome. We will replace reads below this 2.5th percentile (no matter what their wear-time is) with the 2.5th percentile. Similarly, we will replace reads above the 97.5th percentile with the 97.5th percentile. This means we can use a full dataset among non-attrited participants with minimal assumptions around the structure of missing data.
- **Diet**: Within each combination of data collection point, treatment arm and gender, we will replace values of each primary diet outcome below the 1st percentile with the 1st percentile, and values above the 99th percentile with the 99th percentile. We will still exclude administrations that fail any of our attention checks (e.g. completion time of under 3 minutes).

Multiple imputation: If the results differ between the main specification and the simple imputation method above, we will perform multiple imputation with delta adjustment sensitivity analysis.

- **Physical activity**: Having replaced invalid daily reads within the measurement week using valid reads from up to two weeks before/after, we will generate *n* imputed datasets, where *n* is the percentage of incomplete cases rounded up

to the nearest integer (following the rule of thumb suggested by Bodner (2008)[61] and White et al. (2011)[62]). We will use sequential predictive mean matching to impute missing values of physical activity outcomes within each dataset (imputing the outcomes together). As predictors, we will use other reads in the measurement week, all covariates in the main specification, and household- and person-level fixed effects.

We will perform multiple imputation only for the sample of individuals who have at least one valid read in the measurement week. We will then estimate the following equation on each imputed dataset for a range of (fixed) Δ:

$$(3)\ y_{ijd} = \alpha + \beta\, Treatment_j + \gamma_d + \gamma_W week_{ijd} + \gamma_B baseline_{ijd} + \gamma_M missing.baseline_{ijd} +$$

$$\dots \Delta missing.outcome_{ijd} + \gamma_X X_{ij} + \delta_{Cj} + \delta_{Pij} + u_{ijd}$$

$missing.outcome_{ijd}$ is a dummy variable which equals 1 if the outcome is missing and 0 otherwise. We will pool the estimated coefficients and standard errors across the 25 imputed datasets using Rubin's rules[63].

Under $\Delta = 0$, we are assuming that data are missing at random (MAR). This delta adjustment sensitivity analysis informs us to what degree the imputed data could be underestimating the outcome while the findings still hold, for the sample of individuals who had at least one valid read in the measurement week. In other words, it does not capture bias from (differential) attrition.

- **Diet**: We will impute primary diet outcomes below the 1st percentile or above the 99th percentile. As predictors, we will use values of the outcomes in other measurement weeks, all covariates in the main specification, and household- and person-level fixed effects. Again, we will exclude administrations that fail any of our attention checks (e.g. completion time of under 3 minutes).

2.7 Monitoring

2.7.1 Data monitoring

Throughout the trial period, HUL will monitor data on 'consented users' daily and share these data with BIT and DHSC via a live dashboard. As noted above, information about recruitment rate and attrition might be used to inform decisions by HUL and DHSC to potentially increase the resources and efforts allocated to the marketing campaign or to retain users to achieve the target sample size at various trial stages.

[61] Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal* (15) 651-675.
[62] White, I. R., Royston P. and Wood A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* (30) 377-399.
[63] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc.

We do not plan to conduct any interim quantitative analysis for outcomes prior to the completion of the data collection period for the longest follow up. We do not plan to discontinue the study depending on numbers of active users, as many questions in this study pertain to retention and engagement levels over the course of the intervention. While this is a short-term trial with minimal risks of harm we will nonetheless set up an independent data monitoring committee (DMC)[64] to be convened on an ad hoc basis to investigate un-blinded safety and data issues, or to investigate issues raised by the Trial Steering Committee.

Table 13 provides an overview of the different kinds of reporting that DHSC will receive throughout the duration of the trial and after the trial is concluded.

Table 13. Monitoring report by trial phase (TBC)

Trial phase	Monitored data (n)	Reporting frequency	Provided by whom?	In which form
MARKETING: ACQUISITION & ENGAGEMENT (Intervention period)	• Funnel report (vs. target) ○ Registered ○ Consented ○ Randomised	Daily (during launch) Weekly thereafter	HUL	Live dashboard
	• Participant recruitment profile report (aggregate level only) – profile of registered users by: ○ Trial arm ○ Age ○ Gender ○ Suburb / top-level postal code ○ Ethnicity ○ Sub-group segmenters where populated – refer to DPIA for full list of attributes for segmentation ○ Week, month, cumulative	Daily (during launch) Weekly thereafter	HUL	
	• Device connect type	Daily	HUL	
PERFORMANCE REPORTING (Intervention period)	• DAU / WAU / MAU by: ○ Trial arm ○ Age ○ Gender ○ Suburb / top-level postal code ○ Sub-group segmenters (disability status, ethnicity, self-motivation, etc.) – refer to DPIA for	Weekly	HUL	Live dashboard

[64] Chan AW, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, Dickersin K, Hróbjartsson A, Schulz KF, Parulekar WR, Krleža-Jerić K. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. Bmj. 2013 Jan 9;346.

	full list of attributes for segmentation ○ Week, month, cumulative ● Surveys completed + results ○ In-app surveys ○ Intake24	Weekly / monthly / YTD Weekly Fortnightly		
	● CS report: ○ # tickets received ○ # tickets resolved ○ # pending ○ SLA performance vs. target ○ CSAT ● Uptime SLA reporting (monthly %)	Monthly	HUL	Live dashboard
INCENTIVES REPORTING (Intervention period)	● Incentives report, by: ○ Trial arm ○ Age ○ Gender ○ Suburb / top-level postal code ○ Sub-group segmenters (disability status, ethnicity, self-motivation, etc.) – refer to DPIA for full list of attributes for segmentation ○ Week, month, cumulative ● Points earned ● Points redeemed ● Points unredeemed (%) ● Quantity, £ value, and number of incentives redeemed, by: ○ Product category ○ Retailer / manufacturer ○ Physical / virtual ○ SKU ● Top 20 most popular redemption items by volume ● Value of incentives issued ● Mean time between incentive issue to incentive redemption	Monthly (refer to HUL's design plan)	HUL	Live dashboard
HEALTH & BEHAVIOUR CHANGE REPORTING (Intervention period)	● Health dashboards, by: ○ Trial arm ○ Age ○ Gender ○ Deprivation index ○ Sub-group segmenters – refer to DPIA for full list of attributes for segmentation	Daily	HUL	Live dashboard

	● Baseline report ● Weight report: ○ BMI ● Sleep report: ○ Median daily sleep duration ○ Median daily sleep efficiency ● Monthly health changes report: ○ Starting sleep vs. Tn sleep ○ Starting steps vs. Tn steps ○ Starting moderate / intense physical activity vs. Tn moderate / intense physical activity ● Challenges report: ○ Challenges started ○ Challenges status: overachievement success, almost success, failure ○ Challenge validation status (% of rejection?) ○ Recommended and chosen challenge difficulty: maintain or harder ○ Type of challenges (e.g. % of users selected “Step Up”, “Boost veggie and fruits”)	Monthly	HUL	Live dashboard
Post trial - 12 weeks after the end of the trial (incl. end of qualitative research components)	All primary, exploratory, quantitative IPE, and qualitative IPE outcomes described in this protocol	Once (final report)	BIT	Report

2.7.2 Harms

The intervention tested in this study is considered to be a low-risk health promotion intervention. However, unintended consequences (both positive and negative) will be explored as part of the qualitative interviews and focus groups conducted as part of the Implementation and Process Evaluation. As specified in section 4.3.6, participants reporting negative unintended consequences during the interviews will be signposted to relevant resources. We will have a clinical safety group in place throughout the pilot, as well as a Clinical Safety Officer (CSO) to monitor (potential) harms during the study. Participants will be informed at the consent stage that they are free to stop participating in the study at any point, without having to provide any reason.

2.7.3 Auditing

Auditing involves periodic independent review of core trial processes and documents.[65] For this trial, we do not plan any auditing. The trial will be conducted according to the processes described in this protocol. The study protocol has been designed with input and guidance from the DHSC's policy and delivery team and the Design and Evaluation Advisory Group (DEAG – the group of experts from across DHSC diet and physical activity policy and National Institute for Health Research Policy Research Units, established to provide support, advice and steers on the development of the design plan and evaluation protocol). The protocol will subsequently be going through external peer review before being submitted to UKHSA for ethical approval.

The trial will only be launched following ethical approval from UKHSA. The UKHSA ethics committee and the publicly available protocol will be updated should there be any changes to the intervention or study processes taking place after the launch of the study.

3. Ethics and dissemination

The study has obtained ethical approval from UKHSA[66].

Before the trial launch, BIT will register the evaluation protocol on an appropriate registration platform, to be agreed with the funder, DHSC.[67]

The UKHSA ethics committee, the funder (DHSC), and the publicly available protocol will be updated should there be any changes to the intervention or study processes taking place after the launch of the study.

---

[65] Chan AW, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, Dickersin K, Hróbjartsson A, Schulz KF, Parulekar WR, Krleža-Jerić K. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. Bmj. 2013 Jan 9;346.

[66] An update will be provided on this approval in December 2022.

[67] The trial was registered with ISRCTN with Submission number 43198 ahead of the trial launch.

Subject to DHSC’s approval, it is BIT's intention to publish the results from this study in an academic journal. Subject to DHSC’s approval, results of the trial might also be presented at conferences, public meetings, or through other platforms.

4. Implementation and Process Evaluation

4.1 IPE Design

Whilst the impact evaluation will test the effectiveness of the financial incentive scheme, the implementation and process evaluation (IPE) will identify why and how the intervention achieves - or fails to achieve - the expected outcomes in relation to the Theory of Change (ToC).[68] It will also explore potential desirable and undesirable unintended consequences.[69]

Following best-practice guidance from the Medical Research Council (MRC),[70] BIT will conduct a mixed-methods IPE to understand issues relating to the (i) reach of the intervention, (ii) engagement with the intervention, (iii) mechanisms of impact, (iv) and implementation and feasibility.

This mixed-methods approach will incorporate qualitative data from interviews and focus groups and quantitative data from routinely collected in-app user metrics. Using multiple data sources, BIT will enable DHSC to:

- **Gain broad insights across a large number of individuals**: the quantitative IPE will assess key process variables (e.g., engagement with the app) and generate evidence on the mechanisms of impact hypothesised in the ToC.
- **Develop an in-depth understanding of individual experiences:** qualitative methods will further enrich the IPE by generating in-depth insights into the range and diversity of the experiences of different stakeholders.

The methods used for the IPE will be (i) rooted in the details of the Theory of Change and the user journey (Figure 2), (ii) mindful of the needs of the research participants, especially those from underserved communities, and (iii) form part of an integrated plan with the impact evaluation, so that the analysis from the IPE can help to explain any significant or null effects observed.

In the sections that follow, we outline the core research questions for both the qualitative and quantitative IPE (4.1.1), and then describe the methodology for the quantitative (Section 4.2) and the qualitative IPE (Section 4.3).

4.1.1 Research questions

Core research questions

In line with the MRC's recommendations, BIT used the causal hypotheses outlined in the Theory of Change (see accompanying Intervention design & ToC document) and

[68] For the ToC, please see the complementary document.
[69] Public Health England. (2018). *Guidance: Process Evaluation.* Retrieved from www.gov.uk/government/publications/evaluation-in-health-and-well-being-overview/process-evaluation
[70] Moore G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W. et al. Process evaluation of complex interventions: Medical Research Council guidance BMJ 2015; 350 :h1258 doi:10.1136/bmj.h1258

user journey (Figure 2 in the Intervention design & ToC document) to prioritise research topics for investigation.[71] These are:

- **Reach**: what factors affected the intervention's reach?
- **Engagement**: what factors affected the engagement of users with the intervention?
- **Mechanisms of impact**: through what mechanisms does the intervention affect behaviour change?
- **Implementation and feasibility**: how was the intervention implemented and is it scalable?

Table 14 below provides a summary of the high-level research questions for each topic along with the IPE methodology to be employed. As noted, a more detailed breakdown of the quantitative and qualitative methodology are outlined in Sections 4.2 and 4.3 respectively.

Table 14. Summary of research topics, research questions and methodologies.

Research Topic	Research question		IPE	Methodology
Reach What factors affected the intervention's reach?	1.1	To what extent does the intervention reach participants?	Quant	Descriptive analysis of metrics from marketing campaign and in-app engagement (summary statistics)
	1.2	What are the characteristics of intervention recipients?		
	1.3	What was the role of marketing and communications in motivating participants to join the intervention?	Qual	Thematic analysis of interviews and focus groups with recipients and non-recipients.
	1.4	What were the channels of reach and how did they affect sign-up?		
	1.5	How is the intervention perceived by recipients, and non-recipients of the financial incentives?		
	1.6	What are the barriers and facilitators to the intervention's reach?		
Engagement What factors affected the engagement of users with the intervention?	2.1	To what extent does the intervention engage participants?	Quant	i. Descriptive analysis of in-app engagement metrics and participant demographics (summary statistics) ii. Regression analysis of in-app engagement metrics and participant demographics
	2.2	What are the characteristics of those who engage with the intervention for the duration of the trial, those who partly engage, and those who disengage?		

[71] Moore G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W. et al. Process evaluation of complex interventions: Medical Research Council guidance BMJ 2015; 350 :h1258 doi:10.1136/bmj.h1258

	2.3	In what ways did users engage with the app after signing up?	Quant	Descriptive analysis of in-app engagement metrics (summary statistics)
			Qual	Thematic analysis of interviews and focus groups with recipients and non-recipients.
	2.4	What were participants' experiences and perspectives of intervention?	Qual	
	2.5	What are the barriers and facilitators to engagement with the intervention?		
Mechanisms of impact Through what mechanisms does the intervention affect behaviour change?	3.1	To what extent do incentives affect in-app engagement?	Quant	Regression analysis of treatment impact on in-app engagement metrics
	3.2	What barriers and facilitators - both contextual and individual - affect the extent to which the intervention changes behaviours for recipients and non-recipients of the financial incentives?	Qual	i. Thematic analysis of interviews and focus groups with recipients and non-recipients ii. Thematic analysis of focus groups with delivery and reward partners
	3.3	How do features of the incentive scheme affect (or not) the extent to which the scheme changes behaviours?	Qual	
Implementation and feasibility How was the programme implemented and is it scalable?	4.1	What was the process for developing and implementing the intervention among delivery and reward partners?	Qual	Thematic analysis of focus groups with delivery and reward partners.
	4.2	Are the design and delivery processes fit for scaling and sustaining the intervention?	Qual	
	4.3	How was the intervention implemented?	Qual	
	4.4	What are the facilitators and barriers to scaling and sustaining the intervention (including financial incentives) beyond the pilot?	Qual	
	4.5	What was the extent of 'gaming' and data errors in the intervention?	Quant	Descriptive analysis of in-app data on inconsistencies and outlying behaviour (summary statistics)

Using qualitative methods (outlined in section 4.3), BIT will also explore how acceptability, contextual factors and barriers/facilitators to the implementation and outcomes affect each research question to ensure that the IPE addresses all core components of the MRC's framework for process evaluations of complex interventions.[72]

[72] Moore G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W. et al. Process evaluation of complex interventions: Medical Research Council guidance BMJ 2015; 350 :h1258 doi:10.1136/bmj.h1258

4.1.2 Unintended consequences

The IPE will also explore potential unintended consequences of the intervention using qualitative methods. Unintended consequences may include *negative* or *harmful* outcomes (e.g. increased purchase of unhealthy foods given the increase in financial means) or *positive* or *helpful* "spillover" effects (e.g., increased self-reported wellbeing for those participating).

To identify unintended consequences, BIT:

1. **Interrogated the Theory of Change:** in partnership with HUL, BIT has already reviewed the ToC to identify potential unintended consequences - both helpful and harmful. This involved testing the theoretical assumptions of the ToC and a systematic evaluation of the inputs, activities, outputs, outcomes and impact to generate a list of hypothesised unintended outcomes at each stage of the intervention. These were used to inform questions in the topic guide to surface unintended consequences.

2. **Identified any "dark logic" underlying the Theory of Change**: whilst the Theory of Change outlines a path to *positive* or helpful outcomes and unintended consequences, a "dark logic model" aims to anticipate plausible *negative* or harmful unintended consequences. BIT examined the Theory of Change and its assumptions to identify any "dark logic" that might result from the intervention, and used this to inform the topic guides.[73]

3. **Developed a sampling strategy and qualitative methodology that enables identification of unintended consequences:** definitions of unintended consequences may differ between and within populations. BIT has developed a sampling strategy for the qualitative interviews and focus groups to capture the perspectives about negative unintended consequences from a wide range of population subgroups, not just those most likely to have benefitted.

[73] Bonell C., Jamal, F., Melendez-Torres, G.J., et al. 'Dark logic': theorising the harmful consequences of public health interventionsJ Epidemiol Community Health 2015;69:95-98.

4.2 Quantitative methods

Note: As specified in the previous versions of this protocol, we have updated the IPE in line with evolving information about the intervention and marketing design plan as well as information about data availability from HUL. This represents the final version of the quantitative IPE.

4.2.1 Methods overview

The quantitative IPE aims to:

1. Describe the intervention's reach and users' engagement
2. Test key mechanisms of action
3. Explore potential 'gaming' and data errors

Table 15 below summarises how the user journey stages map onto the planned quantitative IPE analyses.

Table 15. Stages of user journey and quantitative IPE analyses

User journey stage	Intervention step	Quant IPE analysis	Topic covered
Marketing campaign and acquisition	1. Downloading app and in-app consent 2. App usage	Marketing reach, acquisition, and usage analysis	• Reach • Engagement • Mechanisms of action • Implementation and feasibility
Onboarding	3. Mandatory details completion 4. Wearable device connection/use	Engagement with the intervention and mechanisms of action	• Engagement • Mechanisms of action • Implementation and feasibility
	5. Intake24 completion	N/A[74]	N/A
	6. Health surveys	Engagement with the intervention and mechanisms of action	• Engagement • Mechanisms of action • Implementation and feasibility
	7. Randomisation	N/A[75]	N/A
Interaction with the intervention (continued app usage)	8. 'New user journey' of app learning/engagement - first 30 days post-randomisation[76] 9. Continued app usage across trial period (completing challenges,	Engagement with the intervention and mechanisms of action	• Engagement • Mechanisms of action
		Analysis of 'gaming' and data errors	• Implementation and feasibility

[74] This data concerns the impact evaluation only and will be analysed as part of that.

[75] This data concerns the impact evaluation only and will be analysed as part of that.

[76] According to the HUL design plan, this includes: (i) Introduction to app features, (daily) mood check-ins and orientation, through email and push notifications; (ii) Prompts to participate in weekly health challenges focused on PA and nutrition habits; (iii) Survey invites, through push notifications and email – which act as part of the user experience as we are able to better tailor health insights and platform experience with this data; (iv) Milestone celebrations, points and rewards (for incentive groups) and NPS / feedback at 30-day mark. *This is specified here for additional context around the user journey; engagement in these activities will be measured by a set of metrics capturing in-app engagement across the intervention period.*

	redeeming incentives etc.)		

In line with the impact evaluation, one cross-cutting focus of the quantitative IPE will be understanding the experience and behaviours of the entire participant sample and of key participant subgroups. For subgroup analyses we will use the same population characteristics and cut offs as those specified for the impact evaluation.

The quantitative IPE makes an important distinction between **engagement with *intervention components*** (e.g. selecting and completing challenges, redeeming rewards), and **engagement with *evaluation components*** (e.g. completing Intake24 surveys). The quantitative IPE is focused on the former; the impact evaluation includes a focus on the latter, as part of attrition management.

The quantitative IPE will use a combination of descriptive analyses (summary statistics and data visualisation) and regression analyses.

4.2.2. Reach, acquisition, and usage analysis

The reach, acquisition, and usage analysis will focus on (i) quantifying the number of consenting users and breaking down this sample by relevant subgroups, (ii) quantifying the amount of engagement with the app during the duration of the pilot and exploring how engagement levels vary by study arm, (iii) identifying study-level and personal-level predictors of engagement with the app.

Table 16 describes the metrics that will be analysed. The feasibility of the planned analyses rely on HUL's ability to share the in-app metrics described in Table 16 with BIT, which has been confirmed by HUL.

Table 16. Metrics underlying the reach, acquisition, and usage analysis

Metric	Data source	Number of timepoints	Number of groups	Which subgroups
Number of consenting users recruited from launch until recruitment finishes (in N)	In-app data	1 (total timespan)	5	Total and by study arm
% of randomised adults that are very active users (defined as app installed and opened at least once over the course of 7 days) per week from randomisation to week 24 (in %)	In-app data	22 (weekly)	5	Total and by study arm
% of randomised adults that are churn-out users (defined as app not opened since more than 30 days from randomisation to week 24) (in %)	In-app data	22 (weekly)	5	Total and by study arm

Number of consenting users from launch until recruitment finishes (in N), by subgroup	In-app data	1 (total timespan)	80	Total and by study arm. Each respectively by deprivation status (2 groups), ethnicity (5 groups), sex (3 groups), age (2 groups), baseline F&V intake (2 groups), baseline MVPA (2 groups).

The above metrics will be **analysed descriptively using summary statistics.**

Additionally, we will conduct a regression analysis to identify the factors predicting the proportion of intervention period during which consenting participants are categorised as being very active (defined as app installed and opened at least once over the course of the last 7 days). The factors explored in the regression will include:

- age,
- sex,
- baseline overweight/obesity status (BMI ≥ 25),
- ethnicity,
- deprivation status,
- disability status (self-reported presence or absence of physical or mental health conditions or illnesses lasting or expected to last 12 months or more - categorised as yes, not, prefer not to say),
- education (degree or above; professional, vocational or other work-related qualifications; no qualification; prefer not to say)
- brand of wearable (Apple watch, Fitbit, Better Health fitness tracker), Garmin, others),
- study arm (control, low incentives, medium incentives, high incentives),
- baseline MVPA, and
- baseline fruit and vegetable intake.

The cut offs for each predictor will be defined following the same criteria used in the impact evaluation (unless specified above).

Estimated coefficients and statistical testing for the predictors will provide robust insights into whether overall in-app engagement differs by any demographics and baseline behaviours. The general equation for estimating the model is as below:

$$(4)\ y_{ij} = \alpha + \beta\, Treatment_j + \gamma_X X_{ij} + u_{ij}$$

$$u_{ij} \sim N(0, \sigma^2_u)\ for\ all\ i, j$$

where

- y_{ij} is the proportion of weeks in the intervention period that participant i within household j is categorised as being very active;
- $Treatment$ is a categorical variable for treatment group (the categories are control group, low incentives, medium incentives and high incentives, with the control group being the reference category);
- X_{ij} are individual level (including age, sex, ethnicity, education, disability status, baseline obesity status, baseline fruit and vegetable intake, baseline MVPA, and brand of wearable) and family level covariates (e.g. deprivation status) - we will take missing values for these covariates as separate categories rather than imputing;
- u_{ij} is an idiosyncratic error term, which is assumed to follow a Normal distribution (denoted as N) of zero mean and a variance of σ_u^2.

Taken together this analysis section will help to answer following questions:

- To what extent does the app reach participants?
- What are the characteristics of people taking up the programme?
- To what extent does the app engage participants?
- What characteristics predict participants' in-app engagement for the duration of the trial?
- To what extent do incentives affect in-app engagement?

We originally also planned to explore the marketing channels through which consenting participants were reached, but we will explore this through the qualitative IPE, as HUL informed us that this data will not be available.

4.2.3. Engagement with the intervention and mechanisms of action

As indicated in the ToC, beyond the initial engagement with the intervention (e.g. providing in-app consent), continued engagement with the app is critical to driving behavioural change in PA and diet activity.

This includes successful onboarding during the baseline stage, selecting and completing challenges, logging behaviours, verifying behaviours, and redeeming incentive rewards.

To explore these themes, the quantitative IPE will analyse the metrics summarised in Table 17. All data will be collected via the app or by reward providers and will be obtained from HUL. The analyses outlined in this section will also rely on HUL sharing the metrics described in Table 17 with BIT.

Table 17. Metrics underlying the 'engagement with the intervention and mechanisms of action' analysis

Metric	Data source	Number of timepoints	Number of groups	Which groups
% of consenting participants completing their profile before randomisation (in %)	In-app data	1 (total timespan)	1	Total sample
% of consenting participants connecting the wearable device before randomisation (in %)	In-app data	1 (total timespan)	1	Total sample
% of consenting participants that are active during the “new user journey” (defined as app opened at least once every 7 days over the 30 days after randomisation) (in %)	In-app data	1 (total timespan)	5	Total and by study arm
Average daily number of hours the wearable device is worn from consent to week 24 (in N)	In-app data	1 (total timespan)	5	Total and by study arm
Average daily number of hours the wearable device is worn per week from consent to week 24, by week (in N)	In-app data	22 (weekly)	5	Total and by study arm
Average number of PA challenges selected per week from randomisation to week 24 (in N)	In-app data	1 (total timespan)	5	Total and by study arm
Average number of PA challenges successfully completed per week from randomisation to week 24 (in N)	In-app data	1 (total timespan)	5	Total and by study arm
Average number of PA challenges successfully completed per week by difficulty from randomisation to week 24 (in N)	In-app data	1 (total timespan)	5	Total and by study arm for three difficulty levels
Weekly number of PA challenges successfully completed from randomisation to week 24, by week (in N)	In-app data	22 (weekly)	5	Total and by study arm for three metrics
Average number of diet challenges selected per week from randomisation to week 24 (in N)	In-app data	1 (total timespan)	5	Total and by study arm
Average number of diet challenges successfully completed per week from randomisation to week 24 (in N)	In-app data	1 (total timespan)	5	Total and by study arm
Average number of diet challenges completed per week by difficulty from randomisation to week 24 (in N)	In-app data	1 (total timespan)	15	Total and by study arm for three difficulty levels
Weekly number of diet challenges successfully completed from randomisation to week 24, by week (in N)	In-app data	22 (weekly)	5	Total and by study arm
Total number of times each challenge has been selected	In-app data	1 (total timespan)	1	Total sample
Total number of times each reward has been selected	In-app data	1 (total timespan)	1	Total sample
Cumulative number of points per week earned (verified) from randomisation to week 24 (in N)	In-app data	1 (total timespan)	5	Total and by study arm

Cumulative number of redeemed rewards from randomisation to week 24 (in N)	Reward partners	1 (total timespan)	5	Total and by study arm
Cumulative £ value of redeemed rewards from randomisation to week 24 (in £)	Reward partners	1 (total timespan)	5	Total and by study arm
Average number of PA challenges completed successfully per week from randomisation to week 24, by subgroup (in N)	In-app data	1 (total timespan)	80	Total and by study arm. Each respectively by deprivation status (2 groups), ethnicity (5 groups), sex (3 groups), age (2 groups), baseline F&V intake (2 groups), baseline MVPA (2 groups).
Average number of diet challenges completed per week from randomisation to week 24 (in N) by subgroup	In-app data	1 (total timespan)	80	Total and by study arm. Each respectively by deprivation status (2 groups), ethnicity (5 groups), sex (3 groups), age (2 groups), baseline F&V intake (2 groups), baseline MVPA (2 groups).
Average cumulative number of points earned from randomisation to week 24 (in N) by subgroup	In-app data	1 (total timespan)	80	Total and by study arm. Each respectively by deprivation status (2 groups), ethnicity (5 groups), sex (3 groups), age (2 groups), baseline F&V intake (2 groups), baseline MVPA (2 groups).
Average cumulative value of redeemed rewards from randomisation to week 24 (in £) by subgroup	Reward partners	1 (total timespan)	80	Total and by study arm. Each respectively by deprivation status (2 groups), ethnicity (5 groups), sex (3 groups), age (2 groups), baseline F&V intake (2 groups), baseline MVPA (2 groups).

Summary descriptive statistics (e.g., averages, proportions) will be calculated using data on all metrics in Table 17 to provide an overview of in-app engagement for the entire participant sample (i.e. all participants who sign up to the app and provide in-app consent, regardless of subsequent data completeness), and split by trial arm.

Additionally, for key metrics related to continued app usage, summary statistics will also be visualised on a weekly basis across the intervention period to provide further insights about how users' behaviour changes over time.

As specified in the table above, to uncover potential differences in engagement between different population segments, **subgroup analyses will be conducted for key metrics, namely:**

- Average number of PA challenges completed from randomisation to week 24
- Average number of diet challenges completed from randomisation to week 24
- Average cumulative number of points earned (**both verified and unverified**) from randomisation to week 24
- Average cumulative value of redeemed rewards from randomisation to week 24

The characteristics and cut offs used for these subgroup analyses will be the same as those employed as part of the impact evaluation.

We will also conduct a regression analysis to identify the factors predicting the cumulative number of points earned from randomisation to week 24, which is used to estimate users' success rate at selecting and completing diet and physical activity challenges. The factors explored in the regression will include age, sex, baseline obesity status (BMI ≥ 25), ethnicity, deprivation status, disability status, education, brand of wearable, treatment arm (control, low incentives, medium incentives, high incentives), baseline MVPA, and baseline fruit and vegetable intake. The cut offs for each predictor will be defined following the same criteria used in the impact evaluation (unless specified above).

Estimated coefficients and statistical testing for the above predictors will provide robust insights into whether overall engagement with incentives differs by any demographics and baseline PA and diet behaviours. The general equation for estimating the model is similar to that of the impact evaluation:

$$(5)\ y_{ij} = \alpha + \beta\, Treatment_j + \gamma_X X_{ij} + u_{ij}$$

$$u_{ij} \sim N(0, \sigma^2_u)\ for\ all\ i, j$$

where

- y_{ij} is the cumulative number of points earned (**both verified and unverified**) *from randomisation to week 24* for participant i within household j;
- $Treatment$ is a categorical variable for treatment group (the categories are control group, low incentives, medium incentives and high incentives, with the control group being the reference category);
- X_{ij} are individual level (including age, sex, ethnicity, education, disability status, baseline obesity status, baseline fruit and vegetable intake, baseline

MVPA, and brand of wearable) and family level covariates (e.g. deprivation status) - we will take missing values for these covariates as separate categories rather than imputing;
- u_{ij} is an idiosyncratic error term, which is assumed to follow a Normal distribution (denoted as N) of zero mean and a variance of σ_u^2.

We will use a compound symmetry covariance structure (as is implied by the random intercepts). Linear regression will be used for continuous outcomes, and logistic regression will be used for binary outcomes.

Taken together this analysis section will help to answer following questions:

- Overall, to what extent does the intervention engage participants?
- What characteristics predict participants' engagement with the interventions?
- In what ways did users engage with specific intervention components after signing up?
- To what extent do incentives affect engagement with specific intervention components?

4.2.4. Analysis of 'gaming' and data errors

The final section of the quantitative IPE will involve a descriptive analysis of 'gaming' and data errors. Table 18 summarises the metrics that will be used to conduct this analysis. These metrics will be operationalised and collected from HUL and our analyses rely on HUL sharing these metrics.

Table 18. Metrics underlying the 'potential gaming and data errors' analysis.

Metric	Data source	Number of timepoints	Number of groups	Which groups
% of consenting participants with unreasonable body measurements (e.g. BMI over 100 kg/m2) at any timepoint from randomisation to week 24 (in %)	In-app survey	1 (total timespan)	5	Total and by study arm
% of non-validated points earned by study arm from randomisation to week 24 (in %)	In-app data	1 (total timespan)	5	Total and by study arm
% of non-validated points earned by study arm, per month from randomisation to week 24 (in %)	In-app data	5 (monthly)	5	Total and by study arm
Subgroup analysis on the % of non-validated points earned by study arm from randomisation to week 24	In-app data	1 (total timespan)	80	Total and by study arm. Each respectively by deprivation status, ethnicity, sex, age, baseline F&V intake, baseline MVPA.

Summary statistics will be calculated using data on the above metrics for the entire participant sample and by study arm. Additionally we will conduct subgroup analyses on the % of non-validated points earned by study arm from randomisation to week 24 to understand if potential gaming and data error varies by deprivation status, ethnicity, sex, age, baseline F&V intake, baseline MVPA. This section of the quantitative IPE will help to quantify the extent of 'gaming' and data errors in the intervention.

4.3 Qualitative Methods

4.3.1 Methods overview

The qualitative component of the IPE will provide rich evidence of individuals' perspectives and experiences of the intervention's implementation to complement the broad insights obtained through quantitative methods.

BIT has identified five target populations who can provide partial perspectives on intervention experience and these will be combined for a comprehensive interrogation of the research questions:

- "**Fully engaged" recipients**: people who have accessed the app at least once every 30 days for the duration of the trial.
- **"Drop out" recipients**: people who signed up to the intervention but have not used the app for more than 30 days consecutively.
- **Non-recipients**: people who were targeted by HUL, but chose not to participate in the intervention. This consequently excludes individuals who were on the waiting list to join the intervention.
- **Delivery partner**: employees of HUL who have been involved in the design and delivery of the intervention.
- **Reward partners**: employees from both corporate and local partners who have been involved in the design and delivery of the incentives to recipients.
- **City of Wolverhampton Council (CWC):** members of staff at WCW - both in leadership positions and frontline positions - to understand their experiences of supporting the implementation of the scheme.

We will use purposive sampling to capture the views of a diverse range of people from the target populations listed above. Selecting participants based on particular characteristics (see sampling frames in Section 4.3.2.1 for more detail) will ensure that the full range of relevant groups are included in the data collection. This enables us to capture a diverse set of perspectives and experiences.

For each population, BIT has selected a qualitative methodology that will best enable the IPE research questions to be addressed (see Section 4.3.2.3 for detail on methodologies). Namely, BIT will conduct: (i) interviews ***and*** focus groups with recipients (fully-engaged and drop-outs) and non-recipients of the intervention; (ii) focus groups with the delivery partner (HUL); and, (iii) focus groups with reward

partners. The key research questions the qualitative component of the IPE will focus on are:

Table 19. Research topics and questions

Research Topic	Research question	
Reach What factors affected the intervention's reach?	1.4	What was the role of marketing and communications in motivating participants to join the intervention?
	1.5	How is the intervention perceived by recipients, and non-recipients of the financial incentives?
	1.6	What are the barriers and facilitators to the intervention's reach?
Engagement What factors affected the engagement of users with the intervention?	2.3	In what ways did users engage with the app after signing up?
	2.4	What were participants' experiences and perspectives of the intervention?
	2.5	What are the barriers and facilitators to engagement?
Mechanisms of impact Through what mechanisms does the intervention affect behaviour change?	3.2	What barriers and facilitators - both contextual and individual - affect the extent to which the intervention changes behaviours for recipients and non-recipients of the financial incentives?
	3.3	How do features of the incentive scheme affect (or not) the extent to which the intervention changes behaviours?
Implementation and feasibility How was the intervention implemented and is it scalable?	4.1	What was the process for developing and implementing the intervention among delivery and reward partners?
	4.2	Are the design and delivery processes fit for scaling and sustaining the intervention?
	4.3	How was the intervention implemented?
	4.4	What are the facilitators and barriers to scaling and sustaining the intervention (including financial incentives) beyond the pilot?

Our approach to conducting this fieldwork will be underpinned by three qualitative strategies: triangulation to assess the qualitative research findings' credibility,[77]

[77]Patton M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. Health services research, 34(5 Pt 2), 1189–1208. Triangulation is a strategy to assess qualitative findings' credibility by enabling "cross-data validity checks". Two forms of triangulation will be employed in this study. Methodological triangulation will be used to establish the findings' validity by testing conclusions across different methods. This will be achieved both through the combination of qualitative and quantitative research and the use of different qualitative techniques (i.e., focus groups and interviews). For example, the quantitative IPE will help us understand "what" happened, whilst the qualitative IPE will help answer "why". Triangulation of sources will build a richer picture of the pilot's implementation by combining data from different users (e.g., recipients of the intervention, reward partners and implementation partners).

ongoing assessment of thematic saturation[78] to guide fieldwork priorities and the use of the Theoretical Domains Framework[79] (TDF) to ensure a comprehensive assessment of factors affecting the intervention's implementation. Anonymous quotations or written summaries of participants' responses may be included in the final report, presentation or other deliverables. However, all identifiable information - including, names, roles etc - will be removed.

Once data collection is complete, BIT will use thematic analysis across the qualitative data to detect themes, patterns and key ideas (see Section 4.4 for more details). We will take an inductive, data-driven approach to analyse patterns and develop themes closely linked to the data. This will be complemented by a contextualist method that takes into account the individual perspective, as well as the social context into account, while maintaining focus on the data collected. Table 20 provides an overview of the data sources, collection methods and analysis strategy BIT will undertake to answer the key research topics for each population .

Table 20. Summary of Qualitative IPE research activities

Research topic	Data collection methods		Data sources	Data analysis methods
1. Reach 2. Engagement 3. Mechanisms of impact 4. Implementation and feasibility	Intervention Recipients / Non-recipients	34 Semi-structured interviews	12 x Fully engaged intervention recipients 12 x Dropout intervention recipients 10 x Intervention non-recipients	Thematic analysis
		3 Focus groups	1 x Fully engaged intervention recipients *(4 participants - 1 per arm)* 1 x Dropout intervention recipients *(4 participants - 1 per arm)* 1 x Intervention non-recipients *(4 participants)*	Thematic analysis
4.Implementation and feasibility	Delivery partner	2 x focus groups	2 x focus groups *(2-5 participants per focus group)*	Thematic analysis
4.Implementation and feasibility	Reward partners	3 x focus groups	2 x focus group with national supermarket	Thematic analysis

[78] Green, J., & Thorogood, N. (2004) Qualitative methods for health research (2nd ed., pp. 198-202). London: Sage Publications. Qualitative findings are not intended to be objective or statistically generalisable. Rather, they aim to uncover the meaning of people's behaviour as defined by themselves. Thematic saturation is achieved when further data collection or analysis do not surface new information to elucidate the research questions. Periodic assessment of saturation at the group level to assess whether no new themes are emerging in new interviews). This will allow BIT to strategically target our qualitative research to address thematic gaps.

[79] Michie, S., Johnston, et al., & "Psychological Theory" Group (2005). Making psychological theory useful for implementing evidence based practice: a consensus approach. Quality & safety in health care, 14(1), 26–33. The TDF identifies twelve domains that explain behaviour change (e.g., knowledge, skill and beliefs about capabilities). Understanding the role of these domains in behaviour change can provide a more comprehensive understanding of the factors that may serve as a barrier or facilitation to a behaviour. The TDF ensure that the interview and topic guides fully address all factors contributing to reach (e.g., social influences of peers on sign-up), engagement (e.g., the extent to which rewards successfully reinforce intentions) and mechanisms of success (e.g., people's belief in their ability to lose weight).

			(2-3 participants per focus group) 1 x national gyms *2-3 participants per focus group)*	
1. Reach 4. Implementation and feasibility	City of Wolverhampton Council	2 x focus groups	2 x focus group with Local Authority *(4-5 participants per focus group)*	Thematic analysis

Fully engaged intervention recipients* are recipients **who remained engaged throughout the pilot and did not drop-out. These interviews and the focus group will be conducted at the end of the pilot.

Since each population has a slightly different journey through the research process, the sections below outline the sampling frame, recruitment strategy and data collection methods for each population (i.e., (i) recipients and non-recipients, (ii) delivery partner and (iii) reward partners) separately.

4.3.2 Data collection with intervention recipients and intervention non-recipients

This section outlines (i) sampling frame, (ii) recruitment strategy and (iii) data collection methods for recipients and non-recipients.

4.3.2.1 Sampling for recipients and non-recipients

Using purposive sampling we will select participants based on two primary characteristics: their level of engagement with the scheme (fully engaged, drop-out or non-recipient) and the treatment arm (control and level of incentives).

We also identified a number of desirable secondary characteristics (e.g, BMI, gender and level of deprivation) which are hypothesised to influence perspectives or experiences of the intervention. The sampling frame below specifies the recruitment targets for both primary and secondary characteristics.

Table 21. Sampling

Minimum sample:46	**16 Fully engaged intervention recipients:** ● 12 Intervention recipients Interviews (3 per treatment arm) ● 1 focus group of intervention recipients (4 participants - 1 per arm) **16 Dropout intervention recipients:** ● 12 Intervention recipients Interviews (3 per arm) ● 1 focus group of intervention recipients (4 participants - 1 per arm) **14 Non-intervention recipients:** ● 10 Interviews ● 1 focus group (4 participants)
Primary criteria	**Target**

Scheme participation	Fully engaged	16
	Dropout	16
	Non-intervention (recipient targeted by HUL to join scheme but who did not sign up)	14
Treatment Arm	Control (no financial incentive)	8
	Low incentive	8
	Medium incentive	8
	High incentive	8
	Non-intervention (recipient targeted by HUL to join scheme but who did not sign up)	14
Secondary criteria (desirable)		**Target**
Baseline BMI	BMI ≥ 25	3 per Treatment Arm
	BMI < 25	3 per Treatment Arm
Gender	Male	2 per Treatment Arm
	Female	2 per Treatment Arm
	Non-binary	1 per treatment arm
Ethnicity	White	1 per treatment arm
	Black	1 per treatment arm
	Asian	1 per treatment arm
	Mixed	1 per treatment arm
	Other	1 per treatment arm
Deprivation	Index of Multiple Deprivation (IMD) ≤ 2	3 per Treatment Arm
	Index of Multiple Deprivation (IMD) > 2	3 per Treatment Arm

The minimum sample size is based on an estimate of the number of participants required for the IPE to achieve thematic saturation, triangulation of findings and a comprehensive understanding of the barrier or facilitation to a behaviour under the TDF.

Once the target minimum sample for recipient and non-recipient interviews has been reached, BIT will:

1. **Employ rapid thematic analysis and triangulation:** BIT will use this to assess whether saturation has been reached and identify any research topics or populations where further qualitative research might reveal new themes.

2. **Use findings from the rapid analysis to inform the ongoing qualitative strategy:** based on the results of the rapid analysis, BIT will conduct:

    a. **Focus groups to confirm saturation hypotheses:** focus groups with recipients and non-recipients will provide an opportunity to surface new themes and confirm whether the hypothesised saturation and triangulation has been achieved.
    b. **Further interviews to explore specific themes or engage with sub-populations:** if the rapid thematic analysis identifies evidence gaps, BIT will conduct further targeted qualitative research until saturation is achieved.

3. **Repeat steps 2 and 3 until saturation is achieved:** BIT will repeat the process of rapid analyses periodically to assess saturation and continue to conduct qualitative research, if required. The maximum sample size in this scenario will be determined by the remaining capacity in the project budget.

4.3.2.2 Recruitment of recipients and non-recipients

Intervention Participants

The recruitment of intervention recipients (both fully engaged and dropouts) for interviews and focus groups will be led by HUL.

In order to encourage and enable participation, each interview participant will be provided with a £30 voucher by BIT at the end of the interview. Those that partake in a focus group will be rewarded with a £50 voucher. Provided a participant attends the interview, they will receive the gift card regardless of whether or not they complete their interview. Participants will be sent the £30 or £50 Tango e-gift card[80] via email from a member of the BIT team. This e-gift card can be used to buy products from multiple retailers.

HUL will be responsible for inviting recipients to participate in the IPE research activities and collecting consent to pass on their contact details to BIT following the recruitment process below:

1. HUL will send an email inviting recipients (both fully engaged and drop-outs) to express an interest in participating in focus groups or interviews which will include:
    - A sign-up form.

[80] Tangocard (www.tangocard.com) is a virtual gift card that allows the recipient to spend a set monetary value (in this case £30 for interview participants and £50 for focus group participants) on a variety of retailers, such as Amazon, Currys, PC World, Tesco, Cineworld or John Lewis.

    - An information sheet that outlines the purpose of the interviews or focus groups, the topics that will be discussed and how their data will be protected.[81]

2. Having been screened based on sampling characteristics and eligibility, HUL will assign a unique study code ID to participants which will be used to link two separate, secure spreadsheets,[82] containing:
    - Contact information for booking interviews.
    - Demographic information.
3. BIT researchers will send each new participant an invitation email with a link to a Google Sheet or Google Form to book an interview slot / select a focus group slot using their ID. A copy of the information sheet will also be included.
4. The researcher will then send the participant a Google calendar invite including a video-conferencing link (and an alternative telephone number to dial in) from their BIT email account.

Non-recipients (pure control)

The recruitment of intervention non-recipients (i.e., individuals that were targeted by HUL to join the scheme, but chose not to sign up) will be led by an external recruitment agency, Acumen following the recruitment process below:

1. BIT will work with DHSC and Acumen to identify local channels through which to recruit.
2. Acumen will contact individuals not involved in the pilot via an invitation email including:
    - A sign up form
    - An information sheet that outlines the purpose of the interviews or focus groups, the topics that will be discussed and how their data will be protected.
3. Having been screened based on sampling characteristics and eligibility, Acumen will assign a unique study code ID to participants which will be used to link two separate, secure spreadsheets, containing:
    - Contact information for booking interviews.
    - Demographic information
4. BIT will provide Acumen with a list of interview and focus group slots which Acumen shares with participants to select a convenient time. BIT will then send out a zoom/G-meet invitation (including telephone option) to participants and BIT researchers on the selected slot together with an email that contains the information sheet again.

[81] It is important to note that we do not need to collect written consent from interview participants. We collect and audio record verbal consent at the beginning of the interview.
[82] Subject to approval by DHSC, the data will be stored securely throughout the evaluation (in a restricted access Google Drive folder) and deleted 6 months after the end of the project.

4.3.2.3 Collection Method

Interviews for intervention recipients and non-recipients

Interviews will be conducted with individuals who signed up for the intervention as well as with people that did not. The interviews will be semi-structured, following a topic guide to ensure that the core components of the scheme are covered, whilst still providing an opportunity for interviewees to provide additional insights and feedback.

The interviews with intervention recipients will address all four research topics: reach, engagement, mechanisms of impact and implementation. In particular, the interviews will explore the interviewees' reasons for participating in the intervention, the ways in which the intervention has affected their behaviours, their general views and perceptions of the intervention, and the facilitators and barriers to them sustaining behaviours following the intervention.

For non-recipients, these interviews will explore the reasons that the individual did not choose to participate in the intervention, including a thorough examination of their perceived acceptability of the intervention.

The interviews will employ the Theoretical Domains Framework[83] (TDF) to explore the capabilities, motivation and opportunities of participants to change their behaviour. For example, a recipient's reasons for joining the intervention could be influenced by their:

- Intentions, goals, readiness to change and beliefs about consequences (motivation)
- Knowledge of the programme (capabilities)
- Behavioural regulation and physical skills (capabilities)
- Social influences and environmental context (opportunity)

The topic guides will provide sufficient time to explore each of these domains of behaviour, and, in doing so, will help identify barriers and facilitators to the scheme.

Interviews will last approximately 60 minutes for intervention recipients and 30 minutes for intervention non-recipients.

Focus groups with intervention recipients

Two focus groups will be conducted:

- **1 focus group** with **4 fully engaged intervention recipients**
- **1 focus group** with **4 dropout intervention recipients**
- **1 focus group** with **4 intervention non-recipients**

[83] Cane, J., O'Connor, D. & Michie, S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implementation Sci 7, 37 (2012)

Focus groups will be conducted to complement the in-depth interviews by enabling an opportunity for discussion and the exchanging of ideas between participants, helping participants to further develop their own ideas and to tease out some nuances in experience and perspectives. As noted above, they will play an important role in helping to determine whether thematic saturation has been achieved.

During these focus groups we will be able to confirm or challenge insights obtained from the prior interviews. The focus groups will explore the same research questions as the interviews, and the topic guides developed will also be grounded in the TDF.

Focus groups will last up to 90 minutes. Both focus groups and interviews will be conducted by video call or by telephone and will be audio recorded and transcribed.

4.3.3 Data collection with the delivery partner (HUL)

This section outlines (i) sampling frame (i) recruitment strategy and (iii) data collection methods for the delivery partner (HUL).

4.3.3.1 Sampling for Delivery Partners HUL

Using purposive sampling, we will select participants based on the primary criterion of their role in the design and delivery of the intervention. This will allow DHSC to gain insights from a range of individuals involved in core delivery partner functions related to reach, engagement and implementation.

Table 22 below provides an overview of the primary criteria and estimated recruitment targets required to reach thematic saturation and triangulation for this population.

Table 22. Sampling frame for Delivery Partners (HUL) Focus Group

Maximum sample:5 -10		**5-10 participants:** • 2 focus groups
Primary criteria		**Target**
Role/Job Title	UX and content development team	1-2
	Strategy team	1-2
	Partnership and incentives team	2-4
	Marketing team	1-2

4.3.3.2 Recruitment of delivery partner

HUL will appoint a recruitment lead who will work with BIT to recruit staff for focus groups. The recruitment process will be carried out as follows:

1. Having identified potential participants,[84] HUL's recruitment lead will share an information sheet and sign-up form with them to collect information on the primary sampling criteria.
2. Participants will be screened based on these criteria and given a unique study ID code. This code and a secure spreadsheet with contact and sampling information will be shared with BIT.
3. BIT researchers will send new participants a link to a Google Sheet or Google Form in which they can book a focus group slot using their participant ID. A copy of the information sheet will also be included.
4. The researcher will then send the participant a Google calendar invite including a video-conferencing link (and an alternative telephone number to dial in) from their BIT email account.

4.3.3.3 Collection methods with delivery partner

Two focus groups will be conducted with HUL staff members to understand their experiences of implementing the intervention. BIT will work with HUL to define the composition of each focus group that will best enable a rich discussion of the intervention's implementation.

Focus groups will enable participants to build on each other's ideas and build a rich picture of the interventions' implementation. Given HUL's role, these focus groups will provide crucial insights into the challenges of implementing the intervention, barriers and facilitators to the scalability and sustainability, and important recommendations for sustaining engagement with the intervention following the pilot.

Focus groups will last up to 90 minutes. They will be conducted by video call or by telephone and will be audio recorded and transcribed.

4.3.4 Data collection with the reward partners

This section outlines (i) sampling frame (i) recruitment strategy and (iii) data collection methods for the reward partners (HUL).

4.3.4.1 Sampling for reward partners

Using purposive sampling, we will select participants based on the primary criterion of the type of reward partner (i.e., whether corporate or local). This will allow DHSC to gain insights from a range of providers involved in the delivery of incentives to recipients.

[84] If there are other relevant delivery partners, BIT will schedule additional focus groups with these partners (1 focus group per additional delivery partners).

Table 23 below provides an overview of the primary criteria and estimated recruitment target required to achieve thematic saturation and triangulation for this population.

Table 23. Sampling frame for reward partners

Minimum sample:8 **Maximum sample:10**		**8-10 participants:** ● 2 focus groups with National supermarkets ● 1 focus group with National gyms
Primary criteria		**Target**
Type of partner	National supermarkets	4-6
	National gyms	4

4.3.4.2 Recruitment of reward partners

The recruitment of incentive providers for focus groups and interviews will be led by HUL. They will appoint an internal recruitment lead who will work with BIT to recruit participants. The expected recruitment approach will be:

1. HUL's recruitment lead will reach out to reward partners to express an interest in participating in focus groups. The email will describe the focus groups at the high level and include:
    - A sign-up form
    - An information sheet that outlines the purpose of the interviews or focus groups, the topics that will be discussed and how their data will be protected
2. Having screened participants based on the sampling and eligibility criteria, HUL will assign a unique study ID. This code and a secure spreadsheet with contact and sampling information will be shared with BIT.
3. BIT researchers will send new participants a link to a Google Sheet or Google Form in which they can book a focus group slot using their participant ID. A copy of the information sheet will also be included.
4. The researcher will then send the participant a Google calendar invite including a video-conferencing link (and an alternative telephone number to dial in) from their BIT email account.

4.3.4.3 Focus groups with reward partners

Focus groups will be conducted with reward partners - both corporate and local - to understand their experiences of implementing the scheme. The focus groups will be as follows:

- **2 focus group** with national supermarkets
- **1 focus group** with national gyms

Focus groups were selected for this population because they will allow comparison and contrast of perspectives and experiences of employees involved in a variety of different roles relevant to the design and delivery of the intervention. They will thus provide insights into the challenges of implementing the intervention, barriers and facilitators to the scalability and sustainability, and important recommendations for sustaining engagement with the intervention following the pilot.

Focus groups will last up to 90 minutes. They will be conducted by video call or by telephone and will be audio recorded and transcribed.

4.3.5 Data collection with City of Wolverhampton Council

This section outlines (i) sampling frame (i) recruitment strategy and (iii) data collection methods for the City of Wolverhampton Council (CWC).

4.3.5.1 Sampling for CWC

Using purposive sampling, we will select participants based on the primary criterion of the type of council member (i.e. leadership role or frontline working staff). This will allow DHSC to gain insights from a range of providers involved in the delivery of incentives to recipients.

Table 24 below provides an overview of the primary criteria and estimated recruitment target required to achieve thematic saturation and triangulation for this population.

Table 24. Sampling frame for City of Wolverhampton Council

Minimum Sample: 8 **Maximum sample:10**		**8-10 participants:** ● 1 focus groups with City of Wolverhampton Council Leadership ● 1 focus groups with City of Wolverhampton Council frontline working level staff
Primary criteria		**Target**
Type of partner	Frontline working staff	5-6
	Leadership	3-4

4.3.5.2 Recruitment of CWC

The recruitment of CWC stakeholders for focus groups will be led by BIT and supported by DHSC. DHSC has provided a list of suitable participants based on their role within the council. The expected recruitment approach will be:

5. DHSC will reach out to the CWC staff on their participant list, informing them of the evaluation aim, focus group aims and BITs role
6. BIT will reach out to these staff members via email. The email will describe the focus groups at the high level and include:
    - A sign-up form
    - An information sheet that outlines the purpose of the interviews or focus groups, the topics that will be discussed and how their data will be protected
7. BIT will assign a unique study ID to each participant. This code and a secure spreadsheet with contact and sampling information will be shared with BIT.
8. BIT researchers will send participants a link to a Google Sheet or Google Form in which they can book a focus group slot using their participant ID. A copy of the information sheet will also be included.
9. The researcher will then send the participant a Google calendar invite including a video-conferencing link (and an alternative telephone number to dial in) from their BIT email account.

4.3.5.3 Focus groups with CWC

Focus groups will be conducted with members of staff at WCW - both in leadership positions and frontline positions - to understand their experiences of supporting the implementation of the scheme. The focus groups will be as follows:

- **1 focus group** with staff in leadership positions
- **1 focus group** with staff in frontline positions

Focus groups were selected for this population due to the enhanced role the CWC has played in the implementation of the health incentives pilot, particularly in supporting the pilot to reach the acquisition and engagement targets. Focus groups will allow comparison and contrast of perspectives and experiences of employees involved in a variety of different roles relevant to the delivery of the intervention. They will thus provide insights into the challenges of implementing the intervention, barriers and facilitators to the scalability and sustainability, and important recommendations for sustaining engagement with the intervention following the pilot.

Focus groups will last up to 90 minutes. They will be conducted by video call or by telephone and will be audio recorded and transcribed.

4.3.6 Data collection - timelines

Figure 4 provides an indicative timeline for the qualitative research activities.

For recipients (both fully-engaged and drop-outs), fieldwork will be conducted at the conclusion of the intervention. This will allow a holistic assessment of their experience and ensure that the qualitative research does not affect the impact evaluation.

For non-recipients, interviews will take place during the intervention's delivery so that their reasons for not signing up are fresh in their minds. This will only happen once the recruitment window for the intervention has closed.

Reward partner and delivery partner interviews will be conducted towards the end of the intervention to allow both populations to share insights and challenges faced at all stages of delivery and engagement with recipients.

4.3.7 Risks

Table 24 outlines potential risks associated with the data collection process along with mitigation measures BIT will take. This risk register will be periodically reviewed and updated during the final stages of the design and delivery of the evaluation.

Table 24. Lists of risks and mitigations

Risk Type	Risk	Mitigation
Methodological	Meeting recruitment targets for qualitative evaluation.	We will employ multiple recruitment strategies simultaneously. If we are unable to reach targets, we will relax our sampling criteria. We will work flexibly around participants' schedules to enable participation. We will emphasise to people that their data will remain confidential, anonymous and presented in aggregate, and any personal information will be removed from reports, slides or other deliverables.
	There may be vulnerable participants, such as people with eating disorders	To ensure that participants are fully informed and empowered when taking part in this research before the interview we will send participants an information sheet. This information sheet will contain: an overview of topics that will be covered in the interviews, the research topics, how the interview responses will be used and the research instruments in a way that is understandable to people participating in this research. The information sheet will also contain information where participants can access support based on the DHSC guidelines.

		Researchers will make themselves available to answer questions from participants via the telephone or email. This allows people to make informed decisions about participating in the research and sharing their experiences.
	Special category data is being collected (such as ethnicity). This personal information is likely to lead to increased levels of harm and stress if there is a data breach or misuse of the data, over what might be caused by the release of less sensitive categories of data.	Security controls as outlined in the Data protection and data security checklist are in place and reviewed periodically. Permissions and personnel involvement will be reviewed regularly to ensure access is only granted to the minimum number of people that need it. The risk of a data breach can never be eliminated but the security controls and organisational procedures result in an acceptable level of risk given the personal data in question.
	Difficulties obtaining consent from people	Consent will always need to be provided verbally at the beginning of the interview to the BIT researchers. Additionally, recruiters HUL may decide to collect consent before the interview in an electronic, verbal or written format. Accessible information sheets i.e. the materials can be sent electronically, posted in person or read aloud verbally by researchers and/or coaches. Ensure consent is granular, and participants can consent to some forms of data collection and processing and not others, if they wish.
	Guaranteeing confidentiality when safeguarding issue is disclosed	In line with BIT's internal safeguarding procedures, participants will be provided with an information sheet outlining sources for support available to participants (e.g, mental health). Further, they will be informed at the start of the interview that, whilst the information shared will remain anonymous, confidentiality may be broken if something raises concerns for their safety or someone else's. In this scenario, BIT will share these details with the Wolverhampton City Council safeguarding lead who will escalate according to their safeguarding policies.
	A participant becomes distressed	Specific mitigations for interviews may include regular 'check-ins' to give participants the opportunity to say or type in a chat if they would like to take a break or stop the interview. We will also provide the option for participants to turn off their video if conducting a video interview. If videos remain on or data collection is conducted face to face, interviewers will be mindful of body language that indicates discomfort with the research. We will also signpost participants to resources provided by HUL.

4.4 Qualitative Data Analysis

Interviews and focus groups will be transcribed and then analysed. A thematic analysis will be employed to code the transcripts and identify emerging themes.[85] These themes will then undergo a further round of classifying, and will be sorted into high-level themes and sub-themes.

A thematic analysis is carried out across the interview and focus group data, following a three-stage process:

1. **Transcripts are coded by research questions within the following** four topics:
    a. Reach
    b. Engagement
    c. Mechanisms of impact
    d. Implementation and feasibility

    This first stage is a 'low-inference and descriptive' process of data management.

2. **Data is coded by themes that respond to the research questions**: these themes are identified both deductively and inductively, using constructs from the literature when supported by evidence, and creating new constructs where it does not. When appropriate, we will also use the TDF to help categorise the themes to inform findings and adopt a contextualist method that takes into account the individual perspective, as well as the social context.

3. **Themes are refined:** this is achieved by reviewing their relation to each other, grouping them into conceptual categories where possible, and ensuring that they comprehensively cover the data.

The predetermined topics of the interview guide will be used to interrogate the data, maintaining a balance between deduction (using existing knowledge and the research questions to guide analysis) and induction (allowing concepts and ways of interpreting experience to emerge from the data).

Researcher bias will be mitigated by using the interrater reliability checker on NVIVO, ensuring that multiple researchers are coding transcripts in the same way for a sample of the transcripts. In the event of discrepancies, researchers will also meet to discuss any discrepancies and agree on a code-book. Verbatim participant quotations and case examples will be used to provide evidence and exemplify the theme(s) discussed in the paragraph before the quotation. Any quotations used will be selected by the qualitative researchers who conduct the data analysis on the basis of how well they exemplify the theme(s) discussed.

[85] This approach will follow an adapted version of Virginia Braun & Victoria Clarke (2006) Using thematic analysis in psychology, Qualitative Research in Psychology, 3:2, 77-101, DOI: 10.1191/1478088706qp063oa

6. Limitations and generalisability

The effect sizes of the impact evaluation will be specific to the financial incentives component of the app: As participants assigned to the control group will have an experience that is identical to that of users in the intervention groups, except for the financial incentives themselves, this trial won't generate evidence on the impact of the app alone or the impact of the app+financial incentives against no intervention. We believe that these questions are important, but outside the scope of this project.

The trial relies on participants accurately self-reporting their dietary behaviours, in addition to their weight at the baseline and each measurement time point. As a result, effect sizes estimated for dietary outcomes could be an overestimate of the true effect of offering financial incentives if participants in the intervention groups are more motivated to report good behaviours. This could be because of desirability bias, erratic eating habits (which make it harder for participants to recall what they have eaten accurately) or because they have an implicit motivation to pay more attention to what they eat (hence being more likely to overcome the typical under-reporting observed in food recall questionnaires). In addition, providing self-report data is effortful and may increase the risk of data attrition throughout the trial.

Providing control group participants with a wearable device and app-based measurements may itself improve their health behaviours through regular monitoring and feedback. This may lead to the evaluation underestimating the scheme's impact when compared to the true "business-as-usual". However, we are confident that this is a worthwhile trade-off as it enables the collection of objective, behavioural outcome metrics in a robust RCT design. It is also the only design which enables isolating the impact of incentives in the context of a digital scheme.

Offering participants 10 diet-related health challenges from which they could freely choose from might dilute the treatment effect across the dietary primary outcomes. The rationale for this limitation is that trying to change too many diet behaviours might then dilute the size of the impact on any of these behaviours. For instance, if all of the challenges were focussing on fruit and vegetables, we would have a narrower impact but might expect greater effects on this particular behaviour.

The trial will take place exclusively in Wolverhampton and might not generalise to other locations. Findings from the pilot can be extended to the other UK locations with high prevalence of low SES population under the assumption that (i) the advertisement campaign and the recruitment process in this pilot are representative of (or similar to) what would happen if they were launched at a national scale (ii) the population of Wolverhampton is representative of the low SES population of the UK (e.g. not more/less skewed towards a particular age group, or towards higher/lower

BMI) (iii) users in Wolverhampton respond to financial incentives similarly to how the average UK participant would.

7. Appendices

Appendix A - Alternatives to Intake24

	Method & description	Pros	Cons
Subjective measures *(short time period)*	**STRONGLY RECOMMENDED** **Automated 24-hour dietary recall:** Online questionnaires	Fewer errors; Quick & low burden for participants; Reduced burden for researchers; Cost-effective; Suitable for large studies; Easier to estimate portion size	Relies on accurate recall
	24-hour dietary recall: Open-ended questionnaires administered by a trained interviewer	Relatively quick & low burden; Procedure does not alter food intake; Sensitive to ethnicity-specific differences	Relies on accurate recall; Expensive due to high interviewer burden and data entry for paper-based survey; Coding and conversion of data to nutrients is time and labour intensive; Participants may change answers due to social desirability bias
Objective measures	**Duplicate diet approach**: Collection of duplicate diet sample and direct analysis	Objective (minimises self-report errors)	Expensive; High individual burden; Unsuitable for large-scale studies; Requires specialist laboratory equipment
	Food consumption record: Observation by trained staff	Objective (minimises self-report errors)	Expensive; Highly intensive for researchers; Limited to specific times and places; Observations may alter individuals usual eating patterns
Subjective measures *(long time period)*	**Dietary history**: Questionnaires administered by a trained interviewer about habitual food intake and dietary behaviours	Energy intake and most nutrients can be estimated reasonably accurately; Only 1 interview necessary	Relies on accurate recall; Data quality depends on interviewer skills; May be difficult for those with erratic eating habits Coding and conversion of data to nutrients is time and labour intensive; Participants may change answers due to social desirability
	Food frequency questionnaire: Self- or interviewer- administered questionnaire about the	Low burden; Easy and flexible to administer; Low cost;	Relies on accurate recall; Not comprehensive to all food consumed; Food list may not be reflective of

	frequency with which food items or groups are consumed over set period		the dietary pattern of the population; Hard to classify pre-prepared meals

Appendix B - Additional information on power calculations

All calculations have been carried out with the statistical software R version 4.1.0, using the packages “tidyverse”, “pwr”, and “data.table”, “ggplot2”.

Allocation ratio across arms

To answer the primary research question, comparisons will be made between all incentive arms (pooled) and the control arm. To optimise power in such a scenario, we have assumed *unequal allocation* of participants to the control and intervention arms at the analysis, such that, with 3 intervention arms, for every 1 participant in each intervention arm there will be 3 participants in the control arm.[86]

Individual or household-level randomisation?

Randomising participants at the individual level or at the household level (i.e. clustered) is a key trial design decision which has implications for both statistical power and the estimation of treatment effects (due to the risk of spillover effects).

With individual-level randomisation, spillover could arise because individuals within a household that are assigned to different trial arms can influence each other’s outcomes (e.g. a household member in the intervention arm can encourage another member in the control arm to do more PA). This can lead to underestimation of the treatment effect. Household-level randomisation minimises the risk of within-household spillover effects because all household members are assigned to the same arm.

With household-level randomisation, it is also important to consider that household members have outcomes that are likely to be non-independent or correlated (e.g. families tend to be active together), and therefore each individual contributes less information to the analysis.[87] The extent of this non-independence or correlation is quantified using the *intra-cluster correlation* (ICC).

The extent to which clustering at the household level affects statistical power depends on the ICC and the sizes of clusters (households) in the trial. In the current trial, cluster size depends on the household sizes signing up to the trial, and on the proportion of people within each household that sign up to the trial. Therefore, there are three key questions that should be answered:

[86] Torgerson, D. (2008). *Designing randomised trials in health, education and the social sciences: an introduction*. Springer. Chapter 10, pp. 108-113.
[87] This is a distinct issue from spillover risk, but both aspects arise from the same underlying issue: household members are not entirely independent in their behaviour and their outcomes.

- **What is the intra-cluster correlation (ICC) within households for the PA and dietary outcomes?** We estimate that intra-cluster correlation is sizable, especially for diet - it could range from 0.2 to 0.5, with this assumption guided by estimates from prior literature.[88] Higher ICCs indicate a stronger association of behaviours among household members. We believe that an ICC of 0.2-0.4 could be realistic for this trial as published estimates based on the Health Survey for England data show an ICC of 0.26 for households, for a binary variable of "moderately/vigorously active in sports"; it also shows ICCs of 0.33 and 0.53 for the binary variables "eats fruit at least 1/day" and "eats vegetables at least 1/day", respectively.[89]

- **Within each household, what is the proportion of people that are expected to sign up?** We assume that if a person within a household signs up, then all the other adults in the household will sign up too (conservative assumption). There is no available data on the proportion of adults within a household that are likely to sign up to the trial, and whether this varies by household size (e.g. whether within 2-person households, on average 2 adults are likely to sign up, whether in 3-person households only 1 adult is likely to sign up...etc.). Therefore, currently we assume that all adults within a household will sign up; this is a conservative assumption as it increases the required household sample size, owing to the need for cluster-randomisation.

- **What is the distribution of household sizes that are expected to sign up?** We estimate that 60% of trial participants could be in the same household with at least another trial participant (conservative assumption). There is no available data on the household sizes that are likely to sign up to the trial (i.e. whether the recruitment campaign will somehow be biased towards small or large households). Therefore, we assume that the household sampling will be random, guided by the distribution of household sizes in Wolverhampton. To obtain an estimate of this distribution, we used Census 2011 data on the distribution of household sizes and household composition in Wolverhampton[90]. The average household size (including children) in Wolverhampton is 2.4 people; the table below shows the distribution of the number of adults per household and proportion of the total number of households (102,177 households in total).[91] For example, in a sample of 1000 participants, we assume that 39% will come from single-person households

[88] A Danish study estimated an ICC of 0.07 for accelerometer-measured MVPA among families with children (Petersen et al. 2020). Estimates based on the Health Survey for England data (1994) show an ICC of 0.26 for households, for a binary variable of "moderately/vigorously active in sports"; it also shows ICCs of 0.33 and 0.53 for the binary variables "eats fruits at least 1/day" and "eats vegetables at least 1/day", respectively (Gulliford et al. 1999).

[89] Gulliford et al. 1999. Health Survey for England data refer to 1994.

[90] https://www.ukcensusdata.com/wolverhampton-e08000031#sthash.0ov8DqyG.dpbs

[91] Various observations suggest that household size in the trial may be on the lower end: Children in larger households are less likely to have high BMI (Dasgupta & Solomon, 2018); children of single-parent households are at a higher risk of obesity (Duriancik & Goff, 2019); the proportion of people who are inactive/overweight increases with age (https://commonslibrary.parliament.uk/research-briefings/sn03336/) while average household size is higher for those 30-49 years old, and then decreases (ONS 2019).

(389 people), 41% will come from households with two adults (409 people across ~204 households), 20% will come from households with at least three adults (202 people across ~67 households).

Table 25. Distribution of the number of adults per household in Wolverhampton (Census 2011 data).

Adults per household	Number of households	% of total (102,177 households)
1	39,777	39%
2	41,784	41%
3+ *	20,616	20%
* set to 3 for simplicity in estimations as Census 2011 data does not provide further granularity		

Adjusting for multiple comparisons

For the primary outcomes, we will adjust for multiple comparisons using the Benjamini-Hochberg procedure within the physical activity outcomes (2 comparisons) and dietary outcomes (4 comparisons) separately. For the exploratory outcomes and questions, we will adjust for multiple comparisons separately within each research question and separately for physical activity and dietary outcomes. For example, when examining effects after 1 month on the primary outcomes, we will perform 2 comparisons for the physical activity outcomes and 4 for the dietary outcomes. We will do the same for the analysis at 3 months.

Baseline measures and other covariates

We assume that the inclusion of covariates such as baseline calorie intake, baseline MVPA min/day, socio-demographic information, and anthropometric information will reduce variance in the outcome measures, and therefore increase power. Based on prior research, the pre-post correlation in the calorie intake measure using Intake24 is expected to be ~0.5; for the measure of MPVA min/day this is expected to be ~0.8.[92] Therefore, we conservatively assume that a correlation coefficient of at least 0.5 is realistic in this trial.[93]

[92] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6722486/; https://journals.humankinetics.com/view/journals/jpah/8/5/article-p668.xml

[93] Note that our estimates are additionally conservative because we do not explicitly factor in multiple measurements per participant during both baseline and outcome timepoints (e.g. participants will be asked to complete two Intake24 measurements during baseline, and will be asked to wear the wearable device every day). We have chosen to not factor this in due to the uncertainty around compliance with the Intake24 and daily wearing of the wearable devices.

Appendix C - Recommended dietary intake according to the Recommended dietary intake according to the Government Dietary Recommendations

Outcome	Current intake	Desired change
Carbohydrates	212 g/day	Increase
Free sugars	51 g/day	Reduce
Total Fat	69 g/day	Met
Saturated fat	25 g/day	Reduce
Fibre	18 g/day	Increase
Salt (Sodium)	5 g/day (Sodium 1985 mg/day)	Reduce
Fruit and vegetables	297 g/day	Increase
Red and processed meat	52 g/day	Reduce

Appendix D - Possible additional analysis subject to contractual agreement

D1. Additional measurement time point at the 12m mark

In addition to the 4 data collection points, we might add another time point at the 12m mark if the pilot programme is extended by DHSC. If the extension takes place, an updated evaluation plan and ethics approval will be sought before data collection. This extension would be subject to contractual agreement.

D2. Sensitivity analysis

We will conduct the following sensitivity analysis subject to contractual agreement. We will test the sensitivity of the results of the physical activity and diet primary outcomes to alternative methods of missing data management (multiple imputation and delta adjustment).

Simple imputation:

- **Physical activity**: Within each combination of data collection point (baseline, 1 month, 3 months, 5 months), treatment arm (no / low / medium / high incentives) and day of the week, for reads based on at least 6 hours of wear-time, we will identify the 2.5th and 97.5th percentiles for each physical activity outcome. We will replace reads below this 2.5th percentile (no matter what their wear-time is) with the 2.5th percentile. Similarly, we will replace reads above the 97.5th percentile with the 97.5th percentile. This means we can use a full dataset among non-attrited participants with minimal assumptions around the structure of missing data.
- **Diet**: Within each combination of data collection point, treatment arm and gender, we will replace values of each primary diet outcome below the 1st percentile with the 1st percentile, and values above the 99th percentile with

the 99th percentile. We will still exclude administrations that fail any of our attention checks (e.g. completion time of under 3 minutes).

Multiple imputation: If the results differ between the main specification and the simple imputation method above, we will perform multiple imputation with delta adjustment sensitivity analysis.

- **Physical activity**: Having replaced invalid daily reads within the measurement week using valid reads from up to two weeks before/after, we will generate *n* imputed datasets, where *n* is the percentage of incomplete cases rounded up to the nearest integer (following the rule of thumb suggested by Bodner (2008)[94] and White et al. (2011)[95]). We will use sequential predictive mean matching to impute missing values of physical activity outcomes within each dataset (imputing the outcomes together). As predictors, we will use other reads in the measurement week, all covariates in the main specification, and household- and person-level fixed effects.

  We will perform multiple imputation only for the sample of individuals who have at least one valid read in the measurement week. We will then estimate the following equation on each imputed dataset for a range of (fixed) Δ:

$$(3)\ y_{ijd} = \alpha + \beta\, Treatment_{j} + \gamma_{d} + \gamma_{W} week_{ijd} + \gamma_{B} baseline_{ijd} + \gamma_{M} missing.baseline_{ijd} +$$

$$\dots \Delta missing.outcome_{ijd} + \gamma_{X} X_{ij} + \delta_{Cj} + \delta_{Pij} + u_{ijd}$$

  $missing.outcome_{ijd}$ is a dummy variable which equals 1 if the outcome is missing and 0 otherwise. We will pool the estimated coefficients and standard errors across the 25 imputed datasets using Rubin's rules[96].

  Under $\Delta = 0$, we are assuming that data are missing at random (MAR). This delta adjustment sensitivity analysis informs us to what degree the imputed data could be underestimating the outcome while the findings still hold, for the sample of individuals who had at least one valid read in the measurement week. In other words, it does not capture bias from (differential) attrition.

- **Diet**: We will impute primary diet outcomes below the 1st percentile or above the 99th percentile. As predictors, we will use values of the outcomes in other measurement weeks, all covariates in the main specification, and household- and person-level fixed effects. Again, we will exclude administrations that fail any of our attention checks (e.g. completion time of under 3 minutes).

---

[94] Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal* (15) 651-675.
[95] White, I. R., Royston P. and Wood A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* (30) 377-399.
[96] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc.

D3. IPE, Quantitative methods

Subject to contractual agreement, we will analyse metrics in

- Table 16 (last metric), section 4.2.2
- Table 17 (last 4 metrics), section 4.2.3
- Table 18 (last metric), section 4.2.4

according to subgroups baseline F&V intake (2 groups), baseline MVPA (2 groups).