Evaluation of a financial incentives scheme to support healthy eating and physical activity

BIT

June 2025

BIT is a global research and innovation consultancy which combines a deep understanding of human behaviour with evidence-led problem solving to improve people's lives. We work with all levels of government, nonprofits and the private sector, applying behavioural science expertise with robust evaluation and data to help clients achieve their goals.

Find out more at bi.team

If you'd like this publication in an alternative format such as Braille or large print, please contact us at: info@bi.team

Authors
Dr Giulia Tagliaferri (lead author), Dr Helen Brown, Niall Daly, Bram Reitsma, Alice Worsley and Hazel Wright

Click the **BIT** logo on any page to return to the Contents

Abstract

The Financial Incentives Scheme pilot, developed by HeadUp Labs and evaluated by the Behavioural Insights Team (BIT), aimed to assess whether offering financial rewards could motivate improving physical activity and dietary habits. Through an app, participants engaged in personalised physical activity and diet challenges for 20 weeks. A randomised controlled trial with four reward arms, including control, tested the impact of varying financial incentives on behaviour. Key outcomes measured included moderate-to-vigorous physical activity (MVPA), steps, fruit and vegetable consumption, fibre intake, saturated fat, and free sugars. The study found modest, statistically significant improvements in diet-related outcomes, notably fruit and vegetable intake (+21 grams/day) and fibre intake (+0.35 grams/day), with greater impacts observed in the high-reward group. However, financial incentives had no statistically significant effect on physical activity levels. Sensitivity analyses were conducted for the physical activity data using data from a fitness tracker worn over a longer period of time. This showed a statistically significant impact on both MVPA (+1.9 min/day) and steps (+256 steps per day). Subgroup analyses revealed disparities in outcomes based on deprivation, age, and incentive levels. The implementation process evaluation highlighted demographic and social influences on participation and engagement, particularly in older and less deprived individuals. While limitations exist, the pilot suggests that financial incentives may promote healthier dietary behaviours.

Author information

Dr Giulia Tagliaferri is a Principal Research Advisor at the Behavioural Insights Team [contact: giulia.tagliaferri@bi.team]

Dr Helen Brown is the Director of Health and Wellbeing at the Behavioural Insights Team

Niall Daly is a Research Advisor at the Behavioural Insights Team

Bram Reitsma is a Research Advisor at the Behavioural Insights Team

Alice Worsley is an Advisor at the Behavioural Insights Team

Hazel Wright is a Principal Advisor at the Behavioural Insights Team

To cite this paper

Tagliaferri G., Brown, H., Daly, N., Reitsma, B., Worsley, A. and Wright, H. (2024). Evaluation of a financial incentives scheme to support healthy eating and physical activity (BIT Working Paper Number 008).

Contents

Glossary **7**

Executive summary **11**

Background **20**

1.1. Health and behaviour change 20

1.2 The evidence for incentives 21

1.3 The project 22

1.4 Our role 22

1.5 Evaluation objectives 23

2. Intervention **25**

2.1 The Theory of Change 25

2.2 The Financial Incentives Scheme 26

2.3 How did the challenges work? 29

2.4 What was the role of points? 30

3. Impact evaluation methodology **36**

3.1 Objectives 36

3.2 Outcome measures 38

3.3 Sample & Randomisation 44

3.4 Analysis plan 46

4. Impact evaluation findings **48**

4.1 Participants 48

4.2 Impact 61

5. Implementation and process evaluation methodology **79**

5.1 Objectives 79

5.2 Study design 80

5.3 Quantitative methods & analysis 83

5.4 Qualitative methods & analysis 84

6. Implementation and process evaluation findings **88**

6.1 Reach 88

6.2 Engagement 93

6.3 Mechanisms of impact 125

6.4 Implementation and scaling 130

7. Discussion **138**

7.1 Overall findings 139
7.2 Recruitment and engagement 140
7.3 Does offering financial incentives improve PA and diet? 141
7.4 How does the impact vary over time? 148
7.5 What is the optimal financial incentive level? 148
7.6 Did all participants respond equally to financial incentives? 150
7.7 Conclusions 152
7.8 Recommendations 155
8. Limitations **164**
8.1 Generalisability (external validity) of the findings 164
8.2 Robustness (internal validity) of the findings 165
9. Appendix **171**
Appendix A: Eligibility Criteria 171
Appendix B: Detailed information about outcome variables 172
Appendix C: Balance checks 177
Appendix D: Additional tables on reach and engagement 180
Appendix E: IPW methodology 184
Appendix F: IPE Research topics and methods 186
Appendix G: Qualitative data analysis procedure 204
Appendix H: Full regression outputs 205
Appendix I: Obstacles to user engagement 244
Appendix J: Additional Quant IPE findings 246
Appendix K: Data collection methods and data quality management 255
Appendix L: Trial information sheet 264
Appendix M: References 272

Glossary

Active users (of the app): participants who have opened (logged in) the app at least once in the previous 30 days. Note that this is not intended to convey whether participants were more or less physically active.

Churn-out (users of the app): at any given point in time, an app user who had not opened the app in the previous 30 days was considered "churned-out" (i.e. no longer actively using the app). The inverse of "active users" at any given point in time.

Cluster randomised trial (CRT): a trial in which pre-existing groups, called clusters, of individuals are randomly allocated to trial arms. In this pilot, each household was considered a cluster. CRTs can be used when individual randomisation to trial arms is not possible or the intervention is naturally applied to a whole cluster.

Cognitive: of, relating to, being, or involving conscious intellectual activity (such as thinking, reasoning, or remembering)

Context-aware approach: refers to the capability of systems (like software applications, services, or devices) to adapt their operations, functionalities, or content based on the contextual information about their environment or user. This approach is significant in creating intelligent and responsive systems that can provide more personalised and efficient experiences for participants. A simple example of a context-aware system could be a smart home setup that adjusts lighting, temperature, and music based on who is at home, the time of day, and even the activity of the occupants.

Drop-off (between consent and randomisation): a measure of consenting individuals not moving from one stage of the onboarding process (e.g. consenting to participate) to the next (e.g. completing registration).This term is only used in sections 4.1.1 and 4.1.2.

Drop-out (users of the app): a term used exclusively to refer to data attrition, where users stopped providing data from one data collection point to the next (e.g. from M3 to M5). This term appears mostly in sections related to the impact evaluation and Appendix E, and is not synonymous with 'churn-out' (see above definition).

External validity: the extent to which the findings of a study involving certain situations, people, settings, and measures can be generalised to other contexts, for example other local authorities, or countries.

Healthfulness: defined [within this study] as participants increasing levels of physical activity (PA) in line with the Chief Medical Officers' (CMOs) guidelines, and improving their diet in line with the Government's Eatwell Guide.

Health-neutral: actions or systems that neither significantly improve nor harm health outcomes. They do not contribute to disease risk but also do not actively promote better health.

Health-promoting: actions or systems that actively contribute to improving physical, mental, or social well-being, making individuals more likely to experience better health outcomes. For example, initiatives that encourage regular physical exercise or the consumption of nutritious foods.

Intention-to-treat approach: a method for analysing results in a prospective randomised study where all participants (for which outcome data are available) who are randomised are included in the statistical analysis and analysed according to the trial arm they were originally assigned, regardless of what treatment (if any) they received. This is in contrast for example to a treatment on the treated analysis, where only participants who made use of an intervention (e.g. redeemed financial incentives) are considered treated.

Internal validity: the degree of confidence that the causal relationship being tested is not influenced by other factors or variables.

Inverse probability weighting (IPW): Inverse Probability Weighting (IPW) is a technique used to minimise selection bias in long-term studies. It operates by estimating the likelihood that participants will continue to be part of the study based on initial data collected. Then, it adjusts the importance of the data from participants who remained in the study. This adjustment ensures that the remaining data is representative and balanced, considering the initial information, across trial arms in the study. Appendix E provides technical information on the methodology.

Linear mixed model: an extension of simple linear models in quantitative statistical analysis, used to analyse data where observations can be correlated or grouped in some way.

MVPA: Moderate and Vigorous Physical Activity (MVPA): defined in Chief Medical Officers' Physical Activity Guidelines as activities that can be done at different intensities. They can be differentiated by the 'talk test': being able to talk but not sing indicates moderate physical activity, while having difficulty talking without pausing is an indication of physical activity being vigorous. This study focuses on daily MVPA minutes as one of the physical activity primary outcomes because cumulative MVPA minutes have been shown to provide various clinical health benefits.

Pooled reward arms: combining or "pooling" of the low, medium and high reward arms.

Prospective randomised trial: a type of study where participants are randomly assigned to one of two or more groups: at least one group receives the intervention being tested, and another group receives an alternative/standard intervention for comparison. This is "prospective" because it follows participants forward in time from the point of their randomisation into the future, observing outcomes as they occur.

Purposive sampling: a non-random method of sampling where participants are selected for inclusion in a sample based on their characteristics.

P-value: a statistical measure that helps determine the strength of the evidence against the null hypothesis, which is that there is no effect or no difference. The null hypothesis is the default position. Lower p-values indicate stronger evidence against the null hypothesis, suggesting that the observed data are unlikely to have occurred by chance alone.

Sensitivity analysis: demonstrates how various sources of uncertainty in a mathematical or statistical model contribute to the model's overall uncertainty.

Statistical significance: acts as an indicator in data analysis, providing a quantifiable measure to gauge whether observed effects in a study are likely to be due to the intervention being tested rather than due to chance. This is usually determined by looking at whether the p-value is below a predetermined benchmark, typically set at 0.05. A p-value below this threshold suggests a statistically significant finding, implying the effect is unlikely due to chance.

Theoretical Domains Framework (TDF): a tool designed to help understand why people change their behaviour. By identifying what factors help or hinder behaviour

change, the TDF helps in creating tailored strategies that effectively address these factors. This makes it easier to design interventions that are more likely to succeed, by directly targeting the underlying reasons people do or do not change their behaviour.

Triangulation: the use of multiple methods or data sources in qualitative research to develop a comprehensive understanding of phenomena.

Wearable fitness tracker: smart electronic devices that are typically worn on the wrist and via sensors collect data on the wearer's activity such as steps, distance travelled and physical activity minutes. These devices often link via the internet or Bluetooth to apps on computers, tablets or smartphones.

Executive summary

Background

The Financial Incentives Scheme was designed to test whether offering financial incentives to adult residents of Wolverhampton could increase their levels of physical activity in line with Chief Medical Officers' (CMOs) guidelines, and improve their diet towards UK government dietary guidelines, as depicted in the Eatwell Guide.

As part of exploring how adults can be supported to live healthier lifestyles, the Department for Health and Social Care (DHSC) commissioned HeadUp Labs (HUL) to develop and implement the app-based Financial Incentives Scheme to provide rewards to participants contingent on performing behaviours related to healthy diets and physical activity. This app was known as Better Health: Rewards. The Behavioural Insights Team (BIT) designed and conducted an independent mixed-methods evaluation of the Financial Incentives Scheme including an impact evaluation and an implementation and process evaluation. The trial protocol for this study was peer reviewed and an independent Trial Steering Committee provided recommendations throughout the implementation of the pilot to ensure quality.

Intervention and trial design

The Financial Incentives Scheme was delivered via the Better Health: Rewards app (from herein, the app) and evaluated using an embedded randomised controlled trial (trial protocol registration number: ISRCTN10465935). Participants voluntarily chose to participate in the trial following exposure to an extensive marketing campaign within Wolverhampton. Participants did so by downloading the free app and registering. Participants had to be 18 years or over and reside in a Wolverhampton postcode. Physical activity data were collected via a wearable fitness tracker (all participants could order a free one via the app if they did not already own one) and dietary intake data collected via Intake24 dietary recall surveys within the app. To be involved in the trial, participants had to complete a mandatory baseline period. The purpose of this period was to gather baseline physical activity and diet behaviours. Participants who completed this baseline period were randomly assigned to one of four trial arms; control arm, low; medium and high reward arms (with the low, medium and high reward arms collectively referred to as "pooled reward arms").

For the next 20 weeks all participants could complete up to 2 weekly challenges from a list of 12 personalised diet and physical activity challenges and earn points for completing the challenges. Challenge content was designed to support participants to get closer to meeting the UK government dietary recommendations and the UK CMOs' physical activity guidelines. To encourage continuous improvement, participants could earn more points for completing challenges at a harder difficulty level.

If challenges were completed and validated, participants would earn points. All participants were awarded the same number of points following the successful completion of a challenge. However, depending on which trial arm the participant had been randomised to, they were awarded **different financial equivalents for those points as per Table i).**

Table i: An overview of the points available to participants across trial arms, and their financial value

Trial arm	1point = (£)	Challenge points available over 20 weeks (n)	Potential points value over 20 weeks (£)
Control	0	5230	0
Low reward	0.005	5230	£26.15
Medium reward	0.025	5230	£130.75
High reward	0.035	5230	£183.05

For the purposes of the evaluation, all participants were also asked to provide diet and physical activity data 1, 3 and 5 months after randomisation. At these timepoints, participants were asked to sync 4 days' worth of physical activity data from their wearable fitness tracker and retake at least one Intake24 survey. Participants were reimbursed financially across **all trial arms** for completing these data entries up to a maximum £70 over the duration of the pilot.

All participants had access to the same rewards store within the app where participants could redeem the financial value of their points and data collection reimbursements for real rewards at any point. This rewards store included thousands

of health-neutral and health-positive rewards, including gift-cards, merchandise options and gym-passes.

Primary outcomes are summarised in Table ii) below. These were:

- Moderate and vigorous physical activity (MVPA) in minutes per day: MVPA is calculated as the sum of vigorous and moderate activity
- Daily steps: measured as the number of steps taken each day
- Fruit and vegetables: measured in grams per day
- Fibre: measured in grams per day
- Free sugars: taken as the % of energy derived from food.
- Saturated fat: measured in grams and calculated as a % of energy derived from food

Secondary outcomes included analysis of daily energy expenditure and daily energy intake, a healthy eating score, self-reported weight, short- and medium-term impacts on primary outcomes (measured at one and three months respectively), and the impact of different reward levels, comparing each to the control arm. Exploratory outcomes included subgroup analyses, mental wellbeing and sleep duration.

Impact evaluation analysis

Analysis of the primary outcomes tested for differences between the pooled reward arms and the control arm five months after randomisation using an intention-to-treat approach and a linear mixed effects model. Analysis took into account individual characteristics of age, sex, ethnicity, education, BMI at baseline, variation in attrition observed across trial arms, and the brand of wearable fitness tracker.

Impact evaluation findings

Impact estimates compared the pooled reward arms and the control arm at five months post randomisation. These are presented in Table ii) below.

Table ii: Summary of primary outcomes and impact estimates at five months post randomisation

Primary Outcome	Sample size (participants)	Effect size (95% CI)
Physical Activity: Steps, total daily steps	3987	+42 steps (-170 - 254)
Physical Activity: MVPA, measured in minutes per day	3934	+1.1 minutes (-0.05 - 2.3)
Diet: Fruit and vegetables, grams per day	6149	+21 grams (14 - 28) **
Diet: Fibre, grams per day	6083	+0.35 grams (0.12 - 0.59) **
Diet: Saturated fat (as % of food calories)	6073	- 0.10 percentage points (-0.30 - 0.09)
Diet: Free sugars (as % of food calories)	6121	-0.21 percentage points (-0.55 - 0.12)

*Stars indicate the adjusted p-values: **:p<0.01.*

- Offering financial incentives did not impact participants' levels of physical activity MVPA. We found small increases that were not statistically significant for both MVPA (+1.1 min/day, 95% CI -0.05 - 2.3) and steps (+42 steps per day, 95% CI -170 - 254).
- Offering financial incentives had a statistically significant impact on fruit and vegetable consumption (+21 grams per day, 95% CI 14 - 28) and a statistically significant impact on fibre consumption (+0.35 grams per day, 95% CI 0.12 - 0.59) compared to the control arm. The remaining two outcome variables showed reductions that were not statistically significant.
- A sensitivity analysis was conducted for the physical activity data in which we replicated the approach to the primary analysis for physical activity outcomes, but included step and MVPA data from wearable fitness trackers worn less than 6 hours per day, which had been excluded from the original analysis plan. This revealed a statistically significant impact on both MVPA

(+1.9 min/day, 95% CI: 1.0 - 2.9) and steps (+256 steps per day, 95% CI: 71 - 442).

- Analysis of secondary outcomes found a positive relationship between incentive level and effect sizes for dietary outcomes. For fruit and vegetables, the high reward arm saw a statistically significant increase of 29 grams (95% CI: 18 - 39) compared to the control arm, and a statistically significant decrease in saturated fat of 0.4% (95% CI: -0.7% - -0.1%) compared to the control arm. When comparing the medium and low reward arm to the control arm, the effects found for fruit and vegetable intake were statistically significant, but smaller (+18g, 95% CI: 7 - 28) for the medium reward arm, and +16g (95% CI: 6 - 26) for the low reward arm.

- Subgroup analysis by participants' level of deprivation found a statistically significant effect on fruit and vegetable intake that was largest among participants who were less deprived compared to those who were more deprived (+25g per day, 95% CI: 15 - 35), while a statistically significant effect on physical activity was found for more deprived participants (+1.7 minutes of MVPA per day, 95% CI: 0 - 3.4).

- Subgroup analysis by participants' age found that those aged 41 and over achieved statistically significant increases in their MVPA (+1.5 minutes, 95% CI: 0 - 3), fibre (+0.4g, 95% CI: 0.1 - 0.7), and fruit and vegetable intake (+26g, 95% CI: 16 - 36), alongside a statistically significant decrease in saturated fat intake (-0.4g, 95% CI: -0.6, -0.1).

A target sample size of 4,200 participants was required at the final data collection point to be powered to detect effect sizes deemed substantive from previous literature. The final sample size for MVPA was 5% below the original target sample size, while for diet outcomes it was almost 50% higher.

Implementation and process evaluation

BIT conducted a mixed-methods implementation and process evaluation of the Financial Incentives Scheme to gain an in-depth understanding of how participants viewed and interacted with the intervention, the mechanisms through which the intervention worked (or the barriers if it did not), and to identify opportunities to further improve the intervention. This sought to understand why and how the

intervention achieved or failed to achieve expected outcomes based on the Theory of Change (ToC). IPE analysis incorporated both descriptive techniques, including summary statistics and data visualisation, and regression analyses.

Implementation and process evaluation findings

- The study met recruitment targets, with 18% of the adult Wolverhampton Local Authority population consenting to participate.
- The characteristics of the randomised cohort were different from those of the general population of Wolverhampton, and were more likely to be female (67% of the cohort) older (with a median age of 41), white (68% of the cohort), and income deprived (50% of the cohort).
- From the start of the pilot to the end of the pilot, attrition was higher amongst men, people aged under 41 and more deprived groups, meaning the month 5 sample set was technically older, less deprived and more likely to be female than at month 1.
- Participants who engaged with the app at least once every seven days were considered active users of the app. Across the duration of the pilot, the high reward arm saw the highest number of participants defined as active, suggesting that this group was the most engaged throughout the pilot.
- Insights from qualitative work found that financial incentives provided motivation for user engagement with the intervention, and encouraged completion of challenges.

Discussion

Overall, the pilot provided some evidence that financial incentives may be an effective way of bringing people closer to meeting guidelines for a healthy diet and physical activity, but with varied results. The greatest effects were found on fruit and vegetable intake, with small statistically significant increases found in participants' consumption of fruit and vegetables. A small statistically significant increase was also detected in fibre intake. The incentive structure also appeared to influence these outcomes, with those in the high reward arm demonstrating greater changes in their fruit and vegetable consumption (+29g per day, 95% CI: 18 - 39), and saturated fat consumption (-0.4g per day, 95% CI: -0.7 - -0.1).

In comparison, the small reductions in physical activity outcomes were not statistically significant, and the effect of the incentive value was less clear. Sensitivity analysis of primary physical activity outcomes generated larger, statistically significant effects for both MVPA and steps.

Those who were closer to guidelines for diet and physical activity at the point of randomisation were less likely to drop out of the pilot (regardless of the trial arm they were assigned to) and more likely to use the app and earn points. This may suggest potential self-selection effects, with those already motivated to maintain good health more likely to join the pilot. It may also suggest that participants were motivated by the app/wearable fitness tracker alone, which they had access to before randomisation; some participants' interview responses support this hypothesis, though it cannot be verified through the pilot given the active-control design.

The pattern of engagement also suggests that both demographic and social factors influenced participation; participants in households with multiple participants earned more points than those who were the only participant in their household, and younger participants were more likely to stop using the app than older participants. Further, the incentive structure appeared to play a role in participation, with the high reward arm having the largest proportion of participants defined as active compared with the other trial arms.

One of the aims of the pilot was to reach typically underserved groups, including minority ethnic groups, older adults, and those classified as most deprived. The pilot broadly met its targets for the composition of the sample for those still actively using the app by M5, with the exception of gender, where women were overrepresented. Trial participants were also more likely to be of white ethnicity, slightly less likely to be among the most deprived, and more likely to be younger than 45 years old compared to pilot targets. Offering rewards had a bigger impact on the outcomes of some demographic groups, such as more deprived individuals, previously physically inactive users, and older adults.

Participants who were classed as inactive at the start of the pilot also significantly increased their levels of physical activity. These findings suggest that demographic factors may play an important role in how effective financial incentives are in practice.

Overall, although these differences are small, they do suggest that financial incentives can move people towards healthier dietary behaviours. In addition, small changes in physical activity and diet sustained over a period of time or over a population can add up to larger impacts.

Limitations

The specific location and time period in which the pilot took place may limit the generalisability of the findings. The pilot's findings apply to the specific user population analysed, which was disproportionately female and older, potentially impacting the results' applicability to other populations.

Data for physical activity was collected from a range of compatible wearable fitness trackers that synced with the Better Health: Rewards app. Although the analysis attempts to control for variation between wearable fitness tracker reads, differences in the functionality of the trackers may have played a wider role in how participants used them, thus impacting wear time and data collection.

Dietary outcomes were collected through a recall questionnaire, Intake24, which introduces concerns about recall bias, honesty, and the tool's ability to comprehensively capture all food consumed.

Recommendations

Further research recommendations:

- Intervention design: Pilot findings suggest that demographic characteristics were a factor in both how effective the intervention was and how much people engaged with it, suggesting there is value in understanding and addressing barriers specific to these groups that may prevent them reaching their goals.
- Pilot design: This pilot was not able to identify the optimal financial incentive level for behaviour change in diet or physical activity outcomes. In a further iteration of the work, it may be valuable to consider increasing the threshold for the low reward arm, and removing caps on the target behaviour for steps, which may act as a disincentive to continued improvement.
- Pilot design: Further research could consider measuring the impact of the wearable fitness tracker and the app, to isolate the impact they have on

behaviour change individually. The current pilot was not designed to determine whether these elements were additive, and understanding their effect may make a valuable contribution to understanding what incentive level is most feasible and desirable.

Background

1.1. Health and behaviour change

Our behaviour is critical to our health. Behavioural risk factors represent our largest opportunity to reduce health burden across the population, making up **more than 50%** of the preventable Disability Adjusted Life Years (DALYs) lost, as estimated by the Global Burden of Disease study (The Institute for Health Metrics and Evaluation 2019). Furthermore, behavioural risk factors have a steep social gradient (i.e. they disproportionately impact those with lower incomes), and are therefore a key contributor to health inequities around the world. Women in the least deprived areas in England live a further 19.7 years in good health than those in the most deprived. For men, it's 18.4 (Merilynn 2021). A similar pattern of disparity is seen by ethnicity, where people from some groups (especially Pakistani, Bangladeshi, White Gypsy, and Irish Traveller groups) are more likely than White British people to report having a long-term condition and poor health. At older ages, health-related quality of life scores are lower and levels of diagnosed ill health are higher among these groups, based on responses to the GP Patient Survey (Raleigh and Holmes 2021).

In particular, **unhealthy diets and physical inactivity are associated with a wide range of chronic conditions**, including cardiovascular disease, metabolic disorders such as type 2 diabetes, and some forms of cancer (Scarborough et al. 2011). Furthermore, the link between these behaviours and our mental health is increasingly evident; a recent systematic review reported that physical activity, for example, is highly beneficial for improving symptoms of depression, anxiety and distress across a wide range of adult populations, including the general population, people with diagnosed mental health disorders and people with chronic disease (Singh et al. 2023).

While most are aware of the benefits of eating healthily (Grajek et al. 2022) and taking part in physical activity (UK Active 2023), **37%** of adults do not reach the recommended 150 minutes of physical activity per week (Sport England 2023), **70%** of adults consume fewer than the recommended minimum of five portions of fruit and vegetables per day (known as 5 A Day), and on average adults consume **more saturated fat and sugar than recommended**. Specifically, the latest National Diet and Nutrition Survey found that only 17% of UK adults aged 19-64 met the

recommendation of no more than 5% of daily total energy intake from free sugars (Public Health England 2020). Similarly, average intakes of saturated fatty acids (saturated fat) exceeded the government recommendation of this constituting no more than 10% of total energy in all age groups to whom the recommendation applied (Public Health England 2020).

1.2 The evidence for incentives

Incentives can be an effective tool for behaviour change because they change short-term motivation. Most behaviour change, particularly health-related behaviour change, requires consistent action - often with a short-term cost or inconvenience - while providing only long-term payoffs. For example, choosing to quit smoking requires consistent, difficult choices in the *short-term* to reduce the risk of developing lung cancer in the *long-term*. Without an immediate reward, people are less motivated to initiate or maintain these behaviour changes. Incentives address this gap by acting as a shorter-term reward that motivates people to engage in the desired behaviours (Kullgren et al. 2016).

The Eatwell Guide and the CMOs' Guidelines for Physical Activity describe the UK Government recommendations for healthy diet and physical activity that the general population should seek to achieve. Previous work has examined how incentives could be used to encourage these recommended behaviours, with some success. A systematic review and meta-analysis of 23 studies from 2014-2018 found that financial incentives increased daily step counts in adults for short and long duration interventions by 607 steps (approximately 10%–15%). Statistically significant daily step count differences were also observed 3–6 months after incentives were removed, with an average difference of 514 steps per day post-intervention. 14 of these 23 studies used wearable fitness trackers (Mitchell et al. 2019). This study built on a 2013 meta-analysis, which found that financial incentives increased exercise session attendance by 11.6% (Mitchell et al. 2012), and subsequent reviews conducted between 2014-2017, which found that financial incentives encourage physical activity (Barte and Wandel-Vos 2015; Giles et al. 2014; Haff et al. 2015; Strohacker, Galarraga, and Williams 2013).

Similarly, a mixed-methods evaluation of a financial incentives scheme with participants in the USA successfully increased fruit and vegetable intake, leading to improvements in participants' general health, with their experience of chronic

conditions to be substantially improved. Participants reported practising more healthy behaviours, such as meal planning, healthier cooking methods, and increasing their levels of physical activity (Waugh. 2021).

There is also evidence supporting the delivery of financial incentives schemes via mobile phone apps. A quasi-experimental evaluation of the 'Carrot Rewards' app found a modest increase in step count; participants walked 116 more steps daily at study week 12 than at baseline (Mitchell et al. 2018).

1.3 The project

To support adults in England to translate these recommendations into practice, **HeadUp Labs (HUL)** was commissioned by the **Department of Health and Social Care (DHSC)** to develop and implement a pilot **app-based Financial Incentives Scheme** to provide rewards to people, contingent on them improving the quality of their diet and level of physical activity in line with the Eatwell Guide and CMO's guidelines.

The app encouraged participants to select weekly diet and physical activity challenges. Participants successfully meeting the challenges earned 'points' at the end of the week. The app allowed some users (those allocated to the pooled reward arms) to redeem these points against rewards. The reward offer was broad, and included many categories of rewards and a broad range of price points, in order to be as appealing and relevant as possible to a range of people / preferences. All rewards were health-neutral or health-promoting, but none were health-negative (i.e. gambling, alcohol, or smoking-related rewards were not included). Rewards included gift cards for supermarkets, vouchers for major retailers (including clothing, homeware, high-street stores, and Wolverhampton FC), 'days out', and other entertainment options (such as cinema or theatre tickets). DHSC also partnered with national supermarkets who offered bespoke gift cards, and local gyms who offered gym passes.

1.4 Our role

The **Behavioural Insights Team (BIT) acted as an independent evaluator** of this pilot scheme to understand whether, and to what extent, financial incentives can encourage behaviour change. BIT also conducted an implementation and process

evaluation of the scheme to gain an in-depth understanding of how participants viewed and interacted with the intervention, the mechanisms through which the intervention worked (or the barriers if it did not), and to identify opportunities to further improve the intervention. This was done using both qualitative (interviews and focus groups) and quantitative methods (analysis of app-based metrics). The next section of the report outlines the evaluation objectives and timeline.

This report shares the findings of the evaluation, covering:

- The intervention itself: description, Theory of Change, user journey;
- The impact evaluation methodology: objectives, study design, user flow and randomisation, outcomes, sample size;
- The implementation and process evaluation (IPE) methodology: objectives, study design, quantitative methods and analysis, qualitative methods and analysis;
- The findings of both the impact evaluation and the IPE, and their limitations.

The final section, the Discussion, sets out recommendations for DHSC policy makers based on the findings from this study.

1.5 Evaluation objectives

The Financial Incentives Scheme was designed to test whether offering financial incentives to adult residents of Wolverhampton could increase their levels of physical activity (PA) in line with CMOs' guidelines, and improve their diet towards UK government dietary guidelines, as depicted in the Eatwell Guide.

The secondary research questions explore the impact over time, the optimal incentive values, and broader effects on PA and diet. The exploratory research questions investigate the impact across different demographics and outcomes where we wouldn't expect to see an immediate impact: sleep quality and mental health.

The purpose of the implementation and process evaluation was to understand how intervention participants, churn-out participants, non-engagers and delivery partners viewed and/or interacted with the intervention. We explored barriers and facilitators

to initial and sustained engagement with the intervention, and mechanisms of impact by which the intervention did or did not change healthful behaviour.

The Trial Steering Committee (TSC) consisted of the following independent members; Professor Amanda Daley (chair) (Loughborough University), Professor Louise Goff (University of Leicester), Professor Deborah Ashby (Imperial College London). Non-independent members were Dr Giulia Tagliaferri (Behavioural Insights Team) and Marylise Talbot (Head Up Lab). The TSC provided oversight for the evaluation design and delivery of the trial.

The study obtained ethical approval from UKHSA on 29th June 2022. The trial protocol (TP) was registered on the ISRCTN platform (registration number ISRCTN10465935) prior to launch on 10th February 2023 and can be found at the registration link (latest version dated 11th Aug 2023): https://www.isrctn.com/ISRCTN10465935.

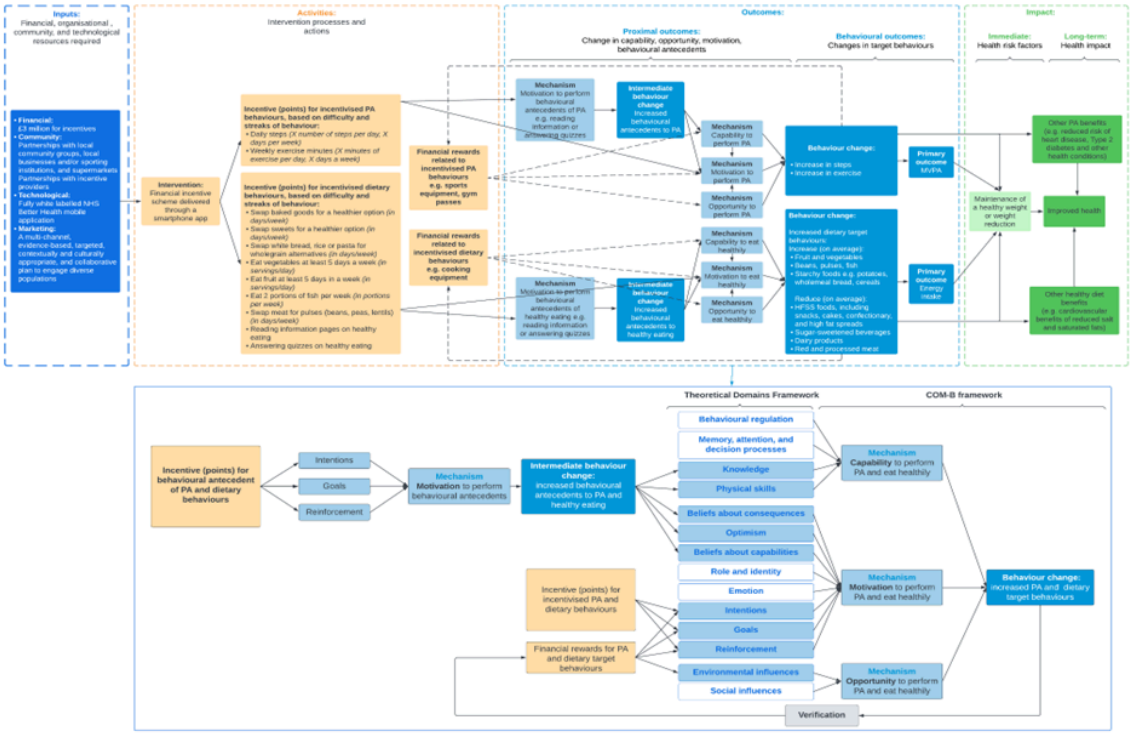
2. Intervention

This was a randomised controlled trial with one control arm and three reward arms. While all participants received access to the free Better Health: Rewards app and a free wearable fitness tracker (if they did not already own one), only those allocated to one of the three reward arms were eligible to earn financial incentives for improvements in their diet and level of physical activity. The three reward arms offered different levels of incentives: high, medium and low.

2.1 The Theory of Change

Figure 1 outlines **how a financial incentives scheme** was hypothesised to increase adherence to healthy eating guidelines and promote greater levels of physical activity as well as how these behaviour changes relate to expected changes in population health.

Figure 1: Overarching Theory of Change



Based on this theory of change, a financial incentives scheme was designed to improve participant's physical activity levels and diet healthfulness, featuring both incentives and reward components. The scheme included:

- Physical activity and healthy eating challenges, in which informative content, tips or recipe ideas would be surfaced, hypothesised to increase the participant's **capability** for change.
- Incentives (in the form of points) contingent on participants demonstrating the desired behaviours, hypothesised to increase the participants' **motivation** to change.
- A mechanism for participants to receive rewards with a real financial value, hypothesised to increase participants' **motivation** to change.
- Health-promoting rewards, hypothesised to further increase participants' **capability, opportunity, and motivation** to perform more physical activity and eat more healthily.

A detailed overview of the final Financial Incentives Scheme, including how it was deployed to investigate the impact of the **financial incentives** specifically, can be found in the following section.

2.2 The Financial Incentives Scheme

The Financial Incentives Scheme was delivered to participants digitally via a specially created free **'Better Health: Rewards'** smartphone app, compatible with both iOS and Android devices.

The scheme was trialled in The City of Wolverhampton Council between Feb 17th 2023 and 13th October 2023. All adults, aged 18 and over, residing within a Wolverhampton postcode, were eligible to take part. The intervention exposure for each participant lasted 5 months.

Site selection

The City of Wolverhampton Council was selected to host the pilot by DHSC after an expression of interest exercise. The Council had been trialling innovative solutions to improve the health and wellbeing of their population, of which one third is classed as being physically inactive and a below-average number of adults in the city are eating their 5 A Day. Additionally, with its large population size and high levels of diversity, trialling the scheme in Wolverhampton was hoped to generate a sufficiently large, representative sample set that would allow the evaluators to

robustly investigate how effective the scheme was for different population sub-groups.

How did the Financial Incentives Scheme work?

- The 'Better Health: Rewards' app was widely advertised in Wolverhampton between February 17th 2023 and March 31st 2023 via a geo-targeted omnichannel marketing campaign featuring digital, social, partnership and out-of-home activities.
- Residents voluntarily self-selected to take part in the scheme by downloading the free app to their phone via their app stores.
- After downloading the Better Health: Rewards app, the resident would be asked to consent to take part in the independent evaluation. To progress in the scheme they had to comply with the following eligibility criteria; 1) enter their address to validate that they lived in Wolverhampton; 2) set up their account and enter basic health and demographic information, 3) self-confirm they met the clinical criteria to take part including reviewing safety disclaimers. See Appendix A for further information. Once participants completed these steps, the resident was considered a **registered participant.**
- The registered participant then had to complete a **mandatory baseline period. The purpose of this period was to gather insight on participant's usual physical activity and dietary behaviours**. To complete the baseline period participants had to:
    - Connect their existing wearable fitness tracker to the app or order a free Better Health: Rewards tracker via the app if needed.
    - Sync at least 4 days' worth of physical activity data, with a minimum of 6 hours wear time per day, to the app.
    - Complete a series of dietary surveys, including a food-frequency-questionnaire and minimum of 1 (maximum of 2) dietary recall surveys via Intake24, at least one day apart.
    - Optionally complete a series of surveys, including: mental health survey, motivation level survey, educational background survey and marketing survey.

- After this point, the participant was **randomised into a trial arm** and could start completing challenges and earning points. If participants did not complete the above requirements, they were not randomised and did not progress any further.
- On a weekly basis, all participants were generated a list of 12 personalised diet and physical activity challenges which, if completed and once evidenced, would earn them points. Participants could choose up to two challenges to work on each week for 20 weeks.
- For the purposes of the evaluation, all participants were also asked to provide diet and physical activity data at 1, 3 and 5 months after randomisation. At these timepoints, participants were asked to sync 4 days' worth of activity data and retake the Intake24 surveys. Participants were reimbursed for completing these data entries.
- Participants could cash in the financial value of their points and data collection reimbursements for real rewards at any point in the intervention via the in-app store. This rewards store included thousands of health-neutral and health-positive rewards, including gift-cards, merchandise options and gym-passes.

How did the scheme isolate the impact of financial incentives?

Theory of Change - rewards pathway: the delivery of the prize with financial value contingent on the behaviour being performed.

The effectiveness of the Financial Incentives Scheme was evaluated via a blinded randomised controlled trial. At randomisation, participants were allocated at a household level to one of four trial arms: high reward, medium reward, low reward, or control.

All participants were awarded the same number of points contingent on their successful completion of a challenge. However, depending on which trial arm the participant had been randomised to, they were awarded **different financial equivalents for those points**.

Table 1: An overview of the points available to participants across trial arms, and their financial value

Trial arm	1point = (£)	Challenge points available over 20 weeks (n)	Potential points value over 20 weeks * (£)
Control	0	5230	0
Low reward	0.005	5230	£26.15
Medium reward	0.025	5230	£130.75
High reward	0.035	5230	£183.05

By controlling for all other intervention features (section 2.2) across trial arms, including the ability to order a free wearable fitness tracker, the app experience and the ability to earn points contingent on challenge completion, we effectively isolated the impact of increasing financial incentives only.

2.3 How did the challenges work?

Challenges were the mechanism by which the app enabled participants to improve their health behaviours closer to the UK government dietary recommendations and the UK Chief Medical Officers' physical activity guidelines. The table below lists the 12 challenges available to participants.

Each week, participants selected up to two challenges to work on and committed to achieving a weekly target for those challenges. The challenge targets were initially based on the participant's baseline level of physical activity and food-frequency-questionnaire answers, and, if successfully achieved, would incrementally increase in difficulty on a weekly basis until the participant achieved the challenge threshold. These challenge thresholds are based on the UK government dietary recommendations and the UK Chief Medical Officers' physical activity guidelines and can be seen below in Table 2.

Table 2: An overview of the 12 in app healthy eating and physical activity improvement challenges

Challenge title	Description	Challenge threshold
Increase your fruit and vegetables!	This goal is to increase the number of fruits and vegetables you have each week	5 servings a day for 7 days in a week
Switch up your drinks	This goal is to make healthier drinks swaps each week.	7 sugary drink free days a week
Brilliant Breakfast	This goal is to make healthier breakfast swaps each week.	Swap a sugary breakfast for a healthier option 7 days a week.
Snack Attack	This goal is to make healthier snack swaps each week.	Swap a sugar snack for a healthier option 7 days a week
Eat leaner meat	This goal is to swap to leaner meat options each week.	Swap out your red or processed meat option for 7 days a week.
Fibre fix	This goal is to swap your starchy carbohydrates for higher fibre choices each week.	Swap to a wholegrain alternative 7 days a week.
Dessert swapper	This goal is to make healthier dessert swaps each week.	Swapping dessert for a healthier option 7 days a week.
Step it up!	This goal is to increase your steps by an average of 500 extra steps each day for a week.	50 and below: 8000 steps/day 51 years and over: 5000 steps/day
Let's get moving	This goal is to do 10 extra minutes of activity each week.	150 minutes exercise/week
Choose lower fat	This goal is to swap to lower fat choices each week.	Swap a full fat dairy product for a low-fat option for 7 days in a week.
Healthier habits on the go	This goal is to make healthier swaps when you're buying food on the go each week.	No convenience foods for 7 days in a week.
Say no to fried food	This goal is to swap fried food for healthier choices each week.	7 fried food free days in a week

During the challenge, the participant would be shown relevant informative content, tips or recipe ideas, which, by enhancing knowledge, was hypothesised to increase their motivation and capability for developing long-term healthier habits.

2.4 What was the role of points?

Theory of change – incentive pathway: the promise of a reward contingent on the performance of a behaviour.

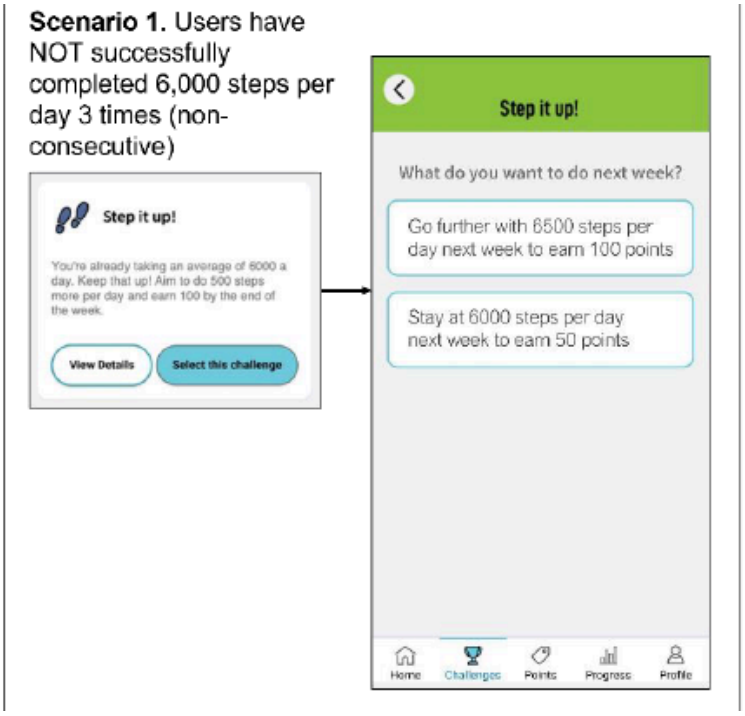
All participants were offered, and could earn, the same number of points by successfully completing their weekly challenges, with more points available to all

participants if they committed to the more challenging behaviour change target that week:

- In their first challenge week, participants were offered the option to select challenges at the harder difficulty only, in exchange for 100 points.
- In all subsequent weeks all participants had the choice to:
    - "**go harder**" on a challenge for 100 points; or
    - **"maintain"** a challenge behaviour for 50 points.
- Participants who selected 'go harder' but only partially achieved the challenge target still received 50 points.
- Participants who chose to maintain but only partially achieved the challenge received 0 points.
- To encourage habit formation, participants earned a marginally higher number of points (+2.75%) when they successfully completed the same challenge, at incremental difficulty, in consecutive weeks.

To encourage participants to continue improving their health behaviours rather than maintaining, participants could earn 50 points for maintaining their behaviour at a certain point, including the challenge threshold, **for up to three weeks only**. After this point, maintaining would be rewarded with 0 points.

Figure 2: Challenge difficulty screen, demonstrated for the 'step it up' challenge



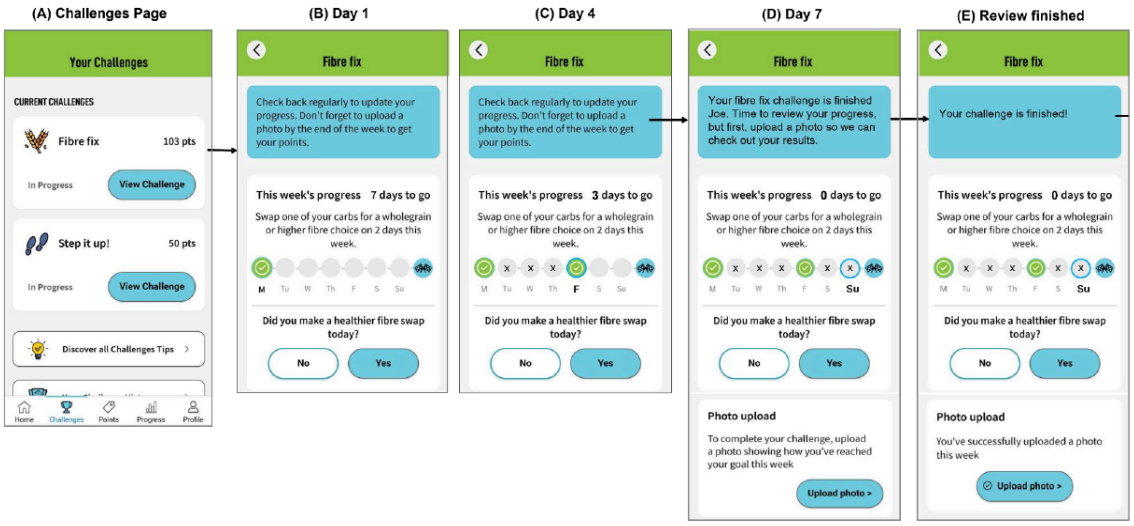
Notes: The participant in the figure is currently walking 6000 steps a day, if they'd like to 'go harder' and challenge themselves to 6500 steps a day, they will earn 100 points. If they'd like to maintain at 6000 steps a day, they can earn 50 points.

How did participants prove they had achieved the challenge targets?

Participants had to complete and validate their challenges to earn points. To do this, participants had to upload data to their challenge page:

- **Physical activity challenges:** Sync wearable fitness tracker data that evidences they had walked the targeted number of steps or had completed the target minutes of MVPA exercise.
- **Healthy eating challenges:** Manually record what swaps had been completed each day and further validate by uploading grocery receipts, photo evidence of swaps or completing a relevant quiz.

Figure 3: Illustration of the challenge experience, demonstrated for the 'fibre fix' challenge



The corresponding points were awarded to the participant if the uploaded data proved that the challenge target had been met or exceeded. Points were awarded to participants as quickly as possible (typically within 24 hours) after they had proved they had successfully performed the desired behaviour.

Data collection reimbursement:

In addition to the challenges, the app asked all participants to provide diet and physical activity data at 1, 3 and 5 months after randomisation, which fed into the impact evaluation.

At these timepoints, participants were asked via the app, and reminded via app and email notifications, to sync 4 days' worth of activity data and retake the Intake 24 surveys. Up to £70 reimbursement, credited to their rewards store, was available to participants for completing all the required data entries. An outline of the reimbursements available can be seen in Table 3 below.

Table 3: A summary of the data collection reimbursements available to all participants

Time point	Data collection request	Reimbursement (£)
Baseline	Complete 1st intake 24 Complete 2nd intake 24	5 5
1-month post-randomisation	Complete 1st intake 24 Complete 2nd intake 24	5 5
3 months post-randomisation	Complete 1st intake 24 Complete 2nd intake 24	10 10
5 months post-randomisation	Complete 1st intake 24 Complete 2nd intake 24	10 10
5 months post-randomisation	Sync wearable fitness tracker 2x	10

What rewards were available?

Participants could cash out the financial equivalent of their data collection reimbursement and their challenge points (if they were in the high reward, medium reward or low reward trial arm) at any point post-randomisation via the in-app rewards catalogue. A summary of the financial values available to participants is shown below in table 4.

The rewards catalogue for this scheme was specially curated to include thousands of health-promoting and health-neutral rewards, including supermarket vouchers, sports equipment, local gym passes, merchandise, high-street gift cards, leisure activities and cooking equipment. Bespoke supermarket vouchers were created to minimise participants' spend on health-negative behaviours, including drinking, smoking, and gambling. This catalogue was hypothesised to positively reinforce participant's capability, opportunity, and motivation to perform physical activity and eat healthily.

Table 4: A summary of the points and reimbursements available to participants across trial arms

Trial arm	'Maintain' challenge points (n)	'Go harder' challenge points (n)	Max challenge points available over intervention (n)	Cash equivalent of points (£)	Max cash available for challenges* (£)	Max data reimbursement (£)	Max. total payout (£)
Control	50	100	5230	0	0	70	70
Low	50	100	5230	0.005	£26.15	70	96.15
Medium	50	100	5230	0.025	£130.75	70	200.75
High	50	100	5230	0.035	£183.05	70	253.05

Notes: the max challenge points available over intervention is computed assuming a participant chooses, and successfully completes, two challenges per week for 20 weeks, at the hardest intensity available to them and makes full use of the 2.75% compound streak. The cash equivalent of points were informed by literature and available budget.

3. Impact evaluation methodology

3.1 Objectives

3.1.1 Primary research questions

The primary objective of the impact evaluation study was to assess the effectiveness of the Financial Incentive Scheme to answer the following primary research questions:

- Physical activity outcomes:
    a. Did the Financial Incentives Scheme increase moderate-vigorous physical activity **(**MVPA minutes/day) for those who received it, compared to the control arm?
    b. Did the Financial Incentives Scheme increase moderate-vigorous physical activity **(**steps/day) for those who received it, compared to the control arm?
- Diet outcomes:
    a. Did the Financial Incentives Scheme improve the healthfulness of the diet for fruit and vegetable intake in g/d, for those who received it, compared to the control arm?
    b. Did the Financial Incentives Scheme improve the healthfulness of the diet for fibre intake in g/d for those who received it, compared to the control arm?
    c. Did the Financial Incentives Scheme improve the healthfulness of the diet free sugars intake in % of food energy/day for those who received it, compared to the control arm?
    d. Did the Financial Incentives Scheme improve the healthfulness of the diet saturated fat intake in % of food energy/day for those who received it, compared to the control arm?

In each case, the outcomes for the pooled reward arms were compared to those of participants allocated to a control arm (no financial incentives offered) at month 5 of the pilot.

3.1.2 Secondary research questions

The secondary research questions for the pilot were:

- **Broader effects on PA and diet:** Comparing the pooled reward arms with the control arm five months after randomisation, what is the impact of offering financial incentives on participants':
    - Energy expenditure
    - Energy intake
    - Score on a healthy eating score based on consumption of key food groups, macro- and micronutrients
    - Weight
- **Shorter-term effects:** What is the impact of offering financial incentives on participants' dietary intake and PA, in the pooled reward arms compared to the control arm at one and three months after randomisation?
- **Optimal incentive value:** What is the impact of each of the incentive levels on each of the primary outcomes specified for PA and diet, comparing the control arm to the pooled reward arms five months after randomisation?

3.1.3 Exploratory research questions

The exploratory research questions for the pilot were designed to allow us to understand whether the Financial Incentives Scheme affected participants' motivation to change their diet and PA behaviours, as well as whether it had any potential unintended consequences. The exploratory measures were:

- **Subgroup analyses:** What is the impact of the pooled reward arms on the primary outcome measures for population subgroups, five months after randomisation, comparing the behaviour of participants allocated to the control arm?

- **Longer-term effects on motivation to change:** What is the impact of the pooled reward arms on participants' motivation to improve dietary intake and to increase their PA level?
- **Unintended consequences:** What is the impact of the intervention on participants' sleep and mental health?

3.2 Outcome measures

The primary outcomes consisted of **2 physical activity outcomes** and **4 dietary outcomes** (see Table 5). Data collected at three moments, 1, 3 and 5 months post randomisation (M1, M3, M5), were used for the impact evaluation.

A summary of the outcome measures and their measurement points are set out below in Table 5.

Table 5: Outcome measures

Outcomes	Category	Metric	Collection method	Research Question (RQ)	Measurement point used
Primary	PA	• MVPA (min/day) • Daily steps	Wearable fitness tracker	Primary RQ: Effect of offering financial incentive on PA	M5
				Secondary RQ: Short-term and medium effects on PA	M1, M3
				Secondary RQ: Impact of different incentive levels	M5
				Exploratory RQ:	M5

				Subgroup analysis	
				Sensitivity analysis to the primary RQ	M5
	Diet	• Fruit and vegetables (g/day) • Fibre (g/day) • Free sugars (% daily food energy) • Saturated Fat (% daily food energy)	24h recall survey (Intake24)	Primary RQ: Effect of offering financial incentive on dietary behaviour	M5
				Secondary RQ: Short-term and medium effects on diet	M1, M3
				Secondary RQ: Impact of different incentive levels	M5
				Exploratory RQ: Subgroup analysis	M5
				Sensitivity analysis to the Primary RQ	M5
Secondary	PA	• Energy expenditure (kcal/day)	Wearable fitness tracker	Secondary research question - Broader effect of offering financial incentive on holistic constructs	M5
	Diet	• Energy intake (kcal/day) • A healthy eating	24h recall survey (Intake24)	Secondary research question - Broader effect of offering financial incentive on holistic constructs	M5

		score based on consumption of key food groups, macro- and micronutrients			
	Weight	• Weight (kg)	Self-reported survey	Secondary research question - Broader effect of offering financial incentive on holistic constructs	M5
Exploratory		• Self-reported mental well-being measured by WHO-5 Index (0~100, 0 = worst, 100 =best)	In-App survey	Exploratory: Unintended impacts	M5
		• Sleep duration (hours/day)	Wearable fitness tracker	Exploratory: Unintended impacts	M5
		• Motivation to change PA • Motivation to change diet	Motivation to change diet	Exploratory: Motivation to change	M5

3.2.1 Primary outcome measure

The primary physical activity outcomes were:

(i) **Moderate and vigorous physical activity** (MVPA) in minutes per day: MVPA is calculated as the sum of vigorous and moderate activity. All wearable fitness trackers define a PA to be "vigorous" if the user's heart rate increases to 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 are detected by accelerometers built in the wearables. Better Health: Rewards tracker users also had to manually select an 'exercise mode' function in the wearable fitness tracker to record MVPA.

(ii) **Daily steps** measured objectively through the users' wearable fitness tracker. This is measured as the number of steps taken each day. In the main data outcome collection weeks, participants were nudged to wear their wearable fitness tracker as long as possible for 7 days and to sync the device with the app.

The primary dietary outcomes were measured through up to two Intake24 (Foster et al. 2019) questionnaires per user. Intake24 is an open-source web-based dietary assessment research tool based on the 24-hour recall method, primarily designed for self-completion. It was created by Newcastle University (UK), funded by Food Standards Agency, Scotland and is now maintained and developed through a collaboration between Cambridge University (UK), Monash University (Australia) and Newcastle University (UK). Intake24 was provided by the MRC Epidemiology Unit at Cambridge University for this study. Participants submitted all food and drinks they consumed on a single day, from which the food groups and nutrients were calculated. Measures included:

(i) **Fruit and vegetables:** measured in grams per day.

(ii) **Fibre:** measured in grams per day.

(iii) **Free sugars:** this was taken as the % of energy derived from food.

(iv) **Saturated fat:** measured in grams and calculated as a % of energy derived from food.

3.2.2 Secondary outcome measure

The secondary outcomes included:

(i) **Daily energy expenditure**: defined as daily calories (kcal/day) burnt from physical activity.

(ii) **Daily energy intake** measured in (kcal/day) and collected via 24-hour dietary recalls (via the Intake24 surveys)

(iii) **A healthy eating score:** collected using 24-hour dietary recalls, based on self-reported consumption of key food groups, macro- and micronutrients. This method is based on the guidelines provided by Scheelbeek et al. 2020, assessing participants' diets against specific nutritional criteria. The scoring system allows for a range from 0 (least healthy) to 7 (most healthy), with points awarded as follows:

1. Fruit and Vegetables: ≥ 400g/day.
2. Red and Processed Meat: ≤ 70g/day.
3. Free Sugars: ≤ 30g/day.
4. Saturated Fat: ≤ 30g/day for males, ≤ 20g/day for females.
5. Fibre: ≥ 30g/day.
6. Total Fat: For males, ≤ 97g (18-64), ≤ 91g (65-74), ≤ 89g (75+); for females, ≤ 78g (18-64), ≤ 74g (65-74), ≤ 72g (75+).
7. Salt: ≤ 6g/day.

Each metric met by a participant scored 1 point, with a maximum of 7. Per the evaluation plan, we excluded two metrics used by the Scheelbeek et al. methodology: oily fish and other fish consumption. We did this as these are expressed on a weekly basis, and we only worked with 24 hours dietary recalls.

(iv) **Self-reported weight:** participants were asked to self-report their current weight at baseline and each measurement time point. The app encouraged participants to weigh themselves when entering weight data.

3.2.3 Exploratory outcomes

The exploratory outcomes included assessment for potential unintended health impact of the interventions. These included:

(i) **Subgroup analyses:** measuring the impact of the pooled intervention on each primary outcome measure at five months by:

- Deprivation level
- Ethnic group
- Age
- Sex
- Baseline diet
- Baseline PA

(ii) **Motivation to change diet:** We collected data on two exploratory outcomes to capture participants' motivation to change their diet before and after the intervention. We used a 5-point Likert scale (1=not at all confident/willing to 5=very confident/willing with 3 being neutral) for each question:

- 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?

(iii) **Unintended consequences:**

- **Sleep duration:** measured in (hours / day) by the wearable, for those willing to wear it at night.
- **Mental well-being:** mental wellbeing was collected at baseline and three data collection points (M1, M3, M5) by prompting participants via push notifications and in-app prompts to assess their mental health using the WHO-5 Well-being questionnaire. This is a simple and validated index consisting of five statements, which respondents rate in

relation to how they have felt over the past two weeks. Details of this measure can be found in Appendix B.

3.3 Sample & Randomisation

Users were randomly assigned to one of the four trial arms. The trial was a randomised controlled trial, clustered at the household level. This minimised the risk of users in the same household influencing each other's behaviour, by ensuring that all individuals in the same household were allocated to the same trial arm.

BIT and DHSC agreed on target effect sizes for this pilot based on the available evidence for similar trials (see section 2.4.2.1 of the TP). A target effect size of 0.2 standard deviations (SD) for the PA outcome and 0.11 SD units for diet outcomes between intervention (any reward) and control (no reward) arms was deemed achievable.

The target sample size for the pilot was calculated based on these target effect sizes. To determine the target sample size, BIT conducted power calculations for the primary research question "*Does the Financial Incentives Scheme improve physical activity (PA) and diet at the five month mark?*". The power calculations focused on the primary outcomes and on the primary analysis, which pools all the reward arms together and assesses their overall difference against the control.

Table 6 summarises the intended minimum detectable effects (MDES) for each outcome, against the number of participants required to achieve this at each stage of the pilot. These calculations were conducted to ensure we were powered to detect the smallest effect across the six primary outcomes. We estimated that at month five, a sample size of around 4,200 participants would be required.

To achieve this, based on learnings from HUL's experience on similar apps and evidence from the literature about the risk of differential attrition across trial arms, it was estimated that approximately 15,500 participants were required at the point of randomisation. This corresponded to 25,800 participants at the beginning of the baseline phase, to account for an expectation that a substantial proportion of those entering this phase would not proceed to randomisation.

We estimated that 1,900 users were required in the high incentive group, 1,950 in the medium incentive group, 2,250 in the low incentive group, and 9,400 in the control group to achieve a sample size of 15,500. To achieve this balance, an algorithm was created to assign 12% of the households to the high incentive group, 12% of the household to the medium incentive group, 15% of the household to the low incentive group and 61% of the households to the control group.

Table 6: Summary of power calculations and the required sample size at each stage of the trial

	Physical activity		**Diet**			
Primary outcome measure	*MVPA*	*Steps*	*Fruit and vegetable*	*Fibre*	*Free sugars*	*Saturated Fat*
Minimum detectable effect size in SD (MDES)	0.2 SD units		0.11 SD unit			
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 %
The minimum effect sizes we are powered (at 80% level) to detect at the 5-month mark in the primary analysis	2.4 min/day	371 steps/ day	18.7 g/day	0.74 g/day	0.63%	0.31%
Number of **overall participants** required to	**4,200** (700 per reward arm; 2,100 in control arm)					

achieve the smallest target MDES at **the 5m mark analysis stage**	
Number of **overall participants** required to achieve the smallest target MDES at **randomisation**	**15,500** (1,900 in the high reward arm, 1,950 in the medium reward arm, 2,250 in the low reward arm, 9,400 in the control arm),
Number of **overall participants** required to achieve the smallest target MDES **to enter the baseline phase**	**25,800** (3,100 in the high reward arm; 3,200 in the medium reward arm; 3,800 in the low reward arm; 15,700 in the control arm)

3.4 Analysis plan

The entire analysis for this trial was based on an **intention to treat (ITT).** This means that we set out to assess the impact of being offered any financial incentive (compared to being offered no financial incentives). Technically, it means that we analysed outcome data for all participants randomised for which valid outcome data were collected and who did not withdraw from the pilot, regardless of their level of participation with the study (e.g. regardless of whether they selected or completed any challenges, earned any points or they redeemed any rewards). As pre-specified in the TP, no outcome was imputed using baseline measure where an outcome at M5 was missing.

In this trial, the unit of analysis was the individual's diet and PA outcomes repeatedly collected over multiple windows of evaluation, and the unit of randomisation was

the household. To account for the clustering of observations within individuals and individuals within households, we used 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. P-values were adjusted for multiple comparisons using the Benjamini Hochberg correction, as specified in the TP. In addition, regressions include key variables that are likely to influence the outcomes, such as whether the diet questionnaire was completed on a weekend or weekday, or the brand of the wearable fitness tracker (see the TP for a full list of the covariates included in the regression models).

In our analysis, we employ inverse probability weighting (IPW), as detailed in Appendix E. Consequently, when discussing control arm averages for outcomes at M1, M3, and M5, we refer to weighted averages. Specifically, each observation is assigned a weight based on the inverse of the probability of having a valid response for the relevant outcome. In discussions of pooled reward arm outcomes, these reflect the estimated weighted average for the control arm under the hypothetical scenario where they would have received the treatment (the reward).

Table 5 in section 3.2 summarises the pre-specified analyses detailed in the TP (see column "Research Question").

The TP describes in great detail the data cleaning rules used to ensure data quality of PA and diet outcomes. Following an initial analysis in line with the TP, DHSC became aware the TP was based on a misinterpretation of the National Diet and Nutrition Survey approach to data cleaning dietary surveys. DHSC consulted with BIT and the Trial Steering Committee to seek views on an appropriate way to rectify the misinterpretation and reduce the risk of excluding valid data. As a consequence the analysis was rerun. This analysis took place after the evaluation team and DHSC were party to the initial results. More details are provided in Appendix K.

4. Impact evaluation findings

4.1 Participants

4.1.1 Sample size: reach

17% of the entire adult (18+) population of Wolverhampton Local Authority (34,900 adults) consented to take part in the study. The targets for recruitment were met, with the exception of a lower proportion of male and older sign-ups than expected.

34,900 adults in Wolverhampton consented to take part in the study - the very first step in the user journey of this pilot. It is worth noting that this represents 17% of the entire adult (aged 18+) population of Wolverhampton, or more than one in six adults living in the area (Office for National Statistics 2021a). The population of Wolverhampton LA aged 18+ in the 2021 census was 202,385, assuming that half the 16 - 19 years old age group were 18 or older.

Demographic data were captured for the 28,281 participants who completed registration (81% of those who consented to participate). Reasons for incomplete registration include not verifying contact details (email address, phone number, home address), not providing other personal data, or not confirming clinical consent.

- Of the 28,281 individuals who completed registration, 64% were female and 36% were male.
- Data on levels of deprivation (using the Index of Multiple Deprivation (IMD)) showed that 54% were classified as being within the 'most deprived' categories (scoring either 1 or 2 on the 1-10 IMD scale), and 46% were classified as being outside this category.
- With regard to ethnicity, 58% identified as White; 25% as Asian or Asian British; 8% as Black, African, Caribbean, or Black British; and 4% as coming from

mixed or multiple ethnic background(s). 41% of those who completed registration were aged 45 or over.

A comparison of demographic data for registered participants with the originally stated targets can be seen in Table 7 below. The targets were largely met, especially with reference to participants living in deprived areas and having a non-white ethnicity. However, a lower proportion than targeted for men, older residents, and those of white ethnicity completed registration.

Table 7: A comparison of demographic data for registered participants with originally stated targets

Ethnicity category	**Target (%)**	**Registered (%)**
White	68	58
Asian or Asian British	18	25
Black, African, Caribbean, or Black British	7	8
Mixed or Multiple Ethnic Backgrounds	5	4
Gender category	**Target (%)**	**Registered (%)**
Female	~50	64
Male	~50	36
Deprivation (IMD) category	**Target (%)**	**Registered (%)**
Most deprived (IMD ≤ 2)	51	54
Other (IMD > 2)	49	46
Age category	**Target (%)**	**Registered (%)**
Aged 45 or over	45	41

4.1.2 Sample size: from consent to randomisation

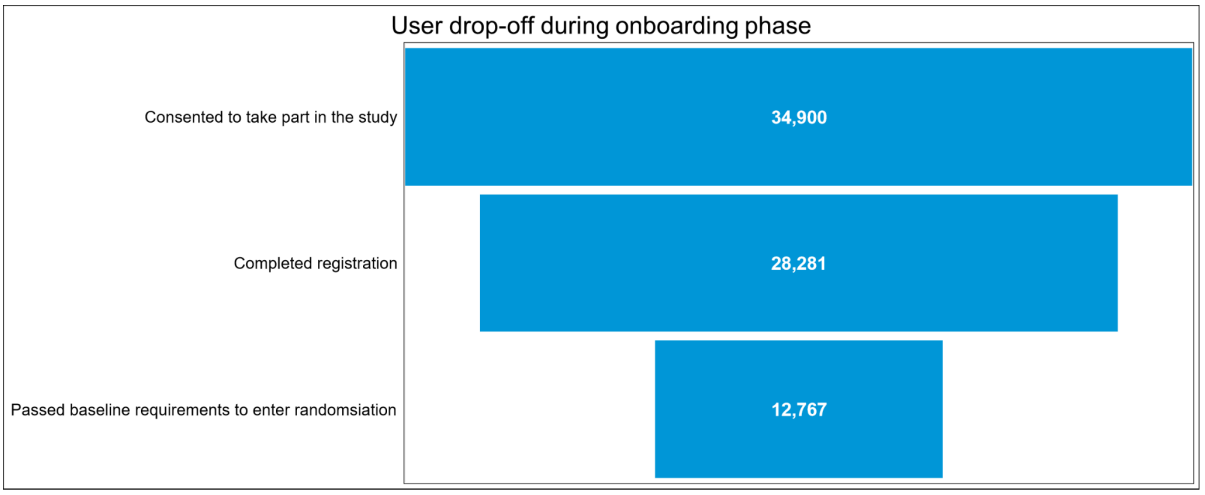
- 37% of participants who consented to the study proceeded to randomisation (N=12,767). At randomisation, 7,791 participants were randomised to the control arm, 1,900 participants to the low reward arm, 1,532 participants to the medium reward arm, and 1,544 participants to the high reward arm. Participants in the trial arms were balanced on most observable characteristics (age, gender, ethnicity, index of deprivation, device used, baseline dietary intake).
- The characteristics of participants randomised was in most cases quite close to the stated targets for demographics. This was particularly true for ethnicity, deprivation level, and age, but much less so for gender (for which a 50:50 split between female and male app users was not achieved at any point).
- The sample of individuals who were randomised were more likely to be female, older, white, and income-deprived than the general population of Wolverhampton LA, and at baseline had both a higher number of steps and a less healthy diet than the general UK population. This was in part due to the characteristics of those who consented to participate in the study initially, but also due to different rates of 'drop-off' during onboarding.

During the onboarding phase, participants were asked to create an account, verify their contact details and address, provide personal data and clinical consent in the app, connect their tracking device (or order a free one if required), and pass some baseline requirements for randomisation. Within this stage, participants dropped off in different stages, as also displayed in Figure 4:

- 28,281 participants completed registration (*6,619 participants not completing one or more steps in registration*).
- 12,767 (37%) participants passed baseline requirements to proceed to randomisation

12,767 participants represents 6.4% (or 1 in 16) of the entire adult population of Wolverhampton LA - a significant achievement.

Figure 4: Funnel diagram showing user drop-off during the onboarding phase



Drop-off between registration and randomisation in the onboarding phase differed across demographic groups. Table 8 shows the proportion of participants by observable characteristics at registration (n = 28,281) and randomisation (n = 12,767), comparing them to the targeted demographic composition of participants. While participants of a white ethnicity were underrepresented (relative to the stated target) in the sample of participants who completed registration (being 'only' 58%), they completed the onboarding phase (and were hence randomised) at a higher rate than those of a non-white ethnicity, and so made up 68% of those randomised - exactly in line with the target.

In a similar vein, those who were defined as most income-deprived (those scoring '1' or '2' in the 1-10 point Index of Multiple Deprivation) were slightly 'overrepresented' in the sample of participants who completed registration, but the 50:50 split between those most income-deprived and others within the randomised sample almost exactly in line with targets. However, this was much less the case for gender, for which a 50:50 split between female and male app users was not achieved at any point, and indeed the sample was even more disproportionately female at randomisation than it was amongst those who had completed registration. Last, the proportion of participants aged over 45 among those who completed registration and were randomised was just slightly lower than the target. For a full presentation of demographic details of the randomised cohort, see Table D1.

Table 8: A comparison of demographic data between those who completed registration and those who passed baseline requirements for randomisation, and the targeted composition of app users as stated in advance of the study

Subgroup	**Targeted composition of subgroups (% in each category)**	**Composition of group who completed registration (% in each category) [n = 28,281]**	**Composition of those randomised and consented to data sharing (% in each category) [n = 12,767]**
Ethnicity category	**Target (%)**	**Completed registration (%)**	**Randomised (%)**
White	68	58	68
Asian or Asian British	18	25	18
Black, African, Caribbean, or Black British	7	8	7
Mixed or Multiple Ethnic Backgrounds	5	4	4
Gender category	**Target (%)**	**Completed registration (%)**	**Randomised (%)**
Female	50	64	67
Male	50	36	33
Deprivation (IMD) category	**Target (%)**	**Completed registration (%)**	**Randomised (%)**
Most deprived (scoring 1 or 2 on the IMD)	51	54	50
Other (scoring above 2 on the IMD)	49	46	50
Age category	**Target (%)**	**Completed registration (%)**	**Randomised (%)**
Aged 45 or over	45	41	42

Trial arm	Target (%)	Completed registration (%)	Randomised (%)
Control	61	*N/A*	61
Low incentive	15	*N/A*	15
Medium incentive	12	*N/A*	12
High incentive	12	*N/A*	12

Notes: Percentages may not sum to 100% due to rounding and some categories not being listed. The RAG (Red, Amber, Green) colour coding of the rightmost column offers a quick visual guide for whether a subgroup within a category in the randomised cohort was either within 2 percentage points of the stated target proportion (green), between 3 and 5 percentage points off the stated target (amber), or 'missed' the target by more than 5 percentage points (red). In effect, those cells shaded red represent the subgroups with disproportionately high or low representation in the randomised sample, relative to stated targets. Full ethnic breakdown at randomisation: 1. White: 8,712; 2. Asian / Asian British: 2,270; 3. Black / African / Caribbean / Black British: 874; 4. Mixed / Multiple ethnic groups: 459; 5. Other ethnic group: 171; No data: 281.

Comparison with Wolverhampton population

The characteristics of the randomised cohort, described above, were different from those of the general population of Wolverhampton, in which women make up 51%, the median age is 38, people of white ethnicity make up 61%, and 21% are categorised as income deprived (using the IMD; Office for National Statistics 2021a, 2023, 2021b). The sample of individuals who were randomised into this pilot was therefore more female, older (median age of 41), white, and income deprived than the general population of Wolverhampton LA.

Comparison of randomised sample baseline PA and diet with Wolverhampton population

12,767 participants provided baseline data on physical activity and food consumption, which enabled them to 'pass' baseline and be randomised. 3,952 (31%) achieved an average of 30 or more minutes of moderate/vigorous physical activity ('MVPA') per week at baseline, with the remaining 8,815 (69%) failing to meet this threshold. DHSC classifies a person reporting less than 30 minutes of moderately intensive physical activity or 15 minutes vigorous activity per week, or an equivalent combination, as inactive, and recommends at least 150 minutes of moderately intensive physical activity (or 75 minutes vigorous activity per week, or

an equivalent combination; NHS Digital 2022). A larger proportion of our sample was classified as inactive than the average found in the UK population (11%; NHS Digital 2022). However, we don't see this as a robust comparison, given the data collected during the pilot is likely to be incomplete and as such, less reliable, this is because only a very small sample, 5% of those randomised, had valid MVPA readings for a full week (7 consecutive days of their baseline period). While most participants provided 3 or 4 valid readings, it is possible that participants did some MVPA in the days during which they did not wear the device. In addition, there is some evidence that commercial wearable fitness trackers (e.g. Fitbit, Apple, Garmin, Better Health: Rewards) work best when measuring steps rather than MVPA, compared to accelerometer and research wearable fitness trackers (such as Axivity or ActiGraph; Mair et al. 2021).

For this reason, steps per day at baseline is a better metric for comparison to national averages. The average number of steps per day at baseline, for those randomised, was 7,438 (minimum: 504, maximum: 22,160) - considerably higher than the UK average of 5,444 cited by Althoff and colleagues (2017).

Of those who provided baseline data on their fruit and vegetable consumption, 2,618 (21%) reported consuming an average of at least three portions of fruit and vegetable per day, while 9,991 (79%) did not. This baseline fruit and vegetable consumption was considerably lower than the average in England; 33% of adults aged 19-64 in England report consuming at least five portions of fruit and vegetables per day (National Diet and Nutrition Survey, 2020).

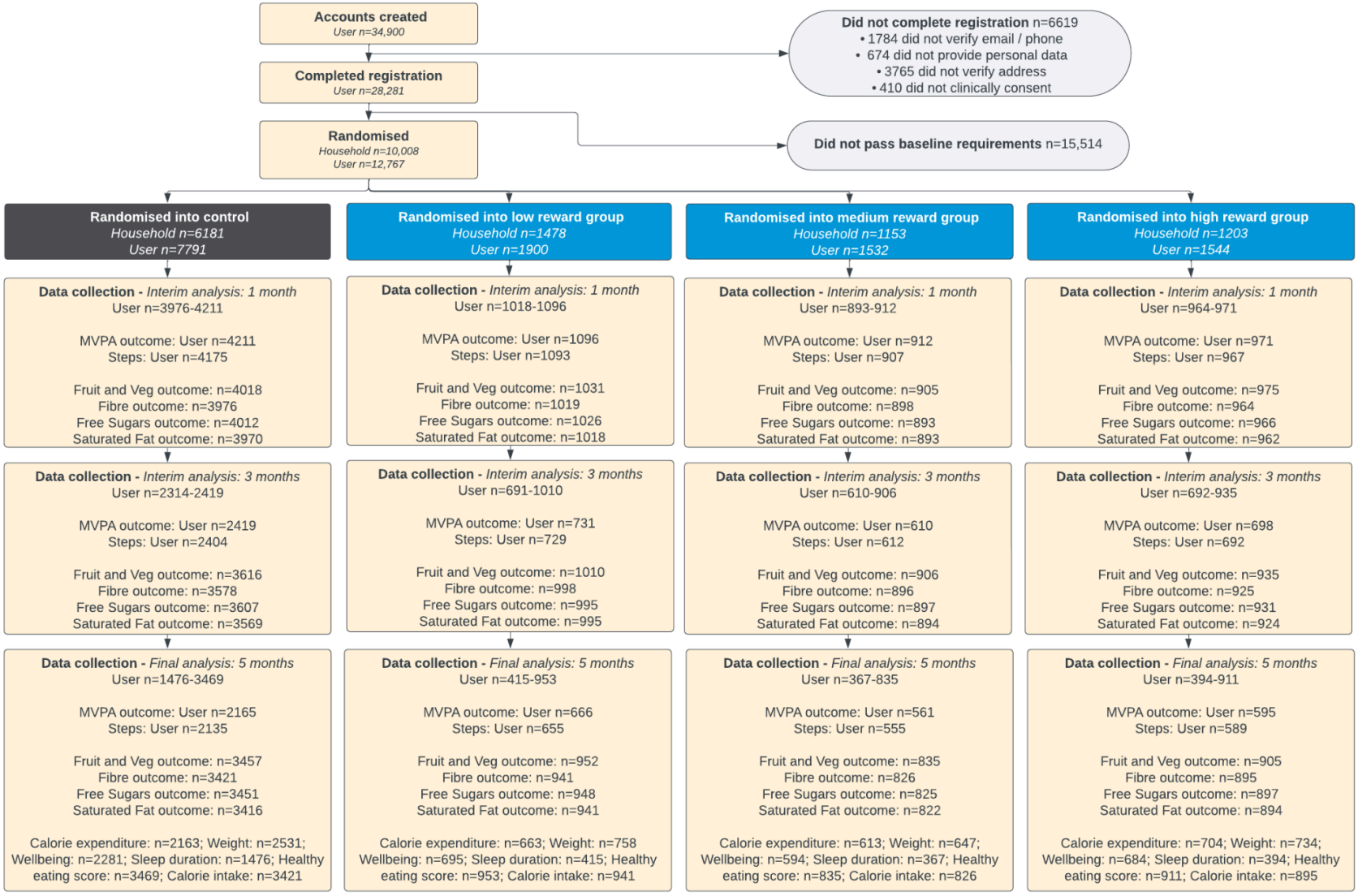
4.1.3 Sample size: from randomisation to the end of the pilot

- Attrition at the final data collection point (month 5) was higher among men, younger people (under 41), people from more deprived areas, and people using the Better Health: Rewards tracker. This means that participants at the end of the pilot were more likely to be female, older, and less deprived than participants at the start of the pilot.
- The final sample size was 3,934 - 3,987 for physical activity (steps and MVPA respectively), 5% below the original target of 4,200, and 6,073 - 6,149 for diet outcomes, substantially higher than the original target.

As explained in the Sample section, the original target sample size was 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 the relevant section in the TP). To achieve a post-attrition sample of 4,200 participants, HUL expected to recruit a sample of around 25,800 participants at the beginning of the baseline stage, given the attrition rate at various pilot stages estimated by HUL.

This target figure was met, with in total 28,281 people completing registration. Completion of registration required the user to download the app; provide personal data, verify their email address, phone number and address; and clinically consent to take part in the pilot. Because more people than expected did not pass the baseline checks, the sample size at the point of randomisation was 12,767; 2,700 lower than the target of 15,500. Section 5.1 provides more information on the characteristics of randomised participants. The CONSORT diagram below (Figure 5) summarises the flow of participants from start to finish.

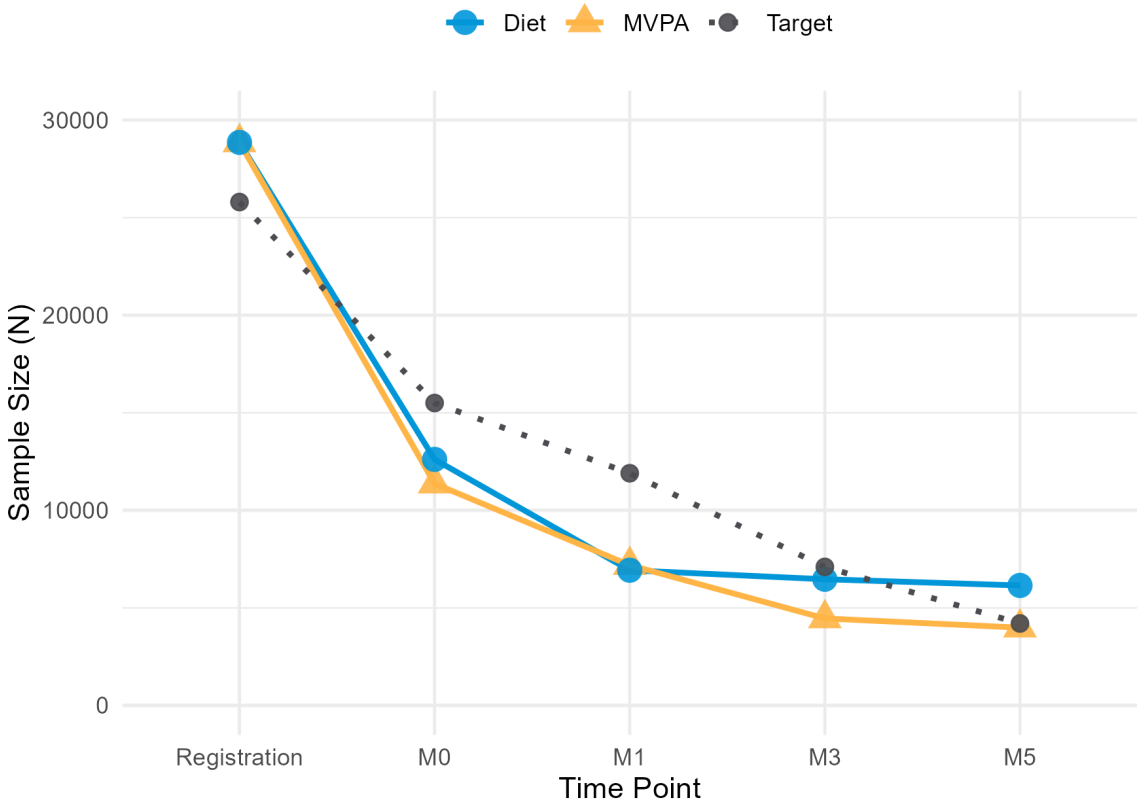
Figure 5: Consort diagram



However, thanks to a lower than expected level of attrition (defined from an impact evaluation point of view as a user failing to provide adequate data for analysis 1, 3, or 5 months after randomisation), especially between the two final data collection points, the final sample size for MVPA was just 5% below the original target sample size, while for diet outcomes it was almost 50% higher. At the final data collection point, 4,937 participants submitted physical activity data, while 6,221 submitted diet data. After applying our filtering rules to remove unlikely or incomplete observations, our final sample sizes are 3,987 for MVPA, 3,934 for steps and 6,073 - 6,149 for diet outcomes. In the remainder of the report, when referring to sample size analysed at M5 in the text, we report only sample size figures for MVPA (to represent physical activity outcomes) and fruit and vegetables (to represent diet outcomes) for brevity. The regression tables in Appendix H include the sample sizes for all analysed outcomes).

Attrition for diet outcomes was substantially higher at M0 and M1. As a result, the decision was made to increase the monetary payment for completing an intake25 survey for M3 and M5, from £5 to £10. This may be the reason for the reduction in the attrition rate after this point. For activity outcomes the attrition between M3 and M5 was lower than expected as well. This could be due to the increased promotional activity around the final data collection point.

Figure 6: Realised versus target sample size throughout the pilot



4.1.4 Balance at randomisation

At randomisation, 7,791 participants were randomised to the control arm, 1,900 participants to the low reward arm, 1,532 participants to the medium reward arm, and 1,544 participants to the high reward arm (ratios consistent with what defined in the TP). Participants in the control and reward arms were balanced on most observable characteristics (age, gender, ethnicity, index of deprivation,device used,baseline dietary intake). Full balance tables focusing on the randomised sample are available in Appendix C.

4.1.5 Attrition across demographic groups

Attrition at month 5 was higher among men, younger people, people from more deprived areas, and people using the Better Health: Rewards tracker.

Table 9 summarises the proportion of pilot participants with a given observable characteristic in the randomised sample and the sample providing their physical activity and diet data 5 months after randomisation, respectively. Participants providing data at the end of the pilot were more likely to be female, older, and less deprived than participants at the start of the pilot. Section 6.2.4 provides some qualitative findings on early engagement with the intervention and discusses barriers to sustained engagement (i.e. factors that contributed to user churn-out).

Table 9: Proportion of pilot participants with a given observable characteristic in the randomised sample and the final sample 5 months after randomisation

	Target (%)	**At randomisation (%)**	**Provided data at M5 (diet) (%)**	**Provided data at M5 (physical activity) (%)**
Women	50	67	68	69
Over 45	45	42	53	60
Deprivation Index < 3	51	50	48	45

Using the Better Health: Rewards tracker	-	71	70	64
MVPA at baseline > 30 min/week	-	31	36	41
F&V at baseline > 240 grams/day	-	21	21	25

As expected, the level of attrition was also influenced by what reward arm participants were randomised into: the higher the rewards, the lower the attrition. The relative differences were lower than expected. As a result, the sample size for the control arm at month 5 was approximately 20% larger than the overall reward arm.

Higher attrition among pilot participants with a Better Health: Rewards tracker is compatible with a few speculative explanations:

- Participants reported concerns about the quality of the Better Health: Rewards tracker (see Section 6.2.4 which discusses software and hardware issues participants experienced with the Better Health: Rewards tracker). It is possible that participants with this brand of wearable fitness tracker were less likely to provide data as a result. However, the qualitative work did not ask participants about their experience with wearable fitness trackers of other brands, so we cannot conclusively conclude that these concerns were unique to the Better Health: Rewards tracker.
- Participants with a Better Health: Rewards tracker were likely not to own another wearable fitness tracker (this cannot be validated in the data). Hence, it is likely that they weren't already used to wearing a device for PA or syncing a device regularly to provide their data.

Table 10: Level of estimated and realised attrition by trial arm

Trial arm	Estimated attrition	Realised attrition Diet at M5	Realised attrition MVPA at M5
High	77%	41%	61%
Medium	78%	45%	63%
Low	81%	50%	65%
Control	87%	56%	72%

Notes: A table with attrition at each measuring point can be found in Appendix D, Table D3

Of those in the pooled reward arms, the drop out rate was 46% for diet outcomes and 63% for activity outcomes, which in both cases was 10% lower than that in the control arm. As a result, and in accordance with the TP, we adjusted the analysis using inverse probability weighting (IPW), a method proven effective in reducing selection bias for longitudinal studies. The IPW works by modelling the probability of successful retention 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 trial arms. Appendix E provides technical information on the methodology.

Apart from deprivation level (the share of participants from more deprived areas decreased less in the high reward arm than in the other trial arms), this higher level of attrition was the same for each trial arm.

4.2 Impact

Table 11 summarises the results of the impact evaluation for the primary outcomes for PA and diet.

Table 11: Summary of impact estimates

Outcome	Sample size (participants)	Comparison	Timepoint	Effect size (95% CI)
Primary outcome Physical Activity: Steps	3987	Pooled reward arms vs. control	5 months after randomisation	+42 steps (-170 - 254)
Primary outcome Physical Activity: MVPA	3934	Pooled reward arms vs. control	5 months after randomisation	+1.1 minutes (-0.05 - 2.3)
Primary outcome Diet: Fruit and veg	6149	Pooled reward arms vs. control	5 months after randomisation	+21 grams (14 - 28) **
Primary outcome Diet: Fibre	6083	Pooled reward arms vs. control	5 months after randomisation	+0.35 grams (0.12 - 0.59) **
Primary outcome Diet: Saturated fat (as % of food calories)	6073	Pooled reward arms vs. control	5 months after randomisation	- 0.10 percentage points (-0.30 - 0.09)
Primary outcome Diet: Free sugars	6121	Pooled reward arms vs. control	5 months after randomisation	-0.21 percentage points (-0.55

(as % of food calories)				- 0.12)

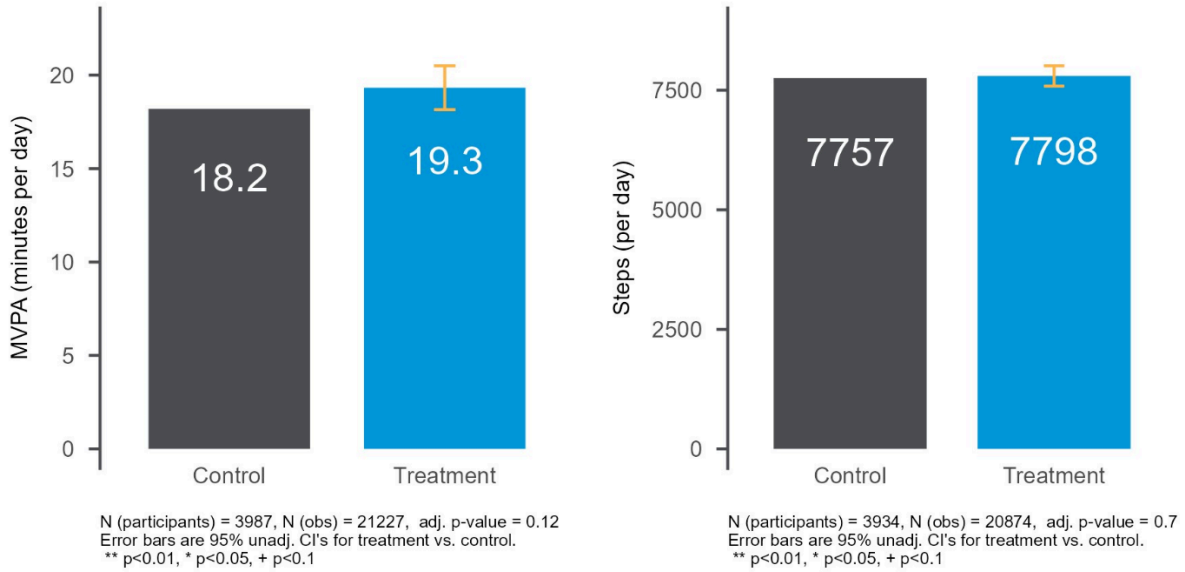
*Notes: Stars indicate the adjusted p-values: **:p<0.01*

4.2.1 Primary outcomes: does a financial incentive scheme increase moderate-vigorous physical activity?

The analysis did not reveal a statistically significant impact from offering financial incentives on participants' PA (MVPA and steps), on average, although the direction of the descriptive effects was positive. The average difference in steps and MVPA per day between the control and reward arm was small (+1.1 min/day of MVPA (95% CI: -0.05 - 2.30) and +42 steps/day (95% CI: -170 - 253)).

Sensitivity analysis did reveal a statistically significant impact on both MVPA (+1.9 min/day, 95% CI: 1.0 - 2.9) and steps (+256, 95% CI: 71 - 442).

Figure 7: Impact of offering any financial incentive on PA outcomes after 5 months among the pooled reward arms (left panel: MVPA; right panel: steps)



Although the direction of the treatment effect was positive, the analysis showed that offering financial incentives didn't statistically significantly impact participants' PA. MVPA saw an increase in 1.1 minutes of exercise per day, (95% CI -0.05 - 2.30) in the reward arm compared to the control arm, from 18.2 minutes to 19.3 minutes, with an adjusted p-value of 0.12. Steps increased by 42 steps per day (95% CI: -170- 253) , from 7757 to 7798, with an adjusted p-value of 0.70. The full regressions for these analyses can be found in Table H1.

On the other hand, the number of minutes of MVPA per day was lower than the average of 48.1 min/day measured among UK Biobank users with accelerometer data (Rowlands et al. 2021). As noted in the Reach and Limitation Section, we think that MVPA from this study have little comparability to the national average, so we should be cautious in interpreting this comparison - steps seem a more reliable metric for external comparisons.

4.2.1.1 Sensitivity analysis

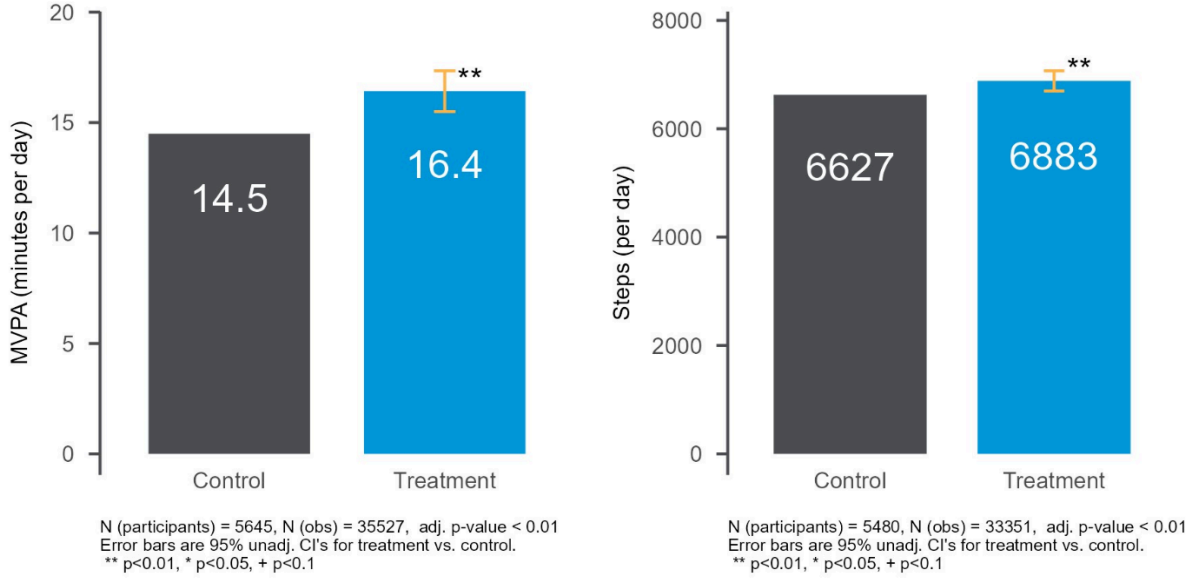
According to the TP, observations on days when the registered wearable time was less than 6 hours were filtered out. Invalid observations were replaced with values from the same day in the 2 weeks before or after, provided valid observations were registered on at least 1 of those 4 days (see section 2.6.3.2 of the TP for further details). These replacement observations had to meet the same filtering criteria as the primary observation, and observations from 2 weeks before or after the main observation were only considered if there were no valid observations in the weeks immediately preceding or following the main week. However, evidence suggests these filtering rules were potentially filtering out a substantial portion of valid data.

A sensitivity analysis was conducted in which observations based on less than 6 hours of wearable time were not filtered out and missing data were imputed with observed values on the same day from the 2 weeks before or after, instead of only replacing invalid reads. This analysis resulted in a sample size increase of 42% participants for the MVPA analysis and 39% participants in the steps analysis at M5, and respectively a 67% (MVPA) and 60% (steps) increase in daily reads. As a result, the composition of this sample was more balanced across several demographic characteristics. It included more men, more participants who were younger, and a greater number of those who were more income-deprived, suggesting the results of this analysis may be more generalisable to the wider UK population.

In this alternative analysis, offering an incentive statistically significantly increases both MVPA and steps (MVPA: effect size of 1.9 minutes (CI: 1.0 - 2.9), adj. p-value < 0.001; steps: effect size of 256 steps (CI: 71 - 442), adjusted p-value < 0.01). Not only were the confidence intervals smaller, which is an expected result from an increase in sample size, the effect size was also larger. It also resulted in a lower level of PA in the control group, likely because it included people who wore the wearable fitness tracker for shorter periods of time. The full regressions for these analyses can be found in Appendix H, Table H11.

This larger treatment effect could be attributed to the inclusion of more people who typically wore their wearable fitness tracker for shorter periods of time, possibly because they wear it primarily to meet their targets and then remove it. These individuals were not included in the main analysis but the higher treatment effect after they were included in the sensitivity analysis suggests they might be more responsive to financial incentives. The absence of discernible differences in wearable time between the pooled reward arms in both analyses, coupled with the increased effect observed in both MVPA and steps - the latter which can be measured without wearing the wearable fitness tracker - suggests there could be a greater efficacy of the intervention among this group.

Figure 8: Sensitivity analysis on the impact of offering any financial incentive on PA outcomes after 5 months between the control arm and the pooled reward arms (left panel: MVPA; right panel: steps)



As described in the TP, we also performed a sensitivity analysis on the diet primary outcome variables. We replaced values of each primary diet outcome below the 1st percentile with the 1st percentile, and values above the 99th percentile with the 99th percentile. This analysis yielded almost identical results, which are available upon request.

4.2.2 Primary outcomes: does offering financial incentives improve the healthfulness of participants' diet?

The quantitative analysis showed some evidence that offering financial incentives can impact participants' diets in a positive direction. The direction of the treatment effects were as expected, and offering financial incentives had a small but statistically significant effect on participants' consumption of fruit and vegetables (treatment effect: 21g, 95% CI 14 g - 28 g, adjusted p-value < 0.01) and fibre (treatment effect: 0.4g, 95% CI 0.12 g - 0.59 g, adjusted p-value < 0.01).

Participants reported a range of improvements to their diet in qualitative findings. Many reflected on the successes of 'swaps' challenges on changing their diet, such as replacing full-fat dairy products with lower-fat or fat-free options, replacing sugary drinks with water or sugar-free alternatives, or replacing unhealthy snacks with fruit. Others noted increased fruit and fibre consumption. Several participants also reflected on changes to their diet as a whole, such as increasing awareness of portion sizes, consciousness of the importance of a balanced diet, and awareness of the nutritional content of food.

The quantitative analysis showed some evidence that offering financial incentives can impact participants' diets in a positive direction. The full regressions for these analyses can be found in Appendix H, Table H2.

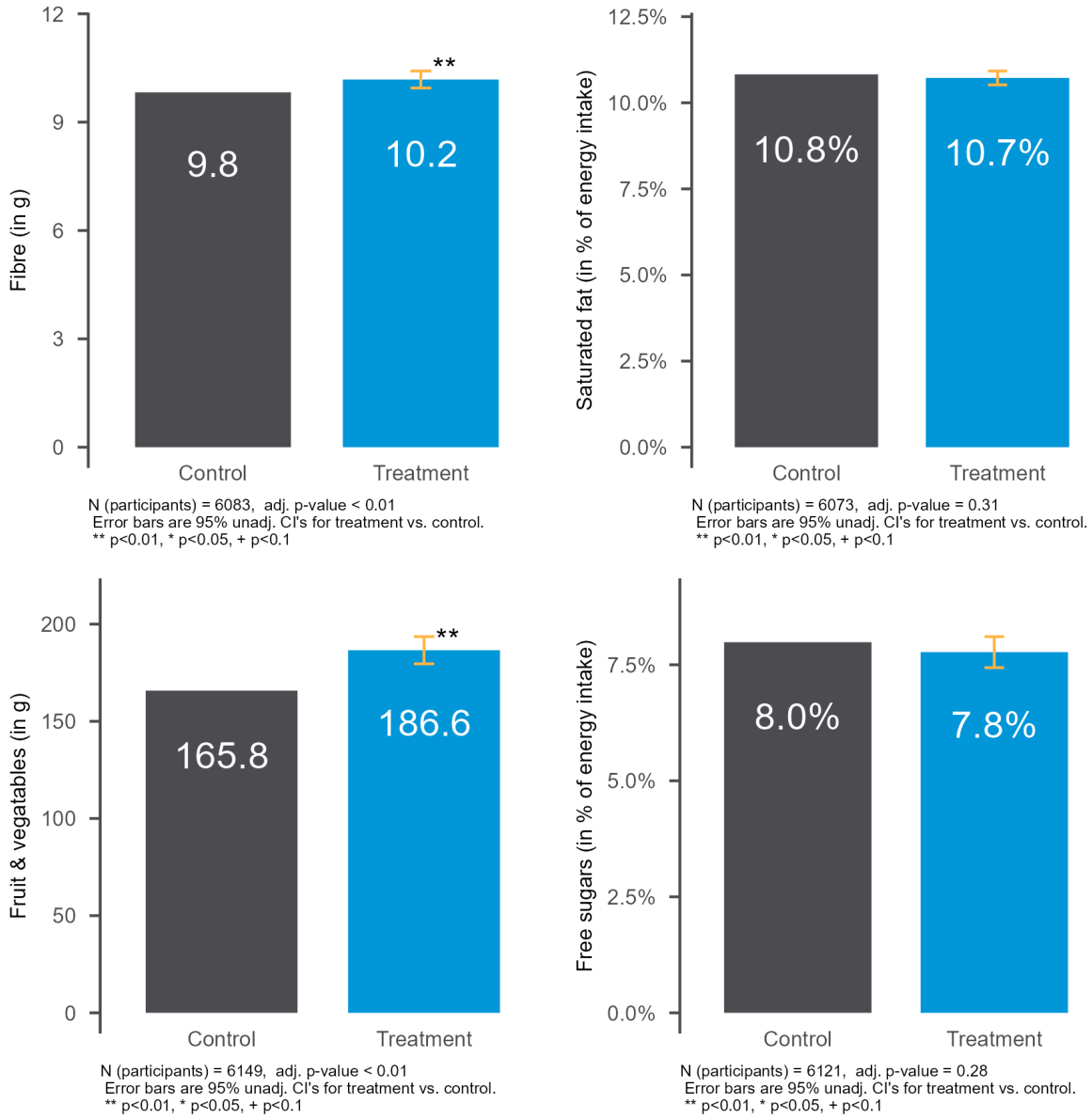
- **Fibre Intake**: The control arm consumed an average of 9.8g of fibre, whereas the average consumption in the reward arms was 10.2g. With an adjusted p-value of 0.01, this was a statistically significant increase (treatment effect: 0.35g, 95% CI 0.12 g - 0.59 g).

- **Saturated Fat Intake**: The control arm average intake was 10.8%, with a treatment effect of -0.10 percentage points (95% CI -0.30 pp - 0.09 pp). The adjusted p-value here is 0.31, which indicates that this difference was not statistically significant.
- **Fruit and vegetable intake**: The control arm consumed an average of 166g of vegetables, whereas the pooled reward arms consumed more, with an average of 187g. (treatment effect: 21g, 95% CI 14 g - 28 g). The adjusted p-value was < 0.001, indicating a statistically significant increase in vegetable intake in the pooled reward arms.
- **Free Sugars Intake**: The control arm's average intake was 8.0%, and the pooled reward arms' was 7.8% (treatment effect: -0.21 percentage points, 95% CI -0.5 pp - 0.1 pp). With an adjusted p-value of 0.28, this suggests that the difference in free sugars intake was not statistically significant.

It is worth noting again that for some dietary outcomes, the sample analysed exhibited healthier habits than the typical British adult at month 5. According to the National Diet and Nutrition Survey, the mean intake of free sugars as a percentage of total energy intake was 9.9% for those aged 19 to 64 years and 9.4% for those aged 65 years and over (Public Health England Data, 2020). Additionally, the study sample consumed less saturated fat as a proportion of total energy than the British population (12.3% and 13.3% for adults aged 19-64, and 65 and older respectively).

However, the study sample consumed less fibre than the typical British adult (19.7g and 18.7g for adults aged 19-64, and 65 and older respectively), and less fruit and vegetables than the typical British adult (311g and 307g for adults aged 19-64, and 65 and older respectively.

Figure 9: Impact of offering any financial incentives on diet outcomes after 5 months between the control arm and the pooled reward arms (top left panel: fibre; top right: saturated fat; bottom right: fruit and vegetables; bottom right: free sugars)



4.2.3 Secondary outcomes: does offering financial incentives significantly affect broader PA and diet constructs?

According to the prespecified analysis, we looked at four constructs aimed at capturing holistic effects on PA and diet, namely 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. The full regression tables of these analyses can be found in Appendix H.

Energy expenditure

The analysis revealed no effect on energy expenditure, measured in calories per day with an effect size of -2.29 (95% CI: -16.0 - 11.4, adjusted p-value = 0.74), on a control arm average of 1,838 kcal per day. However, of the PA outcome measures energy expenditure is the most challenging one to accurately measure, as it depends on a range of inputs, including self-reported factors such as weight, height, age and gender. The full regressions for this analysis can be found in Table H9.

Energy intake

The analysis did not reveal a significant effect on calorie intake for the reward arm. The average daily calorie intake in the control arm was 1204 kcal per day. The estimated treatment effect was 1.7kcal (CI: -24 - 27), which was not significant at the 5% level (adjusted p-value = 0.89). The full regressions for this analysis can be found in Appendix H, Table H10.

Healthy Eating score

The average healthy eating score, which is measured on a scale from 1 to 6, increased by a modest 0.04 (CI: 0.00 - 0.08), which was not significant at the 5% level (adjusted p-value = 0.30).The full regressions for this analysis can be found in Table H10.

Weight

Some participants in qualitative work reported experiencing weight loss following participation in the intervention. For some, this represented further incentive to continue engaging with the programme.

However, the analysis did not find a significant effect from the treatment on the users' self-reported weight (effect size: -0.26kg, CI: (-0.76 - 0.24, adjusted p-value = 0.42), with the average weight in the control arm being 81.0kg. 4,670 participants reported their weight in the M5 period. The full regressions for these analyses can be found in Table H10.

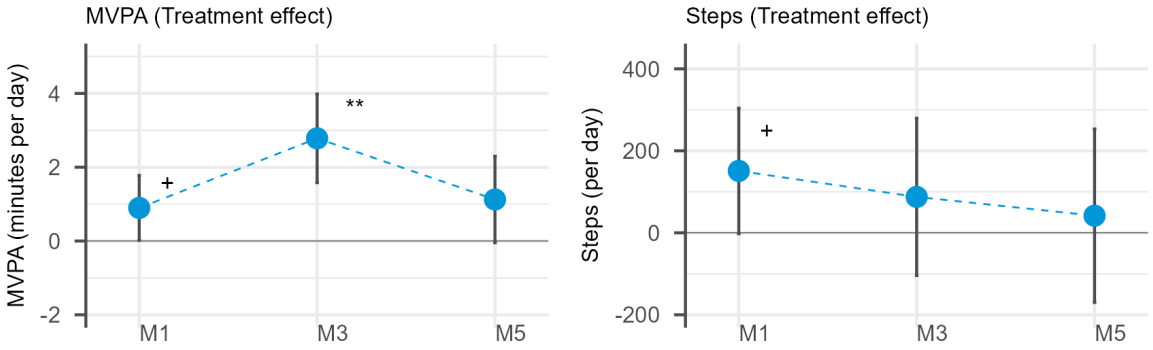
4.2.4 Secondary outcomes: does offering financial incentives affect outcomes after 1, 3 and 5 months?

The analysis does not offer a clear picture of patterns in effect size over time (1-3-5 months from randomisation).

This analysis is aimed at investigating whether offering a financial incentive to encourage PA and diet is more likely to work in the very short/immediate period, versus a longer timeline (e.g. after five months). To do so, the study analysed the effects of offering any financial incentives on the primary outcomes at three points in time: one, three and five months after randomisation.

The results don't offer a clear interpretation because the sample of participants for the analysis is different for each point in time. As time goes by, more participants drop out (stop providing data) - and, as the previous sections described, it is participants who are more deprived and younger that are increasingly more likely to drop out.
With this caveat in mind, the pattern of the treatment effect exhibited an inverse U shape for MVPA (small effects at the beginning and at the end, larger effects half way through) and decreasing effects on steps (bigger effects at the beginning and smaller towards the end of the pilot). This is depicted in Figure 10.

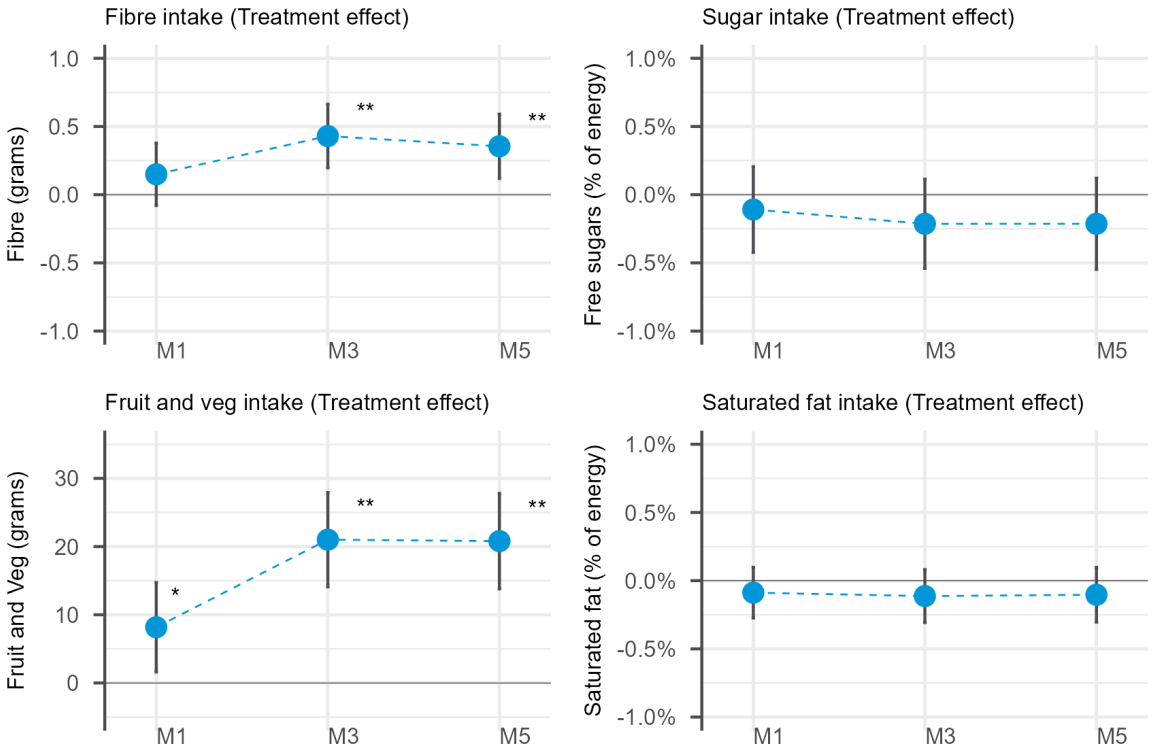
Figure 10: Treatment effects for PA outcomes after 1, 3 and 5 months (pooled rewards arms vs control group). Left panel: MVPA. Right panel: steps



*Notes: **: p < 0.01, *: p < 0.05, +: p < 0.1. p-value adjusted for multiple comparisons*

On the other hand, the effect on diet outcome over time showed a flat or slightly increasing relationship, where the effect seems to increase as time passes by initially, and then remains fairly constant. This is depicted in Figure 11.

Figure 11: Treatment effects for diet outcomes after 1, 3 and 5 months (pooled rewards arms vs control group). Top left panel: fibre intake. Bottom left panel: Fruit and vegetables intake. Top right panel: sugar intake. Bottom right panel: Saturated fat intake



*Notes: **: p < 0.01, *: p < 0.05, +: p < 0.1. p-value adjusted for multiple comparisons*

Table 12 shows point estimates and confidence intervals from these analyses. The full regressions for these analyses can be found in Appendix H, Tables H1 - H6.

Table 12: Point estimate, CI and adjusted p-values for the treatment effects at M1, M3 and M5 among the pooled reward arms

Analysis	Estimated treatment effect	CI (Lower bound)	CI (Upper bound)	Sample size	Adj. P
PA: MVPA					
Effect at M1	0.90	0.02	1.78	7190	0.054
Effect at M3	2.78	1.58	3.98	4458	<0.001
Effect at M5	1.13	-0.05	2.30	3987	0.121
PA: Steps					
Effect at M1	151	-2	304	7142	0.054
Effect at M3	88	-104	279	4437	0.371
Effect at M5	42	-170	253	3934	0.700
Diet: Fibre					
Effect at M1	0.15	-0.08	0.38	6857	0.401
Effect at M3	0.43	0.20	0.66	6397	0.001
Effect at M5	0.35	0.12	0.59	6083	0.006
Diet: F&V					
Effect at M1	8.2	1.6	14.7	6929	0.059
Effect at M3	21.0	14.1	27.9	6467	<0.001
Effect at M5	20.8	13.8	27.8	6149	<0.001
Diet: Sugar					
Effect at M1	-0.11pp	-0.42pp	0.21pp	6897	0.497
Effect at M3	-0.21pp	-0.54pp	0.11pp	6430	0.254
Effect at M5	-0.21pp	-0.55pp	0.12pp	6121	0.282
Diet: Saturated fat					
Effect at M1	-0.09pp	-0.27pp	0.10pp	6843	0.471
Effect at M3	-0.11pp	-0.31pp	0.08pp	6382	0.254
Effect at M5	-0.10pp	-0.30pp	0.10pp	6073	0.314

4.2.5 Secondary outcomes: do different levels of financial incentives achieve different effect sizes on participants' dietary intake and PA?

The analysis provides evidence that there was a positive relationship between incentive level and effect sizes, specifically for the dietary outcomes fibre, fruit and vegetables, and saturated fat intake. However, these differences were not statistically significant.

The effect on PA outcomes was less consistent, with a statistically significant increase in step count found only for participants in the medium reward arm.

A crucial aspect of research on financial incentives to motivate behavioural change is the optimal value of incentives. The question is particularly sensitive because of the cost of scaling up these types of interventions.

Results on user engagement (see Section 6.2) already highlight that participants randomised into the high reward arm were more engaged across a wide array of metrics with the app than participants in the medium and low reward arm, and that participants in the control arm were the least engaged.

An analysis of the impact of different levels of financial incentives against no financial incentive on PA outcomes doesn't reveal a clear indication of a similar pattern. Figure 12 suggests that participants in the medium reward arm respond more to the treatment, and (statistically) significantly so for steps. However, as other indicators, such as challenges completed, do not suggest a stronger effect among this group, it is likely that this effect was due to chance.

On the other hand, the analysis of the impact of different levels of financial incentives against no financial incentive on diet outcomes showed a similar positive relationship between incentive levels and behavioural change - descriptive effects suggest bigger incentives led to a bigger change (see Figure 13). For fruit and vegetables, the high reward arm saw a statistically significant increase of 29 grams (roughly a third of a portion) compared to the control arm, and a decrease in saturated fat of 0.4% compared to the control arm. When comparing the medium and low reward arm to the control arm, the increases in fruit and vegetable intakes

we see are smaller than for the high reward arm, yet still statistically significant. To note, the differences between treatment arms (i.e. low reward vs medium reward) are not significant at the 5% level.

Figure 12: Impact of offering different levels of financial incentives on PA outcomes after 5 months across the different trial arms

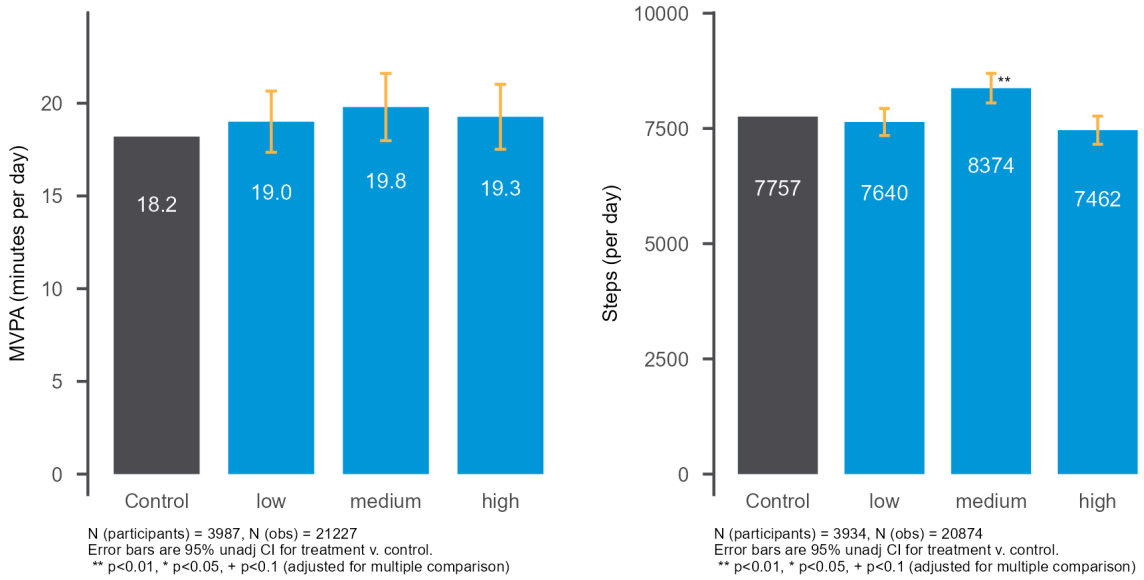
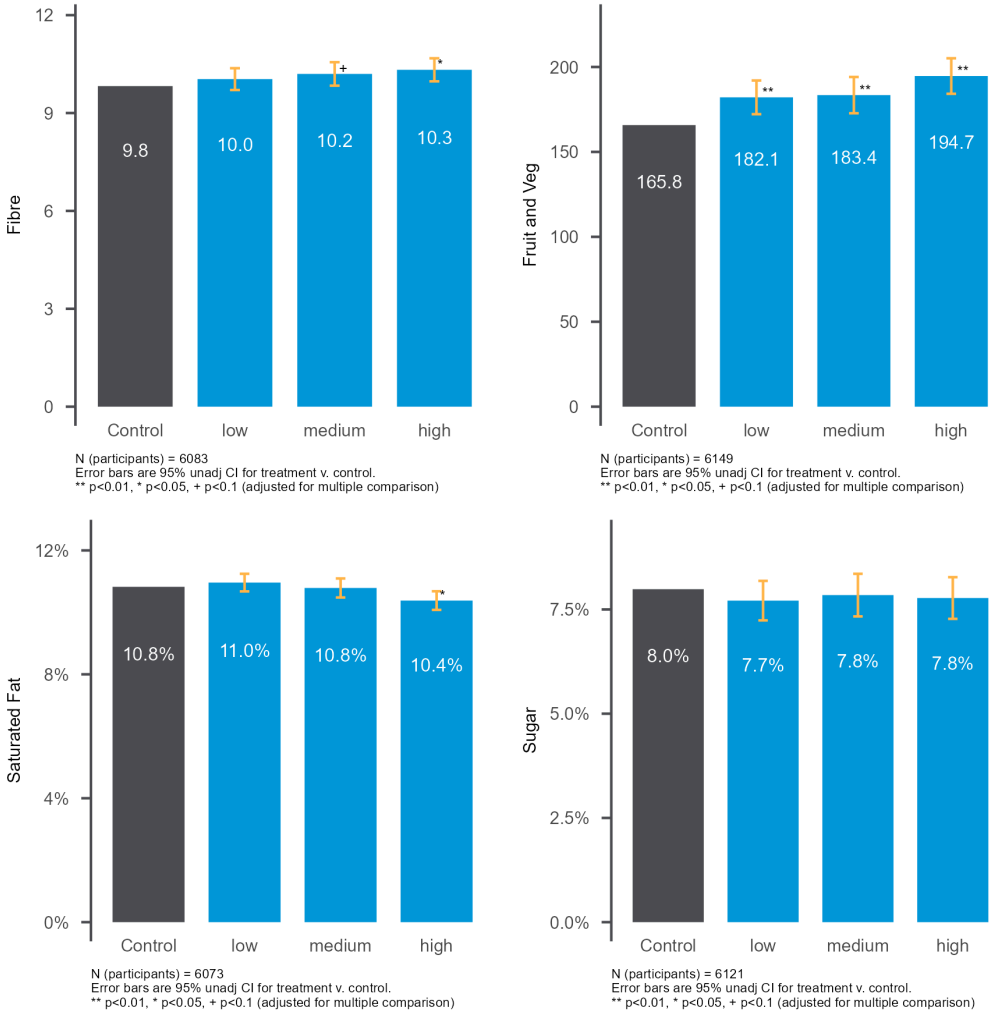


Figure 13: Impact of offering different levels of financial incentives on diet outcomes at M5 across the different trial arms



4.2.6 Exploratory outcomes: does offering financial incentives impact specific population subgroups?

The analysis provides three key insights:

- The impact of financial incentives on physical activity and diet outcomes varied between socio-economic groups, with more deprived individuals being more likely to change their PA and less deprived individuals being more likely to change their diet.

- The treatment's effectiveness on physical activity was notably higher among participants who were initially inactive.
- The influence of financial incentives on both physical activity and dietary patterns appeared to be more pronounced in older participants (aged over 41).

Figure 14 provides an overview of treatment effects when we perform the primary outcome regressions only on specific subgroups of the population. This analysis allows us to dive deeper into the question whether financial incentives are likely to drive behaviour change for specific groups. This analysis yields three key insights:

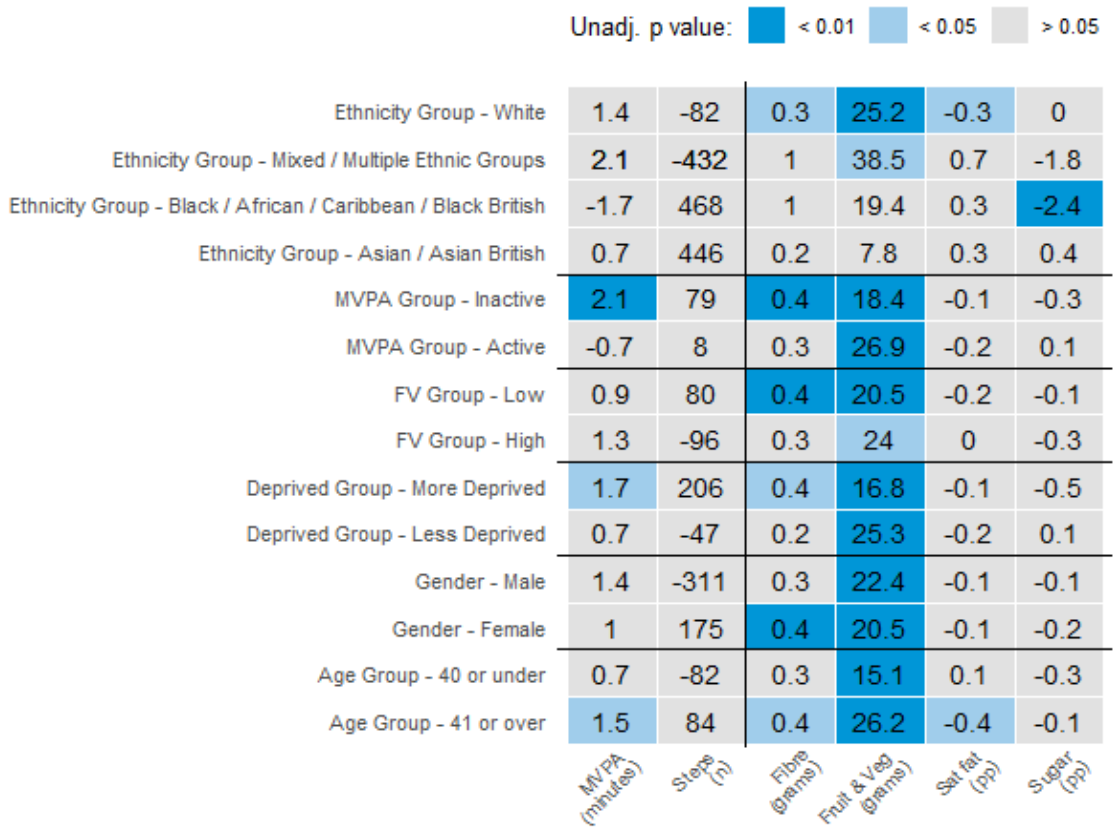
- The first insight revealed a disparity in the impact of financial incentives on physical activity and diet outcomes between socio-economic groups (a different effect for different groups). Individuals from more deprived areas (IMD score 1-2) exhibited a greater increase in MVPA in response to the treatment, averaging an additional 1.7 minutes per day (CI: 0.04 - 3.4), compared to their less deprived counterparts who experienced a smaller increase of 0.7 minutes per day (CI: -1.0 - 2.3). This pattern extends to step count, with a more pronounced treatment effect of +206 steps (CI: -115 - 527) observed in the more deprived group, while the less deprived group's step count decreased by 47 steps (CI: -335 - 240) in the intervention group compared to the control group.
- Secondly, the treatment's effectiveness on physical activity was notably higher among participants who were initially inactive (less than 30 minutes of MVPA during the baseline week). In this subgroup, financial incentives led to an increase in MVPA by +2.1 minutes per day (CI: 0.8 - 3.5), while the participants who were deemed active at baseline saw a decrease in -0.7 minutes per day (CI: -2.7 - 1.4)
- Lastly, the influence of financial incentives on both physical activity and dietary patterns appeared to be more pronounced in older participants (aged over 41, the median age at baseline). In this demographic group, the incentives were associated with an increase in MVPA by +1.5 minutes per day (CI: 0.02 - 3.0), and a rise in fruit and vegetable consumption by 26 grams per

day (CI: 16 - 36). In contrast, the younger demographic saw an increase in MVPA by +0.7 minutes per day (CI: -1.2 - 2.6) and an increase in fruit and vegetable intake of 15 grams (CI: 5.5 - 25).

These insights suggest that the intervention is more effective amongst certain subgroups, but that this is not always consistent across all measured outcomes.

While these findings shed light on potential areas where the intervention has the most substantial impact, the confidence in these results does not match that of our primary analysis. Further research with a focus on specific subgroups and outcomes would allow for firmer conclusions and would provide a stronger foundation for understanding the nuanced effects of behavioural interventions across different populations.

Figure 14: Impact of offering financial incentives (pooled reward arms vs control) on PA and diet across demographic subgroups



Colour indicates significance (not adjusted for multiple comparisons). The number inside each cell displays the estimated average treatment effect. Subgroups with less than 100 people with valid data at M5 were excluded.

4.2.7 Exploratory outcomes: does the Financial Incentives Scheme have any significant impact on participants' sleep or mental health?

The analysis looked at two metrics as unintended consequences: sleep and mental health.

Sleep duration

The quantitative analysis didn't find evidence that financial incentives impacted participants' sleep duration (point estimate: -1.94 minutes; CI: -6.78 - 2.91; p-value: 0.433). However, this data was provided by only 2652 participants, a far smaller sample than the other outcomes. This suggests most participants did not use their wearable fitness tracker to track their sleep.

This was reflected in the qualitative interviews, in which some participants mentioned that they found the sleeping functionality of the Better Health: Rewards tracker confusing or inaccurate (e.g. didn't know how to set it up or did not understand how to interpret the wearable fitness tracker reads). Some participants reported that they might have engaged with sleeping functionality if the app had prompted them, and one user described taking action to improve sleep. None of the participants reported a change in their sleep duration or quality.

Mental health

Fully engaged participants reported feeling better, happier and in improved mental health in the interviews. This effect was also found in the quantitative impact evaluation: participants in the pooled reward arms scored 2 points higher on the WHO Wellbeing scale than participants in the control arm (CI: 0.87 - 3.21, p-value < 0.001). The survey was completed by 4,254 participants. Although the result is significant, it represents only a minor increase: from 59 to 61 on a 100-point scale. In addition, the interpretation of this result should be caveated as this outcome was self-reported in an app-pushed survey, and it is possible that participants with better mental health were more likely to respond to the survey.

Motivation to change physical activity and diet

Surveys measuring participants' motivation were optional, and only completed by 64 people in the M5 period, and only 600 people in total after the baseline period. Due to insufficient participant data availability, BIT was unable to complete analysis

of this outcome. More information is provided in the Limitations section of this report.

5. Implementation and process evaluation methodology

5.1 Objectives

While the impact evaluation tested the effectiveness of the Financial Incentives Scheme, the implementation and process evaluation (IPE) identified why and how the intervention achieved - or failed to achieve - the expected outcomes in relation to the Theory of Change (ToC). It also explored potential desirable and undesirable unintended consequences.

Following best-practice guidance from the Medical Research Council (MRC), BIT conducted 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.

As further outlined in the study design section, this mixed-methods approach incorporated qualitative data from interviews and focus groups, and quantitative data from routinely collected in-app user metrics. Using multiple data sources, the IPE had the objectives of:

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

5.2 Study design

5.2.1 Research questions

In line with the MRC's recommendations, BIT used the causal hypotheses outlined in the ToC and user journey (Section 2 of this report) to prioritise research topics for investigation (Moore et al. 2015). These were:

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

Table 13 below provides a summary of the high-level research questions for each topic along with the IPE methodology that was employed. A more detailed breakdown of the quantitative and qualitative methodology are outlined in Sections 5.3 and 5.4, respectively.

Table 13: Summary of research topics, research questions and methodologies

Research Topic	Research question		IPE	Methodology
Reach What factors affected the Financial Incentives Scheme's reach?	1.1	To what extent did it reach participants?	Quant	Descriptive analysis of metrics from marketing campaign and in-app engagement (summary statistics)
	1.2	What were the characteristics of recipients?		
	1.3	What was the role of marketing and communications in motivating participants to join the Financial	Qual	Thematic analysis of interviews and focus groups with

		Incentives Scheme?		recipients and non-recipients.
	1.4	What were the channels of reach and how did they affect sign-up?		
	1.5	How was the Financial Incentives Scheme perceived by recipients, and non-recipients of the financial incentives?		
	1.6	What were the barriers and facilitators to the Financial Incentives Scheme's reach?		
Engagement What factors affected the engagement of participants with the Financial Incentives Scheme?	2.1	To what extent did it engage participants?	Quant	i. Descriptive analysis of in-app engagement metrics and user demographics (summary statistics) ii. Regression analysis of in-app engagement metrics and user demographics
	2.2	What were the characteristics of those who engaged with the Financial Incentives Scheme for the duration of the pilot, those who partly engaged, and those who disengaged?		
	2.3	In what ways did participants 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 the Financial Incentives Scheme?	Qual	

	2.5	What were the barriers and facilitators to engagement with the Financial Incentives Scheme?		
Mechanisms of impact Through what mechanisms did the Financial Incentives Scheme affect behaviour change?	3.1	To what extent did 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 - affected the extent to which the Financial Incentives Scheme changed 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 did features of the Financial Incentives Scheme affect (or not) the extent to which it changed behaviours?	Qual	
Implementation and feasibility How was the Financial Incentives Scheme implemented and is it scalable?	4.1	What was the process for developing and implementing the Financial Incentives Scheme 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 Financial Incentives Scheme?	Qual	
	4.3	How was it implemented?	Qual	
	4.4	What are the facilitators and barriers to scaling and sustaining the Financial Incentives Scheme	Qual	

		(including financial incentives) beyond the pilot?		
	4.5	What was the extent of 'gaming' and data errors?	Quant	Descriptive analysis of in-app data on inconsistencies and outlying behaviour (summary statistics)

Using qualitative methods (outlined in Section 5.4), BIT also explored how acceptability, contextual factors, and barriers/facilitators to the implementation and outcomes affected each research question to ensure that the IPE addressed all core components of the MRC's framework for process evaluations of complex interventions (Moore et al. 2015). The IPE also explored potential unintended consequences, both helpful and unhelpful, of the intervention using qualitative methods. This was done by reviewing the ToC in partnership with HUL, identifying potential unintended consequences by testing its theoretical assumptions and a systematic evaluation of processes, outcomes, and impact. These were used to inform interview questions. We then developed a sampling strategy for qualitative data collection to capture broad perspectives.

5.3 Quantitative methods & analysis

In line with the impact evaluation, one cross-cutting focus of the quantitative IPE was understanding the experience and behaviours of the entire participants sample and of key user subgroups. For subgroup analyses, the same population characteristics and cutoffs were used as those specified for the impact evaluation.

The quantitative IPE made 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 was largely focused on the former; the impact evaluation included a focus on the latter, as part of attrition management (see section 4.1.5).

The quantitative IPE used a combination of descriptive analyses (summary statistics and data visualisation) and regression analyses. See the TP at the link in Appendix A for further information.

5.4 Qualitative methods & analysis

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

5.4.1 Sample

BIT identified six target populations who could provide their perspectives on intervention experience. These perspectives were combined for a comprehensive interrogation of the research questions:

- **"Fully engaged" recipients**: participants who accessed the app at least once every 30 days for the duration of the pilot.
- **"Churn-out" recipients**: participants who signed up to the app but did not use the app for more than 30 days consecutively.
- **Non-recipients**: people who were targeted by HUL, but chose not to participate in the intervention.
- **Delivery partner**: employees of HUL who were involved in the design and delivery of the intervention.
- **Reward partners**: employees from both corporate and local partners who were involved in the design and delivery of the incentives to recipients.
- **City of Wolverhampton Council (CWC):** members of staff at CWC - both in leadership positions and frontline positions - to understand their experiences of supporting the implementation of the Financial Incentives Scheme.

We used purposive sampling to capture the views of a diverse range of people from the target populations listed above. Purposive sampling is a form of non-probability sampling which involves deliberately selecting participants based on particular

characteristics relevant to the research questions (see sampling frames in Appendix F for more detail). Applying this sampling method ensured that the full range of relevant groups were included in the data collection, enabling us to capture a diverse set of perspectives and experiences.

For each population, BIT selected a qualitative methodology that best enabled the IPE research questions to be addressed (see Appendix F for more detail on methodologies). Namely, BIT conducted: (i) interviews and focus groups with recipients (fully engaged and churn-outs) and non-recipients of the intervention; (ii) interviews with the delivery partner (HUL); (iii) interviews with reward partners, and; (iv) focus groups with the City of Wolverhampton Council (CWC). Table 14 provides an overview of the data sources, collection methods and analysis strategy BIT undertook to answer the research topics for each population.

Table 14: 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	30 semi-structured online interviews (60 minutes for intervention recipients; 30 minutes for intervention non-recipients)	12 x Fully engaged* intervention recipients 9 x Churn-out intervention recipients 9 x Non-engaged	Thematic analysis
		3 online focus groups (90 minutes)	1 x Fully engaged intervention recipients *(2 participants - 1 high arm, 1 low arm)* 1 x Churn-out intervention recipients *(3 participants - 1 low arm, 2 medium arm)* 1 x Non-engaged *(3 participants)*	Thematic analysis
4. Implementation and feasibility	Delivery partner	4 semi-structured online interviews (60 minutes)	4 x Delivery partners	Thematic analysis
4. Implementation and feasibility	Reward partners	6 semi-structured online interviews (5 individual, 1 paired; 60 minutes)	5 x National supermarket partners 2 x National gym partners	Thematic analysis
1. Reach 4. Implementation and feasibility	City of Wolverhampton Council	2 online focus groups (90 minutes)	2 x focus group with Local Authority staff *(7 participants)*	Thematic analysis

Fully engaged intervention recipients are recipients **who remained engaged throughout the pilot and did not churn-out at any point. These interviews and the focus group were conducted at the end of the pilot.*

5.4.2 Data collection

Our approach to conducting fieldwork was underpinned by two qualitative strategies: triangulation of findings from different participant groups to assess the qualitative research findings' credibility (Patton 1999), and the use of the Theoretical Domains Framework (TDF; Michie 2015) to ensure a comprehensive assessment of factors affecting the intervention's implementation.

5.4.3 Analysis approach

After qualitative data were collected, interviews and focus groups were transcribed and then analysed. An adapted version of thematic analysis was employed to code the transcripts and identify themes (Braun and Clarke 2006 and Braun and Clarke 2022). These themes underwent a further round of interpretation, classification, and were sorted into high-level themes and sub-themes.

BIT used thematic analysis across the qualitative data to detect themes, patterns and key ideas (see Appendix G for more details). We took a primarily inductive, data-driven approach to analyse patterns and develop themes closely linked to the data, while incorporating elements of deductive analysis, developing frameworks for data charting informed by topic guides. Themes were subsequently mapped onto the TDF. This was complemented by a contextualist method that took into account the individual perspective, as well as the social context, while maintaining focus on the data collected.

Anonymous quotations or written summaries of participants' responses have been included. However, all identifiable information - including, names, roles, etc. - have been removed.

6. Implementation and process evaluation findings

In this section, we first consider the reach of the Financial Incentives Scheme, by addressing several research questions:

- To what extent does the intervention reach participants?
- What were the channels of reach and how did they affect sign-up?
- What are the barriers and facilitators to the intervention's reach?
- What are the characteristics of intervention recipients (those randomised)?
- How was the intervention initially perceived by recipients and non-recipients of the financial incentives?

We then turn to user engagement with the app and associated challenges, as well as points earned and redeemed by participants, by attempting to answer the following:

- To what extent does the intervention engage participants?
- What are the characteristics of those who engage with the intervention for the duration of the pilot, those who partly engage, and those who disengage?
- What were participants' experiences and perspectives of intervention?
- What are the barriers and facilitators to engagement with the intervention?

In each case, we draw on insights gained from both qualitative and quantitative methods within our IPE (Implementation and Process Evaluation) analysis.

6.1 Reach

In total, 34,900 adults in Wolverhampton consented to take part in the study. The qualitative IPE was aimed at understanding how participants heard about the programme, which campaigns and outreach methods were perceived as more effective, and barriers and facilitators to the pilot's reach.

6.1.1 What were the channels of reach and how did they affect sign-up?

The intervention reached participants via a diverse and expansive range of marketing channels, which made the Financial Incentives Scheme highly visible in Wolverhampton – both physically, socially (through word of mouth), and digitally via social media.

How participants heard about the programme

Engaged, non-engaged, and churn-out participants reported hearing about the programme through a **diverse range of marketing and communication channels**, sometimes in combination. Paid social media and word of mouth appeared to be particularly prominent. The marketing and communication channels used included:

- Social media
- Leaflets and letters through the post
- Local and national press
- The Better Health: Rewards shopfront in Wolverhampton city centre
- The Better Health: Rewards branded bus
- The City of Wolverhampton Council (CWC) website
- Information shared via children's schools
- Billboards in Wolverhampton.

Among marketing approaches, the CWC identified that **paid social media** had the biggest impact on downloads. This was reflected in interviews with participants, among whom hearing about the programme through Facebook was widespread.

In addition to social media, **word of mouth** was identified as an important way of hearing about the Financial Incentives Scheme. This occurred through a wide range of sources, including colleagues, friends and family, and fellow attendees of a local

gym. Discussion of the Financial Incentives Scheme was reportedly prompted by direct conversation or by observing peers wearing the Better Health: Rewards trackers.

Reflections on the success of the marketing and communications campaign

The delivery partner perceived marketing for the intervention to be a success, evidenced by the app exceeding its target downloads. CWC staff reflected that the level of marketing used made the Financial Incentives Scheme highly visible in Wolverhampton during the period of recruitment.

> *"The combination of almost all of those things together, you couldn't really get away from it in Wolverhampton [at that] time. It was so immersive, I think, for the community."* - CWC

Indeed, the findings suggested that (a) the diversity of channels and (b) delivering marketing over a sustained period of time was important in driving engagement. Participants reported engaging with the programme after **encountering it repeatedly** through multiple channels. Participants noted becoming aware of the Financial Incentives Scheme via one channel, not engaging with the information at the time, then re-encountering it through a further channel, which prompted them to look into the programme further.

> *"I had actually seen some advertising, but I hadn't really thought about it, and then when my mum told me about it, and then I saw it, and then I put the two together"* - Churn-out user, Low reward arm

While marketing was perceived to be a success, some participants suggested that the reach of the Financial Incentives Scheme could have been increased by targeting local hospitals and university campuses, which they felt were overlooked in the marketing campaign. It is important to note that these settings were included in intervention marketing, indicating that the reach of messaging across these settings was less successful.

6.1.2 What are the barriers and facilitators to the intervention's reach?

There were no barriers to reach identified by those interviewed, likely because the majority of interview respondents were aware of the intervention. Facilitators included campaign-level facilitators, such as the extensive and diverse marketing effort, which employed consistent and professional design; and individual-level factors, such as desire to improve health, the notion that participants would have 'nothing to lose' in joining the Financial Incentives Scheme, and the novelty of the scheme.

Campaign-level facilitators

The **extensive and diverse marketing effort** applied by DHSC and CWC and the delivery partners was a notable facilitator of the Financial Incentives Scheme's reach. This can be mapped onto the 'knowledge' component of the TDF, as marketing efforts amplified potential participants' awareness and information about the intervention. CWC highlighted that the level of effort put into marketing and recruitment exceeded expectations and normal practice, which might be challenging to replicate.

This effort was enhanced by the value of the **local insight** obtained through meaningful and sustained partnership with local delivery partners, which ensured that messaging was appropriate for, and resonated with, local audiences. This maps onto the knowledge component of the TDF, as more in-depth awareness of the intervention setting allowed greater targeting, as well as the component of social identity. For example, marketing used local accents in any videos about the intervention. Delivery partners felt that marketing materials reflected the diversity of the area, ensuring that the intervention was perceived as inclusive.

Consistent and professional design of these marketing materials was also noted by delivery partners as a facilitator of reach. Effective design of marketing materials promotes attention from potential participants, fostering the TDF component of memory, attention and decision processes.

Individual-level facilitators

On an individual level, participants noted that the Financial Incentives Scheme's **focus on improving health** resonated with them, linking to TDF components of intentions and goals. This particularly motivated participants wanting to improve their health, which in some cases was linked to changes, such as recent diagnoses.

As the intervention was free and easy to use, participants highlighted that they would have **'nothing to lose'** in trying the intervention; this notion was widespread across user groups.

> *"You've not made it too difficult. You're giving people the tools. Added bonus for those who get motivated by this, as well as the incentive. For me, it was like, why would you not?"* - Churn-out user (focus group)

More exceptionally, some participants noted that the concept of the intervention was **novel** and exciting. This caught their attention and encouraged them to seek more information, promoting the TDF component of memory, attention and decision processes.

While some population groups were under-represented among those who consented to participate in the pilot, no barriers to reach were identified by the delivery partner or participants in the qualitative work. However, we must note the likely selection bias in those agreeing to be interviewed. Additionally, while these populations may have been reached by intervention marketing and messaging, they may have chosen not to participate due to a variety of barriers to initial engagement, discussed in Section 6.2.4.

6.1.3 How was the intervention initially perceived by recipients and non-recipients of the financial incentives?

Participants felt motivated by the incentives offered by the Financial Incentives Scheme. There was some indication that these incentives enhanced the appeal of the intervention for those in more deprived communities, who are typically harder to reach.

Initial impressions of the intervention were largely positive among engagers, churn-outs, and non-engagers. Even among those who chose not to engage, the intervention was seen as a 'good thing' with the potential to be useful.

One delivery partner reflected that the success of the intervention's reach exceeded expectations, particularly among typically underserved communities. They suggested that this was largely due to the **reward offering**, mapping to the TDF component of reinforcement.

> *"We were very amazed that we managed to reach the people that we wanted to reach so well. I think the targets in terms of deprivation index, age, baseline BMI and diets were mostly met, I think. So, it was good to see even the hard-to-reach people in some areas of Wolverhampton taking part [...] I think that's the reward part. My gut feeling is if we had just gone with the health and fitness app, we might not have engaged them that much"* - Delivery partner

This is reflected in participants' perception of **the appeal of what the intervention was offering**, particularly the offer of a free wearable fitness tracker and the rewards available. Participants from areas of greater deprivation in particular noted that the cost of purchasing a wearable fitness tracker would have been prohibitive. One fully-engaged male user from an area of higher deprivation reflected that the wearable fitness tracker *"was one of the reasons I actually signed up for it because I'd always wanted a tracker watch like that."* This may explain higher levels of attrition among these groups if the wearable fitness tracker was a greater motivation for engagement than participation in the intervention.

6.2 Engagement

6.2.1 To what extent does the intervention engage participants?

Overall, the evidence suggested that the app was engaging to participants. When measuring engagement in the app, we looked at both "very active" and "active"

users.

- **Engagement with the app:** This was measured by the number of participants who were "very active" and opened the app at least once per every seven days. Engagement fell from 100% in the first week after randomisation, to 44% by the end of week 22, and to 32% by the end of the pilot (week 24).
- **Engagement with challenges:** Throughout the pilot, participants still active at M5 (i.e. who had opened the app at least once in the previous 30 days) selected 161,402 challenges across PA and diet. Of these, 82% were completed (finished, as not abandoned). The challenges selected more often were the two PA challenges. However, given that more diet challenges were available to participants than PA ones, overall, diet challenges were selected more than twice as often as PA challenges. This pattern is mirrored in the number of times PA and diet challenges were completed.
- **Engagement with points and rewards:** On average, participants active at M5 (i.e. who had opened the app at least once in the previous 30 days) earned 63 points per week from challenges (of a different pound (£) value depending on trial arm), and £54.80 overall for providing their data. In total, between week 1 and October 13th (when the pilot ended) they redeemed 31,517 rewards for a total value of £524,953.7 (using points earned via challenges and the monetary amount provided in exchange for users providing data). This is equivalent to £71.01 per user.

As indicated in the ToC, beyond the initial engagement with the intervention (e.g. providing in-app consent and successfully completing onboarding), continued engagement with the app is critical to driving behavioural change in PA and diet activity. This includes logging behaviours, selecting and completing challenges, and redeeming rewards.

6.2.1.1 Engagement with the app

According to the TP, we defined engagement based on three definitions:

- Very active users: participants who have opened (logged in) the app at least once in the previous 7 days;
- Active users: participants who have opened (logged in) the app at least once in the previous 30 days;
- Churned out participants: participants are not active, i.e. have not opened (logged in) the app at least once in the previous 30 days.

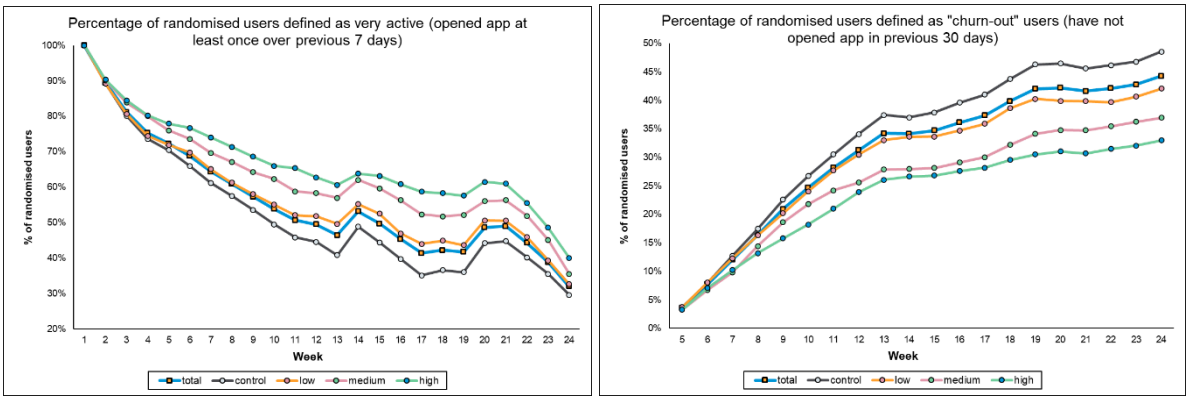
Login data showed that engagement with the app itself fell relatively consistently over the 24 weeks following randomisation. 61% of participants were classed as 'very active users' over the 30 days after randomisation. Figure 15, left panel below shows the proportion of the full randomised sample classed as 'very active users' (i.e. who had opened the app at least once in the previous seven days) fell from 100% in week 1 to 44% by week 22, and to 32% by week 24.

In comparison, when we consider the proportion of the randomised sample classed simply as 'active users' (i.e. who had opened the app at least once in the previous 30 days, which is the inverse of 'churn-out'), the decline is understandably less dramatic. The proportion of users meeting that classification fell from 100% in weeks 1-4, to 58% by week 22 and to 56% by week 24.

PA and Diet challenges could be completed up to the end of week 21, while points from these challenges were awarded up to the end of week 22, which was the M5 data collection point. From week 23 and up to the closure of the app on October 13th, participants could still use the app to sync data and access the online store to convert points earned into redeemable rewards and spend the money reimbursed for providing data.

In week 14 and week 20 we saw a small uplift in the proportion of these users. This is likely due to the fact that these weeks participants received notifications to fill in the Intake24 survey (for which they were reimbursed £10 x 2 for providing data at both the M3 and M5 data collection points). When looking at not active and churn-out users (i.e. who have not opened the app in the previous 30 days) per week across the pilot, it is evident how the proportion of these participants increased in time, with a steeper slope ('faster churn out') in the first period of the trial (Figure 15, right panel).

Figure 15: Left panel: The percentage of randomised adults that are very active users (defined as app opened at least once in the previous 7 days) per week from randomisation to week 24, split by intervention arm. Right panel: The percentage of randomised adults that are churn-out users (defined as app not opened in the previous 30 days) from randomisation to week 24, split by intervention arm



Demographic composition of cohort

Table 15 compares the target sample, the sample at randomisation and the sample still active at M5 (not churned out, i.e. participants who opened the app at least once in weeks 19 - 22, which are the last 30 days of the pilot). The RAG (Red, Amber, Green) colour coding of the rightmost columns offers a quick visual guide for whether a subgroup within a category at M5 was either within 2 percentage points of the stated target proportion (green), between 3 and 5 percentage points off the stated target (amber), or 'missed' the target by more than 5 percentage points (red). In effect, those cells shaded red represent the subgroups with disproportionately high or low representation in the sample still active at M5, relative to stated targets.

As with the full cohort that was randomised and started the pilot (n = 12,767), the composition of subgroup categories among those still active at M5 (n = 7,387) was in most cases quite close to what was targeted in advance of the study. For ethnicity, deprivation level, and age, the cohort still active at M5 was remarkably close to possessing the composition of these demographics stated as a target before the pilot began. This was less the case when considering the balance across trial arms, or

when considering gender, given the fact that the sample was unbalanced across female and male participants from registration onwards.

Table 15: The composition of subgroups for participants randomised, participants active at M5, and those providing valid PA and diet data at M5; and the targeted composition of app users as stated in advance of the study

Subgroup	**Targeted composition of subgroups (% in each category)**	**Composition of those randomised and consented to data sharing (% in each category)** ***[n = 12,767]***	**Composition of those active at M5 (% in each category)** ***[n = 7,387]***
Ethnicity category	**Target (%)**	**Randomised (%)**	**Active at M5 (%)**
White	68	68	67
Asian or Asian British	18	18	18
Black, African, Caribbean, or Black British	7	7	7
Mixed or Multiple Ethnic Backgrounds	5	4	3
Gender category	**Target (%)**	**Randomised (%)**	**Active at M5 (%)**
Female	50	67	68
Male	50	33	32
Deprivation (IMD) category	**Target (%)**	**Randomised (%)**	**Active at M5 (%)**
Most deprived (scoring 1 or 2 on the IMD)	51	49	48
Other (scoring above 2 on the IMD)	49	49	51
Age category	**Target (%)**	**Randomised (%)**	**Active at M5 (%)**

Aged 45 or over	45	42	46
Aged less than 45	55	58	54
Study arm	**Target (%)**	**Randomised (%)**	**Active at M5 (%)**
Control	50	61	57
Low incentive	17	15	16
Medium incentive	17	12	13
High incentive	17	12	14

Notes: Percentages may not sum to 100% due to rounding and some categories not being listed

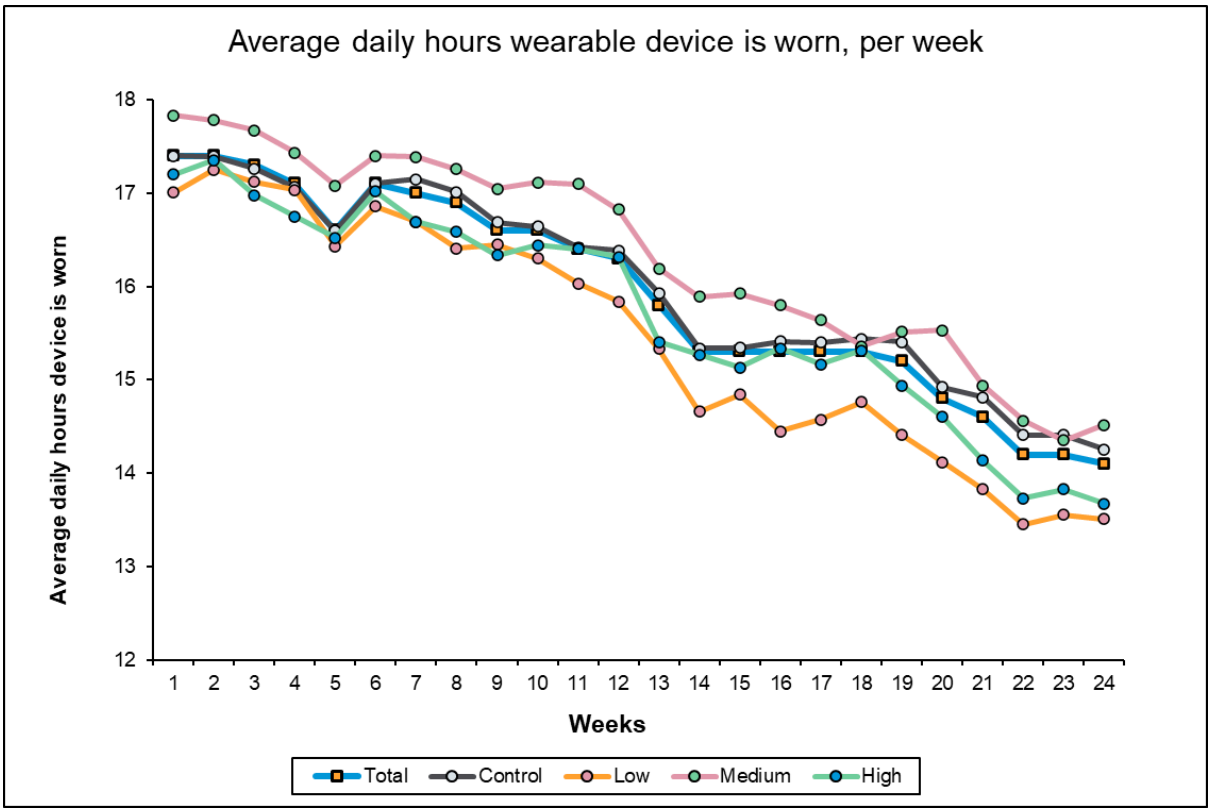
Wearable fitness tracker usage

Another important way to measure engagement with the intervention was to consider participants' use of the wearable fitness trackers. Figure 16 shows the average number of hours that the wearable fitness trackers are worn among the sample of participants who were still active at M5, on days for which participants synced activity data. It shows this by week and study arm. Overall, wearable fitness trackers were worn an average of 17.4 hours per day in the first two weeks following randomisation, with this figure falling relatively consistently to 14.1 hours per day by week 24. It is worth noting also that the number of weekly wearers fell, again quite consistently, from over 7,000 participants in week 1 to fewer than 2,400 in week 24.

While some interview participants did not mention concerns with the wearable fitness tracker, those who did reported **discomfort** and **skin irritation**, especially when sleeping. Participants also discussed **forgetting** to **wear** or **charge** the wearable, as well as software issues such as syncing the wearable fitness tracker with the app, and challenges recording exercise through the wearable.

> *"The material it was made out of was unnecessarily harsh. It kept irritating my wrist."* - Churn-out participant, control arm

Figure 16: Average daily hours wearable tracking device is worn, by intervention arm and per week, from randomisation to week 24



Factors predicting engagement with the app - a regression approach

It is worth emphasising that the analysis we have outlined thus far is purely descriptive, and cannot control for the fact that some of the characteristics discussed may be correlated (for example, it is possible that participants of some ethnic groups may have systematically higher levels of deprivation than others). Therefore in aiming for a greater level of analytical robustness, and in line with the TP, we also conducted statistical regression analysis to identify the factors associated with app users being very active (defined as app installed and opened at least once over the course of the last 7 days) for a larger proportion of the pilot.

Our analysis found that of those still active at M5, those who are older (0.42%, 95% CI: 0.39 - 0.46), and females (2.23%, 95% CI: 1.20 - 3.26) were very active for a (slightly) larger proportion of the trial, while participants with white ethnicity were slightly less

active (-2.41%, CI: -3.69 - - 1.13). There was no correlation between app usage and education or household size. Conversely, there was a statistically significant (but very small) positive correlation between app usage and both baseline MVPA and baseline fruit and vegetable consumption: those with higher levels of MVPA and higher levels of fruit and vegetable consumption at baseline were more likely to be active users for longer. Last, participants who used the Better Health: Rewards tracker were active for a larger proportion of the pilot than those who used other wearable fitness trackers. A full regression output is available in Appendix J, Tables J5 and J6.

6.2.1.2 Engagement with challenges

Another crucial facet of engagement with the study was the number of challenges that participants both selected and completed to earn points. Participants could select and complete PA or diet challenges, for a maximum of two challenges per week, from week 1 to week 21 of the pilot. Once participants had completed their first week of challenges participants could select challenges at either the "maintain" level, or a "harder" level (which was more challenging to complete, but which offered greater rewards to the user).

PA vs diet challenges

Throughout the pilot , participants still active at M5 selected 161,402 challenges across PA and diet. Of these, 132,131 (82%) were completed (finished), To note, started challenges could have one of four mutually exclusive 'status' outcomes by the end of the trial: finished, abandoned, rejected for invalid photos, expired. We define a challenge as 'completed' if its status is 'finished'.

Table 16 below provides an overview of the number of times each challenge was selected by participants still active at M5. . The challenges selected more often were the two PA challenges ('Let's get moving' and 'Step it up!'). However, given that more diet challenges were available to participants than PA ones, overall, diet challenges were selected more than twice as often as PA challenges by participants still active at M5 (total diet challenges selected: 110,686; total PA challenges selected: 50,716). This pattern was mirrored in the number of times PA and diet challenges were completed (total diet challenges completed: 90,673; total PA challenges completed: 41,458).

Table 16: Comparison of all challenges in terms of number of times selected and completed, sorted by number of times a challenge was selected

Challenge name	Type	Total number of times challenge selected	% of total selected	Total number of times challenge completed	% of total completed
Let's get moving	PA	26,716	16.54	21,759	16.47
Step it up!	PA	24,000	14.85	19,699	14.91
Boost Fruit And Veggie Intakes	Diet	22,192	13.74	17,670	13.37
Switch Up Your Drinks	Diet	16,787	10.39	14,189	10.74
Snack Attack	Diet	15,565	9.63	12,566	9.51
Go Lower Fat	Diet	13,156	8.14	10,784	8.16
Fibre Fix	Diet	9,711	6.01	7,904	5.98
Brilliant Breakfast	Diet	7,781	4.82	6,082	4.60
Dessert Swapper	Diet	6,957	4.31	6,131	4.64
Say No To Fried Food	Diet	6,788	4.20	5,731	4.34
Eat Leaner Meat	Diet	6,588	4.08	5,207	3.94
Have Healthier Food On The Go	Diet	5,161	3.19	4,409	3.34

6.2.1.3 Engagement with PA challenges

Challenges selected

As described above, participants could pick either the 'Step it up!' or 'Let's get moving' physical activity challenge. On average, those who were active at M5

selected 0.33 PA challenges per week during the weeks in which they could select challenges, from week 1 to week 21 inclusive.

Challenges completed (not abandoned)

On average, those who were active at M5 completed 0.27 PA challenges per week during the weeks of the trial in which they could select challenges. The left panel of Figure 17 below displays the pattern of challenge completion by pilot week, where the week denotes when the challenge was selected.

Table J1 in the appendix provides a breakdown by subgroup. There are some small differences in the number of challenges completed for some demographic groups. The biggest difference across demographic groups was observed by age. Older participants (aged more than 41) completed on average 0.30 PA challenges per week, whilst younger participants completed 0.23 PA challenges per week.

Challenge difficulty

When disaggregating for challenge level (maintain or harder), those who were active at M5 completed 0.07 PA challenges of level maintain per week, and 0.2 challenges of level harder during the weeks of the trial that they could select challenges.

6.2.1.4 Engagement with diet challenges

Challenges selected

Similarly, when we considered the diet challenges available to participants, ten separate challenges related to diet could be selected and completed to earn points during the pilot. On average, participants who were active at M5 selected 0.71 diet challenges per week during the weeks in which they could select challenges.

Challenges completed (not abandoned)

On average, those who were active at M5 completed 0.58 diet challenges per week during the weeks of the trial in which they could select challenges. The right panel of Figure 17 below displays the pattern of challenge completion by pilot week, where the week denotes when the challenge was selected.

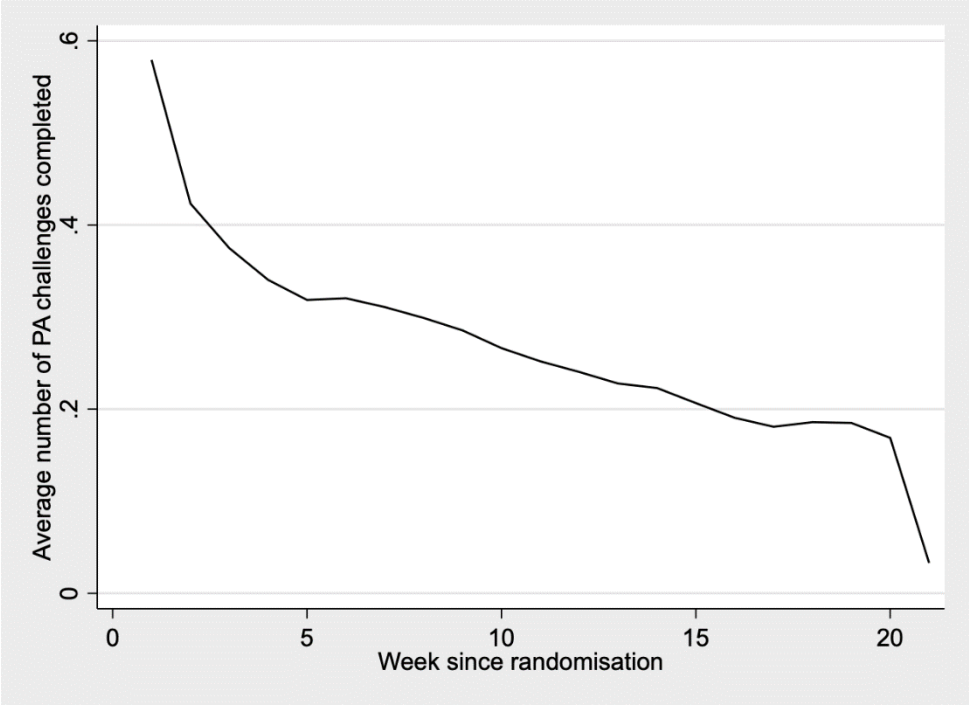
Diet challenges showed a similar pattern of completion across subgroups as PA challenges, but with slightly more pronounced(more detail in Table J1 in the Appendix). In terms of average number of diet challenges completed per week among participants active at M5:

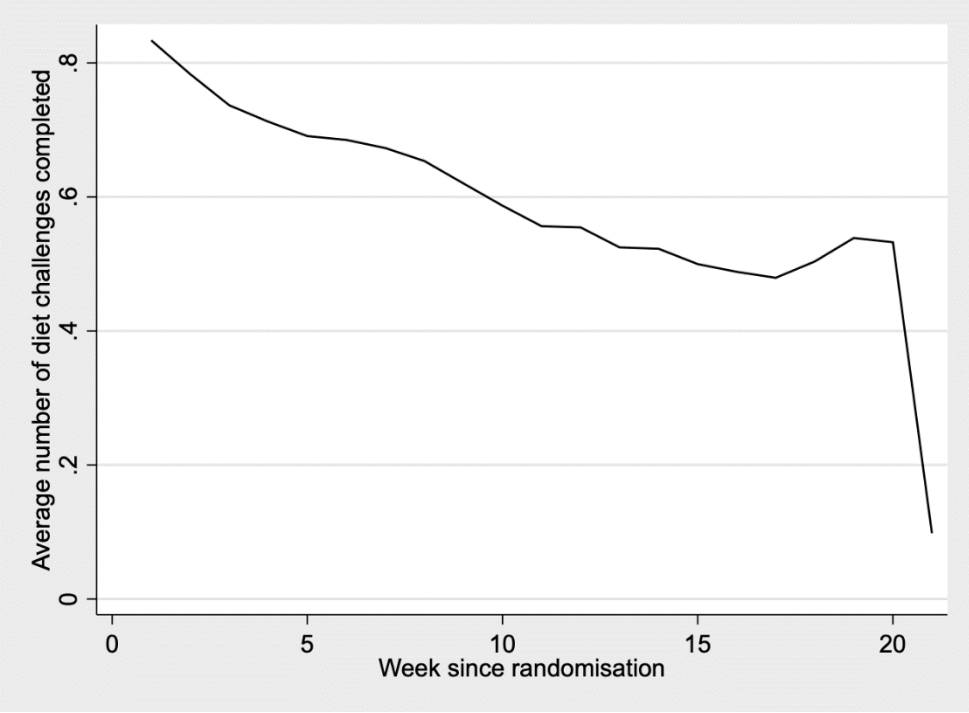
- Older participants (Over 41 years) completed fewer diet challenges per week than younger participants (0.63 vs 0.53)
- Participants with healthier habits at baseline completed more diet challenges per week than participants with less healthy habits (high baseline fruit and vegetable consumption: 0.67 diet challenges completed; low baseline fruit and vegetable consumption: 0.56; high baseline MVPA: 0.71; low baseline MVPA: 0.52)
- Participants of White ethnicity completed more diet challenges (0.61) than the other groups, especially participants of Black, African, Caribbean, or Black British ethnicity (0.47).

Challenge difficulty

When disaggregating by challenge level (maintain or harder), participants who were active at M5 completed 0.13 'maintain' level diet challenges per week, and 0.45 challenges of 'harder' level challenges during the weeks of the pilot in which they could select challenges.

Figure 17: Left: The weekly number of PA challenges successfully completed, by week (in N). Right: The weekly number of Diet challenges successfully completed, by week (in N)





6.2.1.5 Points and rewards

Points, rewards, and value of the rewards

Participants were rewarded with points for completing challenges. Participants were also reimbursed with a monetary sum (not in the form of points) for providing data (e.g. filling in Intake24 survey questionnaires and syncing the wearable fitness tracker). Participants in the control arm could use the monetary sum reimbursed for data collection to redeem rewards. Participants in the reward arm could use the monetary sum reimbursed for providing data and points earned through challenges to redeem rewards, where points had a conversion rate to pounds (£) depending on the study arm they were in, as described in Section 2.4.

The total number of points earned through challenges by those still active at M5 was 10,729,498 points, equal to 63 points per week on average per user (week 1 to week 21 inclusive). The total points earned by participants still active at M5 is equivalent to a monetary value of £142,489.54, this is equal to an average of £19.29 per participant.

Similarly, throughout the pilot participants still active at M5 were reimbursed a total of £404,825 for syncing their wearable fitness tracker and filling in the Intake24 diet questionnaire, equal to £54.80 per person throughout the pilot (minimum reimbursement to a user active at M5 was £5 and maximum was £70). More information on how users could be reimbursed for data collection is available in Appendix K.

Overall, participants still active at M5 had a total of £547,314.54 (from both points and reimbursements for data provision) that they could redeem for rewards. This is equivalent to an average of £74.09 per participant.

Between week 1 and October 13th, participants active at M5 redeemed 31,517 rewards for a total value of £524,953.70 (using points earned via challenges and the monetary amount reimbursed for providing data). This is equivalent to £71.06 per user.

The vast majority (99.5%) of rewards selected were gift cards, and only 0.5% were merchandise. Table 17 below shows how often each gift card category was selected.

Table 17: Number of times each gift category was selected by participants active at M5

Reward Category	N times selected	%
Food & Drink	25,804	82%
Retail	4,436	14%
Entertainment	527	2%
Home & Personal Services	383	1%
Travel & Transportation	78	>1%
Gifts	52	>1%
Spa & Wellness	5	>1%
Tourism	2	>1%
Spa & Pampering	1	>1%

Factors predicting the cumulative number of points earned - a regression approach

Certain groups accumulated more points than others: of those who were active at M5, participants who were older (an increase of 11 points per year they are older, 95% CI: 8.9 - 13.1) and female (+127, 95% CI: 67 - 187) cumulatively earned more points.

Further, those with a third-level degree earned more points than those without, and those of white ethnicity earned more than those of Asian or Black ethnicity.

Note that these groups (except white participants) were also more likely to be very active for longer (see section 'Factors predicting engagement with the app - a regression approach'). The pre-specified analysis presented here does not control for participants' levels of activity within the app, so we are unable to disentangle whether this is because these groups were logging in more than other participants or

if they selected more challenges or if they were more likely to earn points from challenges. However, the evidence suggests that these factors may be at play for some participants but not for others. For example, as detailed in the 'Engagement with PA challenges section' and 'Engagement with Diet challenges' sections, older participants were more likely to complete challenges. On the other hand, White participants seem to be very active for a slightly shorter period of time within the app - suggesting that they may earn more points with less engagement.

Participants in households with multiple participants earned more points than those who were the only user in their household, which *may* suggest that engaging with such PA and diet challenges alongside other people could help participants to complete them. Participants who used the Better Health: Rewards tracker earned more points during the pilot than those who used other wearable fitness trackers, holding all else equal. Lastly, there is a statistically significant (but very small) positive correlation between points earned and both baseline MVPA and baseline fruit and vegetable consumption. A full regression output is available in Tables J3 and J4 in Appendix J.

6.2.2 To what extent do incentives affect in-app engagement?

Overall, we saw participants in higher reward arms engage more with the app across all the metrics analysed (e.g. app log-ins, PA and diet challenges selected and completed, points earned per week, money reimbursed for providing data, number of rewards redeemed, total value of redeemed rewards), except wearable usage.

The regression analysis found that participants in the high arm who had similar demographics to participants in other study arms earned more points: participants in the high arm earned 250 more points than similar participants in the medium reward arm and 768 more points than similar participants in the low reward arm.

This section describes how user behaviour in terms of app log-ins, challenges selected and completed, points gained and rewards redeemed varies by pilot trial arm.

6.2.2.1 App log-ins by incentive levels

Those in the 'high' incentive study arm were consistently more engaged (as measured by app log-in and wearable fitness tracker usage) than those in the 'medium' arm, followed by the 'low' and 'control' trial arms respectively. This is according to a variety of metrics:

- Long term engagement: the 'high' incentive study arm saw a higher proportion of very active users than the 'medium' arm, followed by the 'low' and 'control' trial arms respectively. This is graphed in Figure 5, left panel. In addition, we note that such a relationship is visible in the first month of trial. 61% of participants opened the app at least once every 7 days over the 30 days after randomisation. This can be disaggregated into the control (59%), low incentive (60%), medium incentive (66%), and high incentive (70%) trial arms respectively. This demonstrates that differences in engagement between trial arms arose from the beginning, implying that the incentive structures had an immediate effect on app usership.The regression analysis presented in Table J3 corroborates these descriptive findings with participants remaining very active for a higher number of weeks in higher-paying trial arms. Those who were in the low, medium, and high incentive trial arms were very active for a larger proportion of the intervention, respectively being very active for 3.12% (95% CI: 1.78% - 4.45%), 7.99%% (95% CI: 6.55% - 9.43%), and 10.0% (95% CI: 8.62% - 11.39%) more weeks than those in the control arm, holding all else equal.

- Churn out: These same patterns emerged when looking at the level of churn among participants, as graphed in Figure 13, right panel above.

- Wearable fitness tracker usage: There was a less obvious relationship between incentive structure and wearable usage (for days that people submitted activity data) over the course of the pilot, as graphed above in Figure 12. Those in the medium reward arm had a higher average level of 'daily hours worn' than those in the other pilot trial arms in most weeks.

6.2.2.2 Challenges by incentive levels

Table 18 below provides the average number of challenges selected and completed (separately for PA and diet) and the average number of challenges completed by difficulty, by pilot trial arm (participants active at M5).

A similar picture emerged across PA and diet. A lower number of PA and diet challenges per week were selected and completed in the control arm, followed by low, medium and high incentives in order. This pattern seems to be stronger for 'harder' challenges.

Table 18: Average number of challenges selected, completed, completed by maintain/harder level of difficulty, for PA and diet challenges per week by intervention arm (sample active at M5)

	Control	Low incentives	Medium incentives	High incentives
PA selected	0.30	0.35	0.35	0.38
PA completed	0.24	0.29	0.30	0.33
PA completed, maintain	0.07	0.07	0.07	0.06
PA completed, harder	0.17	0.22	0.23	0.27
Diet selected	0.57	0.74	0.96	1.02
Diet completed	0.44	0.61	0.84	0.91
Diet completed, maintain	0.11	0.14	0.18	0.19
Diet completed, harder	0.33	0.47	0.66	0.72

To see how these patterns evolve in time, see Appendix J, Figure J1. Comparing the left and right panel (average number PA challenges and diet challenges completed by week by pilot trial arm for the sample active at M5, respectively), we can notice two patterns: (a) PA: participants in all trial arms completed fewer PA challenges per week as more time passed by; (b) Diet: participants in the control arms completed

more or less the same number of diet challenges per week as time passed by, whilst participants in pooled reward arms completed more challenges per week as time passed by, until the final weeks of the trial when all trial arms saw a decrease in the number of diet challenges per week completed.

6.2.2.3 Points and rewards by incentive levels

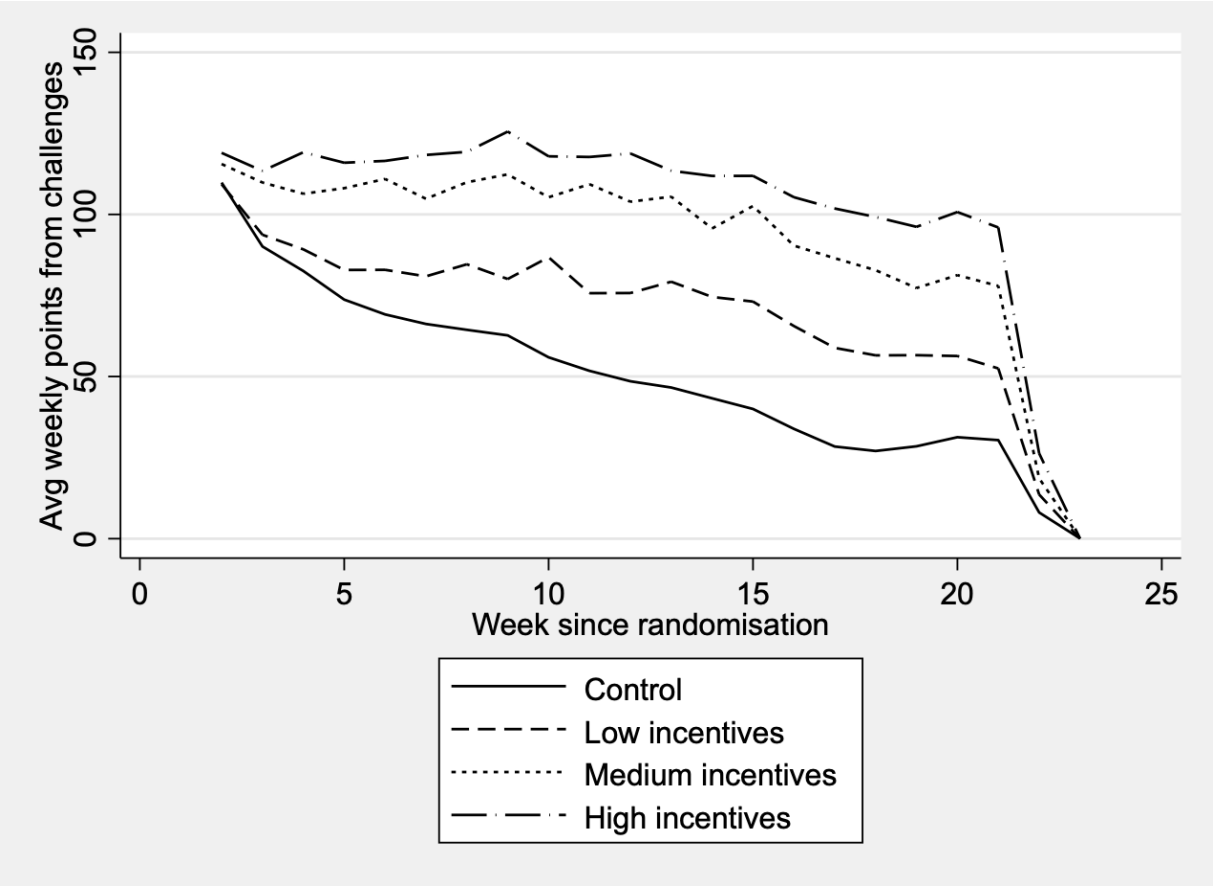
Table 19 below shows earned points and their monetary value per user, by arm (participants still active at M5). For the totals per arm, see Figure J2 in Appendix J. The higher the level of the incentive, the higher the average number of points per week earned by a user. This is observed both in a descriptive, aggregated analysis (Table 16 above) and in the more sophisticated regression analysis. The regression analysis found that participants in the high reward arm earned 250 more points compared to similar participants in the medium reward arm and 768 more points compared to similar participants in the low reward arm. The differences between all trial arms are statistically significant. The full regression output can be found in Table J3.

It is worth noting that although all participants could earn the same monetary value for providing data, more participants in the high reward arm chose to complete data collection activities. However, the statistical relationship between incentive level (i.e. study arm) and monetary reimbursement for providing data was much less strong than the relationship observed between incentive level and the average number of points earned per week. Across pilot arms, participants in higher reward arms tended to redeem more rewards, and for a higher total value.

Table 19: Statistics for points and rewards per person, by intervention arm (sample active at M5)

	Control	Low incentives	Medium incentives	High incentives
Average amount of points per user per week	48	66	88	98
Total pounds (£) for providing data per user on average Note that all participants were rewarded with the same amount for providing data, regardless of the pilot trial arm	£53.74	£54.95	£56.51	£57.25
Total N of rewards redeemed p user on average	3.95	3.41	5.00	5.77
Total value of redeemed rewards per user on average Note that the monetary value of points gained via challenges differed across pilot trial arms.	£51.87	£58.28	£102.30	£131.82

Figure 18: The weekly number of points earned from randomisation to week 21, by week and split by intervention arm



Notes: The picture excludes week 1, in which very few points were awarded.

6.2.3 What were participants' experiences and perspectives of the intervention?

Participants engaged with the app in a variety of ways. These can be summarised across four idealised 'persona' types: competitive participants, participants who learned a lot, participants who were along for the ride, and participants who have made repeated attempts to improve their health.

In what ways did participants engage with the app after signing up?

Participants engaged with the app in a variety of ways, based on differing motivations and varying baseline health and fitness knowledge. We have collated these experiences into four composite 'personas', drawing on insights from a combination of participants.

Personas were developed inductively from qualitative data, but do not map onto quantitative IPE data or equivalent numerical evidence, or to specific individual interview participants. This differs from a qualitative typology analysis in that our personas are illustrative and representative, rather than exhaustive.

Competitive

- Competitive participants are primarily driven by achieving targets and completing challenges
- They experience intrinsic motivation to achieve the app's targets in addition to incentivisation through rewards
- They would describe themselves as 'competitive' or 'determined'
- They have a moderate-high baseline level of health and health knowledge
- They select challenges which offer the highest number of points, or which represent the greatest personal challenge

"It wasn't so much the points, it was more the fact that it just set me a target. It made me do it, like a challenge for myself." - Fully engaged user, 'Low' reward arm

"My wife and I would get into a competition as to who could earn the most points and who would do the most steps." - Fully engaged user, 'Medium' reward arm

"I like a challenge. Once I start something like this, I'm going to see it through to the bitter end. That's the kind of personality I am." - Fully engaged user, 'Low' reward arm

Learning a lot

- Learning a lot participants are new to the idea of increasing physical activity and healthy eating and were incentivised to join the Financial Incentives Scheme by the prospect of financial incentives
- They have learned a lot from the challenges and the health information, and applied this to change their lifestyle
- They are sometimes concerned about not being able to achieve challenges, and may select challenges that they thinks are feasible for them

"It's been a revelation. It's been fantastic for me" - Fully engaged user, 'Low' reward arm

"I've literally changed the whole of my diet [...] I'm more conscious of the protein content, the fat content, the carbs content, the content or the nutritional side of the food now. Rather than, oh, this looks nice or this tastes nice" - Fully engaged user, Control arm

"It wasn't so much the points, it was like, right, what can I manage to change? What's going to be easy for me to do, because if I start a challenge, I don't want to fail it." - Fully engaged user, Control arm

Along for the ride

- Along for the ride participants engage with the app by continuing their existing activities to do the minimum required to gain rewards, without making any notable changes to their lifestyle
- They choose challenges based on what is easy, given their existing diet and level of physical activity
- They typically have a moderately high baseline level of physical activity and healthy eating, although may not be explicitly aware of this - they are not a

'fitness fanatic'

- If they have a lower level of baseline physical activity and healthy eating, the challenges may still prompt some improvement in diet and level of exercise

"*As I've proven, you can just get it and then not really follow through*" - Fully engaged user, 'Low' reward arm

"*I was already exceeding what it wanted me to do anyway*" - Fully engaged user, 'Medium' reward arm

"*After a while of doing it, it was like I was just doing it because of the rewards. I lost the real focus. [...] I didn't care about my eating habits*" - Churn-out user, 'Medium' reward arm

Repeated attempts to get healthy

- Participants with repeated attempts to get healthy have tried many times to improve their health, but without lasting success; they see this programme as a new approach that might be an easy and effective solution
- They may disengage with the programme if they do not see immediate change

"*Anytime you see something new, you think, oh, could this be the one? Could this be slightly different? Could this be the thing that actually cracks it and works?*" - Fully engaged user, 'Low' reward arm

"*It waned for me. Once I'd done a couple of the three things, or the three tasks, then that was it, really, to be honest, once I'd done that, and then I probably reverted back to my app on my phone, the normal Google Fit.*" - Fully engaged user, 'Low' reward arm

"*I have looked into this many, many times, different variety of ways of doing it through gym, through PT, instructors, through different apps. Do you know what I*

mean? By now, I do know what's good for me and what's not." - Fully engaged user, 'Low' reward arm

6.2.4 What are the barriers and facilitators to engagement with the intervention?

Key barriers to initial engagement with the intervention identified from qualitative work were: concern about the degree of challenge, time requirement, or complexity of engaging with the intervention; and lack of awareness of the benefits available. Key facilitators of initial engagement were: the offer of rewards and free technology; desire to improve health; and a simple sign-up process.

Key barriers to sustained engagement identified through qualitative work were: Better Health: Rewards Tracker technical issues; forgetting to charge, wear, or sync the wearable fitness tracker; constraints on time or cognitive bandwidth, representing excess mental load; and inconvenience accessing technical support. A key facilitator identified was the simplicity of the app.

Our analysis distinguishes *barriers* (factors that contributed to potential participants not engaging or churning out) from *obstacles* (initial concerns raised that were subsequently overcome and therefore did not lead to lack of engagement/churn-out).

A summary of obstacles can be found below, and further discussion of obstacles can be found in Appendix I.

Obstacles to initial engagement:

- Concerns about legitimacy of the Financial Incentives Scheme
- Concerns that the intervention would be disappointing

Obstacles to sustained engagement:

- Complexity of reward redemption
- Issues with the wearable fitness tracker recording certain activities as exercise
- Confusion among control group participants about points

6.2.4.1 Barriers to initial engagement

We identified key barriers to initial engagement, as follows:

- Concern about what the intervention would entail
- Misunderstanding of the intervention
- Negative feedback

Barriers to initial engagement resulted in potential participants not engaging with the programme. It is important to note that qualitative work was not conducted with participants who disengaged with the programme during the baseline period, hence barriers to engagement specific to this period may be understated.

The **degree of challenge** was cited as a barrier to engagement by non-engagers, mapping to the TDF component of memory, attention, and decision processes. Concerns were reported that the intervention would set targets that might put pressure on them and that failing to achieve these targets might be discouraging and demotivating. This was also flagged by some fully engaged participants, but counteracted by the sentiment that they would try their best.

> *"I'm not sure I'm ready to sign up for anything that will put pressure on me every day to fulfil targets. I like targets, but when I think of health and fitness, I don't ever want to feel pressured trying to do it."* - Non-engager

Non-engagers, particularly those with busy lives such as parents, expressed concern that the intervention would be too **time-consuming** to engage with. Relatedly, concerns were raised that the app would be **complex** and technically challenging to use. This assumption was sometimes reinforced by reports from peers that they had struggled to use the app, raising questions about whether it was worth the effort.

> *"Working full-time, having a child and stuff, I was just like, no, I'm too busy to even consider it."* - Non-engager

Misunderstanding of the intervention was a widespread barrier reported by non-engagers. Some non-engagers, particularly those who already engaged in healthy behaviours, **could not see how the intervention was different or superior** to existing wearable fitness trackers or health promotion apps that they had engaged with. Moreover, a lack of awareness **of the unique benefits** of the intervention, namely the financial incentives on offer, was also identified as a barrier. Some non-engagers with mental health conditions or other physical disabilities reported concerns that the programme **may not be accessible** to them. These concerns were linked both to their physical ability to complete challenges, and the motivation to be able to do so.

Further **mistaken assumptions** about the intervention were raised by some non-engagers, indicating that **unclear marketing** represented a barrier to engagement. For example, one non-engager reflected that they would have liked the ability to pick challenges appropriate to their requirements and/or activity level. However, it *was* possible for participants to select challenges based on their requirements. These barriers arising from misunderstanding of the intervention can be considered within the TDF component of knowledge, arising from the absence of accurate knowledge.

Negative feedback about the intervention, both from peers and from reviews and ratings of the app, was a less widespread, but nonetheless present, barrier to initial engagement. This relates to the TDF component of social influences.

> *"When I want to download an app, if I'm seeing it has 3.2 stars and a lot of negative reviews, I probably won't bother downloading the app"* - Non-engager.

6.2.4.2 Facilitators of initial engagement

We identified two key intervention-level facilitators of initial engagement, as follows:

- Intervention offering
- Simplicity

We identified three key individual-level facilitators of initial engagement, as follows:

- Desire to improve health
- Privilege/scarcity of pilot
- Positive feedback

Intervention-level facilitators

A frequently mentioned facilitator of initial engagement was the **intervention offering**, both the prospect of earning rewards, linking to the TDF component of reinforcement, and the option to receive a free wearable fitness tracker.

Intervention **simplicity** was also noted by participants as a facilitator of initial engagement. The **straightforward sign-up process** was widely reported across sample groups, and the perception of the intervention as a **simple app** with a **manageable level of requirement** further encouraged initial engagement. Limited procedural knowledge required links to the TDF component of knowledge.

> *"I didn't have to pay to take part or I didn't have to already have a Fitbit or some expensive equipment to be able to take part. I didn't have to do a whole load of face-to-face or telephone interaction with people every week to check in. It just recorded it*" - Churn-out user, 'Medium' reward arm

Positive peer feedback about the intervention also encouraged participants to engage, representing a component of the TDF component of social influences.

> *"She raved about it really, and she thought it was really good for her. The way she talked about it, I thought, well, crikey, I might give this a go.*" - Fully engaged user, 'Low' reward arm

Individual-level facilitators

On an individual level, participants were encouraged to participate in the intervention out of a **desire to improve their health**, aligning with the aims of the programme and linking to the TDF component of intentions.

> *"I wanted to lose weight, and then someone will now reward me losing weight, so I was interested in the whole idea from before the word go."* - Fully engaged user, 'High' reward arm

A further individual facilitator identified was the **sense of privilege** participants felt being part of a pilot. In some cases, this was linked to a sense of **local pride**, as Wolverhampton had been selected as the area for the pilot, or to a sense of **curiosity** about how the programme would operate.

> *"It was quite exciting. It's like, 'Oh, something that's just trialled here.' Feel like you're quite privileged to be part of it."* - Fully engaged user, 'Low' reward arm.

6.2.4.3 Barriers to sustained engagement

We identified key intervention-level barriers to sustained engagement, as follows:

- Technical issues
- Challenge-related barriers
- Insufficient benefit
- Inconvenience of reporting

We identified key individual-level barriers to sustained engagement, as follows:

- Time/cognitive bandwidth
- Mental health

Intervention-level barriers

Wearable fitness tracker issues were widely reported as a barrier to sustained engagement. Churn-out users reported technical issues with the software, including **difficulty syncing** with the app, **difficulty recording exercise** on the Better Health: Rewards Tracker, and perceived **inaccurate step count.** Some churn-out users reported issues with wearable fitness tracker hardware, noting that the material was flimsy and irritated their skin.

> *"Just connecting [the wearable fitness tracker] to the app, and it was just - I couldn't be bothered with it, to be honest. It was just a bit too hard."* - Churn-out user, Control arm

Perceived **inconvenience of accessing support** prompted participants to churn-out rather than seek solutions to technical issues, reflecting the TDF component of beliefs about capabilities. However, participants who did use the support processes in the app reported that the process was straightforward.

> *"I just didn't want to go into the emails and wait for a couple of days to get a response and all that, I just stopped it."* - Churn-out user, Control arm

Churn-out users reported **boredom with challenges** that they felt were repetitive, particularly for participants who were unable to choose all challenges, such as those who do not eat meat, or who do not typically choose puddings. Some participants also noted concerns with **challenge difficulty**: some felt they were too challenging, linking to the TDF component of beliefs about capabilities, while others felt they were insufficiently challenging. Excessive challenge difficulty was particularly impactful for picky eaters, participants who felt that their eating habits were set, and those who typically snack rather than eat meals.

> *"I feel like they could add more challenges. I did definitely get bored. In the end, I just stopped using it completely."* - Churn-out user, 'Medium' reward arm

Unfair challenge failures contributed to some participants' decisions to disengage with the programme, a form of perceived sanction linking to the TDF component of reinforcement. For example, one user reported that the app rejected a photo that

he felt was a healthy meal, leading him to disengage with the programme. However, infrequent or 'fair' challenge failures did not prompt churn-out.

> *"After two weeks of not achieving, when, actually, I thought I was doing quite well health-wise, [...] not achieving meant that I didn't really want to use it anymore"* - Churn-out user, 'Medium' reward treatment arm

Some participants felt that the **rewards offered were not worth the effort required**, representing insufficient reinforcement. This was particularly pertinent for participants allocated to the 'low' reward arm, who felt that challenges offered too few points, translating to too small a reward value. It is important to note that some interview participants used 'points' as a way of referring to reward value, particularly those who found the conversion rate between points and financial reward value difficult to understand. The perception of challenges offering too little value also applied to participants who lived in less deprived areas, who noted that they would be more likely to pay for a subscription to a health-focused programme than to be motivated by the offer of rewards. Nonetheless, other participants who did not 'need' the rewards found them exciting. Some churn-out users reported losing motivation to pursue challenges when the numbers of points available for challenges decreased.

> *"I think £20 took me about four, five weeks to earn, and I think, for me - I know that everyone's feeling the squeeze a little bit - but over five weeks, £20 for me, possibly wasn't worth it."* - Churn-out user, 'Medium' reward arm

Some churn-out participants who used other health and fitness apps reported that the Better Health: Rewards app did not offer the same **level of detail or approach** that they received from other apps, such as measuring heart rate. This was also noted by some fully engaged participants, although it did not prompt them to churn out, and noted by some participants who initially benefited from the intervention but felt they had **gained as much as they could**, replacing the intervention with more specialised technology. Some participants also began to replace the intervention with alternatives given its defined end date.

Inconvenience of reporting in the app was also identified as a barrier. Churn-out users, particularly those whose jobs limited phone use, discussed **forgetting or being unable to take photos of their meals**, which was required for some validated challenges. Forgetting to report links to the TDF component of memory, attention and decision processes, while inability to report speaks to the component of

environmental context and resources. Failing challenges as a result of failures to report discouraged participants from continuing. The **inconvenience of the Intake24 survey**, which was described as laborious and long-winded, was also highlighted by churn-out and fully engaged participants.

> *"It was quite cumbersome. You couldn't just put a mixed salad, because I was trying to eat a lot of salad, and it wanted to break every part of the salad down"* - Fully engaged user, 'High' reward treatment arm

Individual-level barriers

Lack of time and cognitive bandwidth was reported by churn-out users as limiting their ability to engage with the intervention, particularly when experiencing other time pressures such as long hours at work or caring for children. However, participants liked that challenges ran over a week, so that if they were unable to achieve the challenge task on one day, they could make up for it on other days. This barrier links to the TDF component of memory, attention and decision processes.

> *"I was looking after three kids; I was working at the time. I just can't be bothered to cook something healthy"* – Churn-out user, 'Medium' reward treatment arm

Linked to lack of cognitive bandwidth, churn-out participants reported that repeatedly **forgetting to charge, wear, or sync their wearable fitness tracker** increased their likelihood of churning out, as this would lead to participants failing step count challenges. This barrier further maps to the TDF component of memory, attention, and decision processes. Participants also reported that the wearable fitness tracker battery often did not last long, requiring frequent charging.

> *"When I stopped using the app, it's probably when it had run out of battery, and maybe either forgot or just, yes, couldn't be bothered maybe to keep charging it."* - Churn-out user, 'Medium' reward arm

Mental health-related barriers prompted some churn-out participants to disengage from the app, due to loss of motivation and self-confidence. Participants who experienced mental health-related barriers reflected that when they lost motivation due to a decline in mental health, the requirements of completing challenges felt too much. This concern was also flagged by non-engagers with experience of mental health challenges as a reason why they decided not to engage with the

intervention initially. For example, one non-engager who experiences poor mental health assumed that the app would require going outside and to the gym, which they felt unable to do.

> *"I was struggling with my mental health anyway, and then I just couldn't be bothered anymore."* - Churn-out user, 'Medium' reward arm

6.2.4.4 Facilitators of sustained engagement

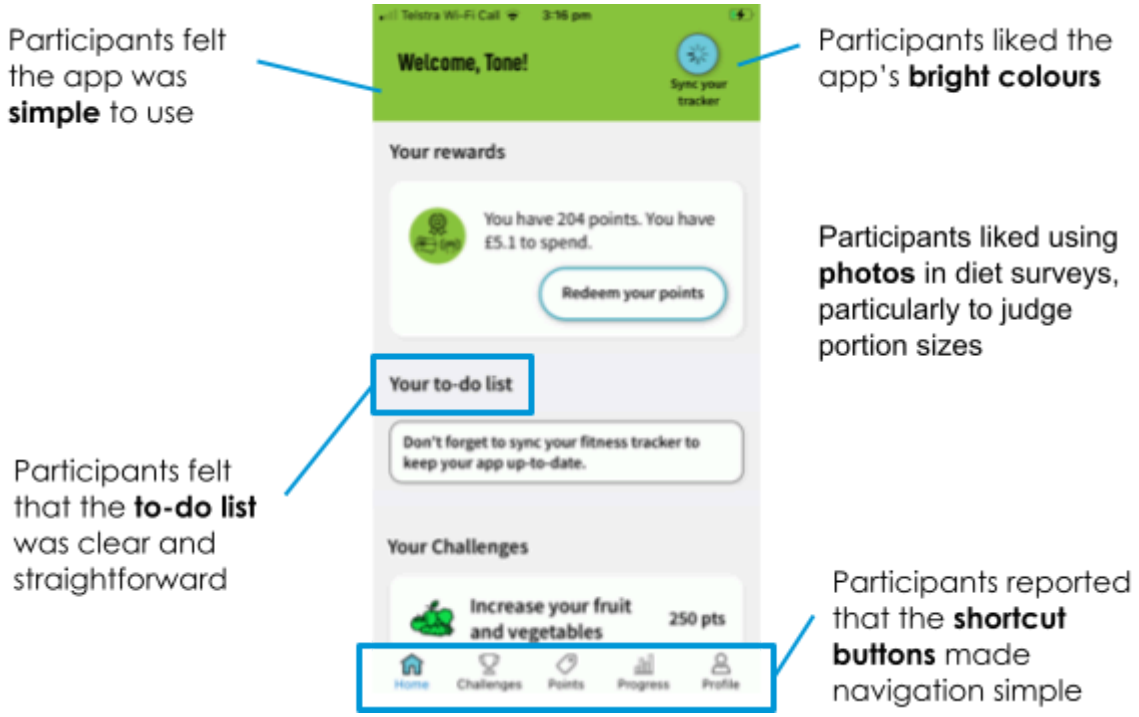
We identified key intervention-level facilitators to initial engagement, as follows:

- App content
- Achievable tasks
- Reminders

Participants reported that **the app was simple to use**, with no unnecessary complications; it was 'good enough' for the requirements. This maps to the TDF component of knowledge of the task environment, and beliefs about capabilities. Participants reflected that they liked the **tone** of messaging in the app, which they felt was informative and supportive, not patronising or critical.

> *"There was a lot of hand-holding, but as I say, not in a patronising way. It was more educational"* - Fully engaged user (focus group)

Figure 19: App homepage



Fully engaged participants reported that their perception of the intervention tasks as '**achievable** but meaningful' encouraged them to continue engaging with the intervention. As with the 'simplicity of the app', this relates to both knowledge of the task environment and beliefs about capabilities. The ability to **select from a range of options** for physical activity and dietary challenges reinforced this perception of achievable tasks.

> *"There was always an achievable option, so you never felt like you were failing even if it was a difficult week."* - Fully engaged user (focus group)

Reminders, both in-app push notifications and email reminders, were a useful prompt to participants to encourage them to **complete their activities** within the intervention, particularly those with busy lives or those who described themselves as forgetful. In addition to prompting participants to complete specific tasks in the app, reminders also served as a **general reminder to keep healthy and active**. These reminders acted within the TDF component of memory, attention and decision making.

> *"They were really good prompts and reminders for me because my brain's like, honestly, it's like a sieve."* - Fully engaged user, Control arm

Some participants preferred in-app notifications, while others preferred email reminders. Participants generally reported that the frequency of reminders was appropriate. The personalisation of reminders made participants feel special, encouraging them to continue.

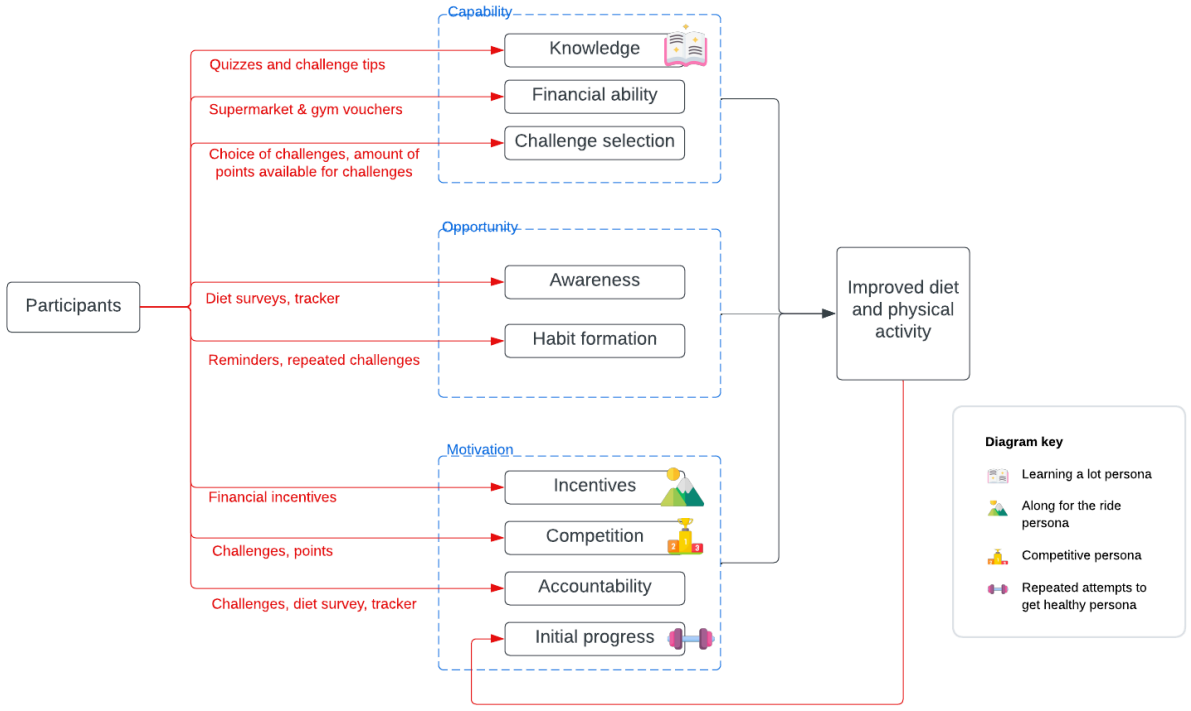
6.3 Mechanisms of impact

6.3.1 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?

How do features of the Financial Incentives Scheme affect the extent of behaviour change?

Among participants for whom the intervention elicited behaviour change, this took place through a range of mechanisms of impact. These mechanisms were derived from inductive analysis of qualitative data, and subsequently mapped onto the COM-B framework. One such mechanism was the offer of financial incentives. This can be seen, alongside other relevant mechanisms of impact, in Figure 20 below. The mechanism of impact most relevant to each persona is highlighted in the diagram.

Figure 20: Intervention mechanisms of impact summarised within the COM-B framework



Capability

Increased **knowledge** from quizzes, information pages, and challenge tips facilitated healthful behaviour change, particularly for 'Learning a lot' personas. Participants for whom this information was new were engaged by it, while others who were familiar with the information viewed it as a useful reminder of how to pursue healthful behaviour.

Some participants viewed the information as patronising. This included those who felt they knew enough or who didn't attribute unhealthy decisions to a lack of knowledge but their own choices.

> *"I now know that each meal should consist of fruits, vegetables, protein, and carbohydrates, not just a lot of carbohydrates."* - Fully engaged participant, 'High' reward treatment arm

> *"I don't think it told me anything I didn't know already, but I did read things through, and it just refocuses you from time to time with certain things"* - Drop out participant (focus group)

Participants who said that they could not otherwise afford certain healthful dietary and physical behaviours may have been **financially enabled** by rewards such as supermarket vouchers. It should be noted participants were not limited to using supermarket vouchers on healthy food. However, some interview participants suggested that they felt encouraged to purchase healthy food as they were aware that they had earned the vouchers through a scheme promoting healthfulness, this was not universal across participants. Additionally, we do not have data on the healthfulness of food purchased with reward vouchers as this was out of the scope of this study.

Relatedly, the provision of a wearable fitness tracker overcame potential financial barriers to participation for those who would not be able to afford to purchase one.

> *"Having that extra £10 gift card, I could go to Tesco, buy a nice box of grapes"* - Churn-out user, Control arm

The **mechanism of selecting challenges** facilitated change across a range of healthful behaviours. The process of challenge selection was designed in a way which retains individual choice over which challenges to select, therefore incorporating self-regulation. Users of the app were encouraged to select two challenges each week. Participants could select a new challenge, or continue an existing challenge. Participants who selected harder challenges, or participants who were 'on a streak', were offered a greater number of points. Participants could alternatively choose to maintain the same level of challenge difficulty and earn fewer points if they wanted. Participants who were motivated by earning points typically selected challenges offering the most points, while others who were less motivated by earning points could select challenges based on areas they wanted to improve, challenges that would be achievable, or challenges that they would find interesting or engaging. However, some participants who focused on achieving points felt that the focus on points detracted from focusing on health.

> *"I picked the same challenges but then saw that the points decreased, so then I was like, actually, no, why do I want to keep it easy? I want the maximum points possible. Then I would go and pick a challenge that had more points."* - Fully engaged user, Control arm

Opportunity

Increased **awareness** of diet (from Intake24 diet surveys) and physical activity levels (from the wearable fitness tracker) was identified as a further mechanism for behaviour change. The app and wider components of the intervention (such as the wearable fitness tracker) provided material resources for participants to measure and understand their actions. Gaining this information enabled participants to reflect on their habits and correct misperceptions about their diet and activity levels.

> *"Making you write down everything that you've had in that day puts it into perspective a little bit. Like, I didn't realise I ate so much there, or I didn't really eat much that day, and how much the calories can tally up."* - Fully engaged user, 'Medium' reward arm

Participants who repeated actions and challenges during the intervention reported that **habit formation** enabled them to develop new healthful habits or renew old habits. The ability to repeat challenge types across multiple weeks within the physical (digital) environment thus promoted habit formation. This related particularly to checking step counts, maintaining dietary swaps, and ensuring that meals contained healthful components.

> *"It's become habitual now where you start looking at, as [user] said, the five portions of fruit and veg. It's become an automatic thing now where you're subconsciously arranging your meals around that model."* - Fully engaged user (focus group)

Motivation

The ongoing automatic motivation offered by the offer of **incentives** was identified as a widespread theme across sample groups. The cost of living crisis and rising food costs were recurring themes, particularly among participants from areas of higher deprivation, which contextualised the value of shopping vouchers in particular. Some more affluent participants reported feelings of guilt as they did not need rewards, although this did not impede their motivation. Participants allocated to the 'high' reward arm commented on the generosity of rewards available.

> *"When there's money involved, I think that definitely spurs you on more"* - Fully engaged user, 'Medium' reward treatment arm

However, participants who were primarily motivated by a desire to improve health and lose weight reflected that they would have remained motivated in the absence of rewards.

Relatedly, some participants who felt that incentives represented an initial hook to prompt their engagement in the programme suggested that longer-term engagement with behaviour change would require additional motivation as well as incentives, such as observing initial progress.

> "I think it was a motivator to get up and running." - Fully engaged user, 'Low' reward arm

Indeed, participants who observed **initial progress** in mental health and weight loss felt that this reinforced their motivation to improve their diet and physical activity. For these participants, improvements in exercise and diet were mutually reinforcing. For example, some reported feeling more comfortable exercising in public following initial weight loss.

Moreover, initial progress in achieving challenges and improving step count enhanced participants' beliefs about their capabilities. This contributed to reflective motivation to continue engaging with the intervention. In addition, reinforcement through observed success such as weight loss or feeling healthier, provided automatic motivation.

> *"Overall I felt better in myself, and then you start thinking, right, I'll stick with this, I can stick with this."* - Fully engaged user, 'High' reward arm

Participants who described themselves as 'competitive' or 'disciplined' reflected that **competition**, either with themself or with others, motivated behaviour change. Competitive motivation arose from the desire to achieve challenges and accumulate points, from personal commitment to complete challenges that they had selected, and, for those whose peers were also using the app, comparisons with friends and family. Reinforcement through success or failure in varying forms of competition contributes to automatic motivation.

> *"It was exciting. Me and my friend, we set goals between us, like, 'Oh, I'm going to reach my goal. Are you going to reach your goal?' It was exciting."* - Churn-out user, 'Medium' reward treatment arm

Accountability offered by the intervention was reported as motivation to change behaviour. Participants valued the sense of accountability and structure offered by regular weekly challenges and frequent diet surveys. Interview participants reported that knowing that they would have to record their actions in the app encouraged them to try to improve their behaviour. Participants' awareness of the need to report actions on the app, and assessment through the app of challenge success or failure influenced their beliefs about consequences, contributing to automatic motivation.

> *"I quite like the idea of being responsible to somebody else for what I was doing [...] I then had to justify what I'd done that day."* - Fully engaged user, 'High' reward arm

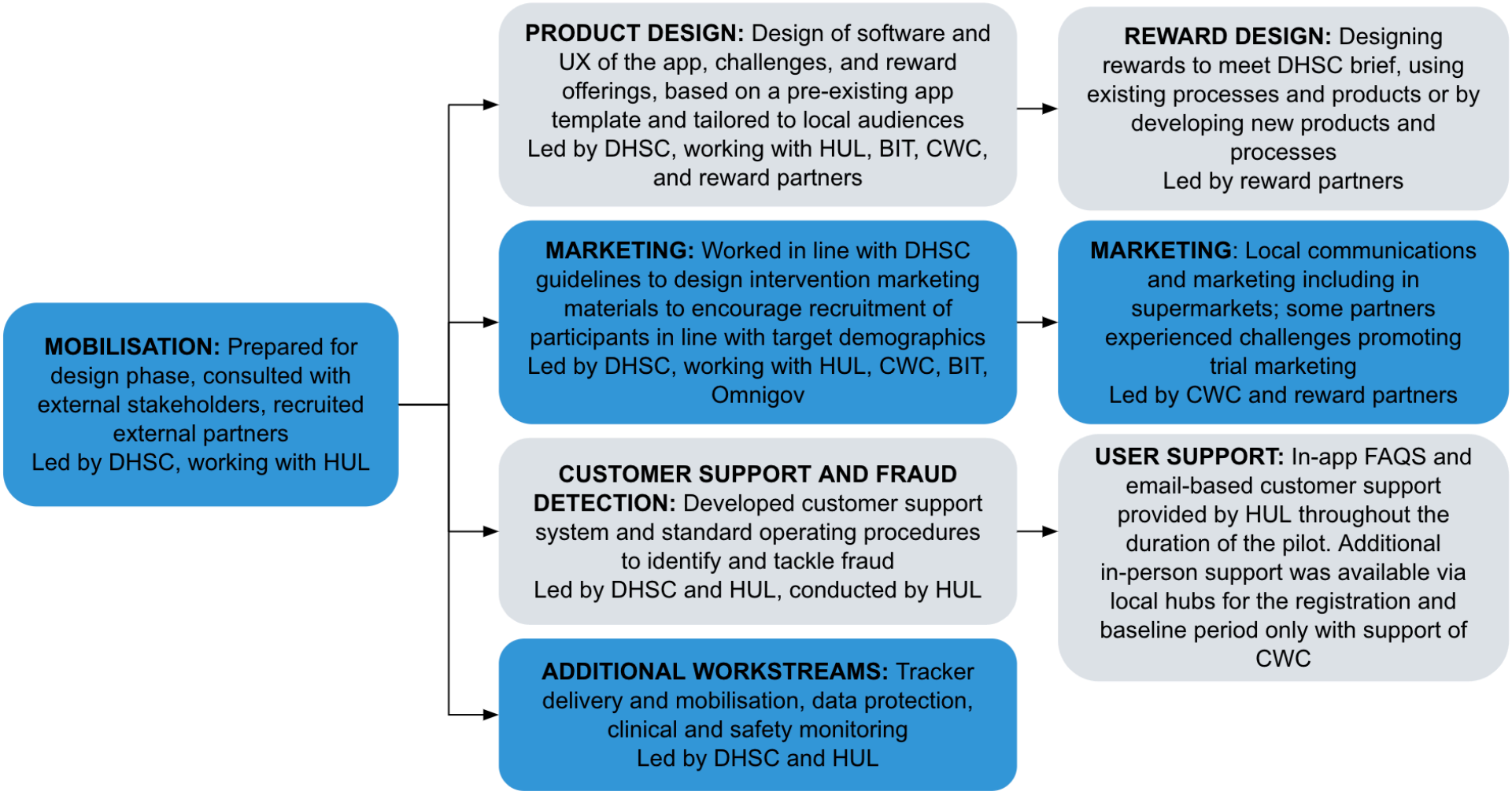
6.4 Implementation and scaling

6.4.1 What was the process for developing and implementing the intervention among delivery and reward partners?

A process map detailing the process for developing and implementing the intervention can be found in Figure 21.

This section first discusses the process for developing and implementing the intervention for the delivery partner (HUL), who developed and implemented the app, then for reward partners, who provided incentives for intervention participants to achieve, and finally for the City of Wolverhampton Council (CWC), who oversaw implementation and communications at a local level.

Figure 21: Process map



Delivery partner (HUL)

Findings relating to the process for developing and implementing the intervention for the delivery partner are reflected in discussion of the barriers and facilitators to intervention scaling (Section 6.2.4). In addition to this, we identified the following reflections from interviews with the delivery partner.

The development and implementation of the intervention by the delivery partner required multiple parallel workstreams, and coordination with a range of external stakeholders. One challenge reported by the delivery partner was ensuring that ways of working were aligned across different stakeholder groups with different experience of designing technology products. This navigation of new working relationships across partners with different approaches to processes could at times lead to prolonged timelines and risks of inefficiency.

Reward partner

Selection of reward to fit DHSC requirements

Some reward partners already had existing health promoting or health neutral voucher options. These partners were able to use these existing vouchers for the app, which represented a more straightforward process to implement, requiring minimal additional resources.

Reward partners who did not have a suitable existing voucher option developed bespoke reward options. This required more upfront resource commitment to develop the voucher and the processes for participants to redeem it. The use of a third-party reward aggregator increased the ease of implementation for reward partners.

> *"The e-gift card was set up and ready to go. It just needed to be turned on."* - Supermarket reward partner

Supermarket reward partners experienced challenges meeting DHSC requirements to restrict the voucher to certain product groups and promote healthful products. Some supermarkets did this by using product restrictions on existing gift cards that prohibited the purchase of items like alcohol or tobacco, or by implying restrictions existed through messaging. Alternatively, images of fresh produce were used to

promote healthful foods. It was argued that going beyond this, such as creating a bespoke gift card, would have been more resource intensive.

Amplification of intervention

Supermarket reward partners were also asked to display promotional materials to amplify the intervention. Some supermarkets were able to do this without difficulty, while others with more restrictions on what can be displayed in stores struggled to prioritise the materials as they were not deemed sufficiently linked to the supermarket due to DHSC branding.

City of Wolverhampton Council

Much of the work done by the City of Wolverhampton Council is reflected in more detail in Section 6.1, where intervention reach is discussed. In addition to this, we identified the following reflections from focus groups with CWC.

Local delivery partners felt that their role involved presence 'on the ground', both contributing to intervention and marketing content, and subsequently encouraging engagement and offering user support through the 'Better Health: Rewards Hub'. The approach taken by CWC was very broad, and required significant resource and time input, which focus group participants reflected would be challenging to replicate.

> *"That's where your population-level approach and your digital element meets local nuances"* - CWC

6.4.2 Are the design and delivery processes fit for scaling and sustaining the intervention?

Interviews with delivery partners (HUL), reward partners, and CWC indicate that design and delivery processes are fit for scaling and sustaining the intervention. Technical processes can largely be scaled with minimal difficulty, and reward partners indicated that they would be able to scale existing processes. However, some elements would require adaptation to be delivered effectively at scale, such as automating currently manual customer service and fraud detection processes. Further detail on potential barriers and facilitators to scaling can be found in the following section, Section 6.4.3.

While current processes are *fit* for scaling, certain improvements could be made to optimise these processes.

6.4.3 What are the facilitators and barriers to scaling and sustaining the intervention (including financial incentives) beyond the pilot?

Table 20: Barriers and facilitators of scaling the intervention beyond the pilot

	CWC	Reward Partners	Delivery Partners
Commitment due to partners' aligned values with intervention aims			
Ability to dedicate more resources to larger scale programme			
Extent of marketing requirement			
Limited degree of partnership			
Manualised processes			
Lack of feedback			
Need for greater consideration of diversity and inclusion			
Technical issues			

Notes: Green - facilitator experienced by stakeholder; red - barrier experienced by stakeholder; CWC - City of Wolverhampton Council

6.4.3.1 Facilitators of scaling and sustaining the intervention

A key facilitator to scaling and sustaining the intervention identified was the commitment of reward partners and local delivery partners to the project due to **aligned values** with the aim of the intervention.

> *"It's just great to support a project which is looking to shift healthy behaviours because as a business, that's also our target too, so I think, yes, getting involved with something like that was almost like a no-brainer."* - Reward partner

Extensive efforts to complete foundation work for the pilot project were expected to increase the relative ease of scaling. As much of the initial time- and resource-intensive baseline work was completed during the pilot (such as dedicating technical resources to overcome challenges in creating unique codes), it was thought that scaling up would not require a repeat of resource investment on product design and delivery processes.

> *"Could be scaled quite easily as well because as I said, because the framework has now been written, the actual programming bit is quite easy."* - Reward partner

Reward partners also reflected that they would likely be able to **dedicate greater-value resources to a larger-scale programme**; reward partners highlighted that, due to the unique small-scale nature of the pilot, they were limited in their approach to promotion of the intervention.

6.4.3.2 Barriers to scaling and sustaining the intervention

Local delivery partners and some reward partners reported **concern about fulfilling the requirement for marketing and communications at scale**. As discussed in Section 6.4.1, CWC reported concerns that the scale of resource effort employed for the pilot could not be replicated at scale, while some reward partners reflected concerns about their ability to prioritise the intervention marketing over existing organisational material.

For reward partners, some of this concern arose from the **limited degree of partnership** in intervention marketing. As marketing was DHSC-branded and not co-produced with reward partners, reward partners felt that they might experience difficulty prioritising the project. Reward partners suggested that this could be overcome by a more in-depth partnership, such as linking rewards to existing supermarket loyalty programmes. A more in-depth partnership could also involve tracking intervention participants' purchases directly, removing the requirement for challenge validation.

> *"At the end of the day everything is Department of Health branded, so it's a bit trickier in terms of getting it into store plans"* - Reward partner

Delivery and reward partners reflected that some elements of the pilot relied on **manualised processes**, which, while appropriate at the level of a pilot, would be challenging to scale. This included processes relating to customer service, both in-person and in-app, and fraud detection and response. In addition, delivery partners experienced a greater need for customer support than anticipated, and reflected that careful selection of third parties might minimise this.

While reward partners mostly indicated that they would be able to scale their reward offering with limited changes, those who experienced fewer redemptions than anticipated noted that they would require **feedback** around reward selection and redemption to enable adaptation to their approach before committing to scaling.

> *"To build a business case around it, we need to have that data or understand, yes, the potential and how many are going to be purchased."* - Reward partner

One delivery partner reflected that while the pilot incorporated a substantial awareness of **diversity and inclusion**, this would need to be considered more substantially at a larger scale. For example, a nationally-scaled intervention would need to consider language barriers. This requirement is reflected in part in the reflections of the local delivery partner of the value of a hyper-local focus.

> *"At larger scale, there's a lot more that needs to be worked on, so it is more inclusive"* - Delivery partner

A final barrier to scaling was **technical issues**, mentioned by delivery partners, and reflected in interviews with participants. These related primarily to the wearable fitness tracker, but also to the software of the app. The delivery partner reported challenges to standardisation arising from the use of different wearable fitness tracker types, which record exercise and steps in different ways, while participants reported a range of issues with wearable fitness tracker software and hardware, as well as occasional bugs and 'glitches' in the app.

6.4.4 What was the extent of 'gaming' and data errors in the intervention?

One metric which we specified beforehand might help to detect the extent of 'gaming' and data errors is the proportion of participants reporting unreasonable body measurements at any point during the pilot. Before the study began, we had also considered examining the percentage of points earned that were non-validated, however this was not possible as we did not receive data on non-validated points. Specifically, we have calculated the percentage of participants who reported a BMI of over 70 kg/m^2 or under 10 kg/m^2 was 0.24% across the entire randomised sample. This varied across trial arms, from control (0.27%), to low reward (0.16%), to medium reward (0.26%), to high reward (0.19%). It is clear therefore that this phenomenon was not a serious issue during the pilot, and that there was no clear relationship between 'unreasonable' body measurement data and the incentive structure a user faced.

Intervention participants suggested that it might be **possible** to misreport behaviour, as much of the intervention relies on user honesty. However, across sample groups this was viewed as **counterproductive**, reflecting the view that those who did so would only be cheating themselves.

Participants felt that there was no better way to record diet, as an approach based on more rigorous monitoring of diet would not be feasible or ethical. Additionally, participants felt that the existing level of reward would not encourage fraud; some suggested that increasing the reward value or introducing competitive leaderboards might increase motivation to misreport.

Delivery partners felt that **fraud detection policies were effective**. They felt that the fraud risk guidance standard operating procedure was clear, and that levels of fraud were consequently well-monitored and at a low level. However, the process for detecting fraud was largely manual, which would be challenging to replicate at scale.

> *"I don't think it's [fraud] been a big risk on this programme and it's been very well monitored"* - Delivery partner

7. Discussion

Summary of Key Findings:

- The study reached its recruitment targets at the point of registration, though the number of male participants, older individuals, and those of white ethnicity was lower than anticipated.
- 12,767 participants proceeded to randomisation. Those randomised were more likely to be female, older, white, and income-deprived than the general population of Wolverhampton LA. They were also more active (with a higher daily step count at baseline), but had a less healthy diet than the general UK population.
- By the end of the pilot, attrition was lower among those who received the largest incentives. Participants who provided their data at M5 and were hence included in the impact evaluation were in line with targets for ethnicity. However, those who were still active at M5 were older, less income-deprived, and more likely to be female than set out in pilot targets.
- Pilot analysis showed that offering financial incentives did not significantly impact participants' PA (MVPA and steps) on average. The average difference in steps and MVPA between the control and pooled reward arms was small, at +1.1 min/day of MVPA and +42 steps/day.
- Under sensitivity analysis, offering an incentive statistically significantly increased both MVPA and steps, with substantive effect sizes increasing to 1.9 minutes per day and 256 steps per day respectively.
- The pilot provided some evidence that offering financial incentives can increase participants' fruit, vegetable and fibre consumption. Offering financial incentives significantly led to small statistically significant increases in participants' consumption of fruit and vegetables (treatment effect: 21g per day) and fibre (0.4g per day).
- Offering participants larger financial incentives did appear to increase effects on dietary outcomes, with those in the high reward arm

demonstrating greater changes. This relationship was particularly strong for fruit and vegetable consumption (an increase of 29g per day), and saturated fat (a reduction of 0.4g per day). The effect on PA outcomes was not consistent, with a statistically significant increase in step count found for participants in the medium incentive arm alone.

- The pilot provided some evidence that incentive structures had a positive effect on the engagement and completion of diet challenges, with participants in higher reward arms completing more challenges (for both diet and PA).
- The analysis does not offer a clear picture of patterns in effect size over time (1-3-5 months from randomisation).
- The results did however suggest that offering financial incentives varies between socioeconomic groups: with more deprived individuals being more likely to change their PA and less deprived individuals being more likely to change their diet. Larger effects on physical activity were found among participants who were initially inactive, and impacts for both physical activity and dietary patterns appeared to be more pronounced in participants aged over 41.

7.1 Overall findings

This pilot set out to determine whether financial incentives are an effective means of increasing physical activity and diet healthfulness. **Overall, the pilot provides some evidence that financial incentives may be an effective way of bringing people closer to guidelines for a healthy diet and physical activity, but with varied results.** The pooled intervention did not significantly impact participants' physical activity outcomes, but did lead to small increases in their intake of fruit and vegetables, and fibre.

The pooled intervention had a positive impact for specific groups, such as more deprived individuals, previously inactive users, and older adults. Incentives were more effective in increasing fruit and vegetable intake among participants who were less deprived, and older participants achieved statistically significant increases

in their MVPA, fibre, and fruit and vegetable intake, alongside a statistically significant decrease in saturated fat intake compared to their counterparts. Financial incentives led to a small significant increase of 2.1 minutes per day in MVPA for inactive users, compared to a null effect observed for active users.

Larger incentives led to bigger impacts for dietary outcomes. All incentive levels for fruit and vegetable intake yielded positive, statistically significant effects, while the high reward arm showed a statistically significant reduction in saturated fat intake. While the effect of the incentive structure on PA outcomes was limited, those allocated to the medium reward arm were an exception, with a strong positive effect on step count.

7.2 Recruitment and engagement

The pilot broadly met its targets for recruitment and engagement, though the demographic composition of the sample changed over time. While 34,900 adults consented to take part, 12,767 completed all onboarding steps and the baseline period to be randomised into the study. Those who made it to this stage had different characteristics to the general population of Wolverhampton; they were more likely to be female, older (with a median age of 41) and white (68% of those randomised). They were also twice as likely to be income deprived; where 21% of the population of Wolverhampton meet this definition, this group comprised 50% of the pilot sample. Overall, while the sample did not perfectly represent the general population, limiting the generalisability of the findings, the diverse marketing campaign for the pilot was effective. The majority of the targets for the demographic composition of the sample were met, with the exception of gender where women were overrepresented (67% of the sample against a target of 50%).

The pilot successfully reached participants who were falling below official guidelines for diet healthfulness. Those randomised had a less healthy diet than the national UK population; 79% of the randomised sample reported eating fewer than five portions of fruit and vegetables per day, the minimum recommended intake, while 67% of UK adults do not meet this recommendation (National Diet and Nutrition Survey years 2016/17 – 2018/19) This suggests that the intervention effectively reached individuals who were not achieving UK government dietary recommendations on diet, satisfying

the primary objective of the pilot. Participants themselves reflected that the focus on improving health strongly resonated with them, and that they recognised it was a 'good thing' that would be useful.

On the other hand, those randomised were more likely to be physically active (measured by steps per day at baseline) than the national UK population. This may reflect a pattern of positive self-selection into the pilot, where participants with healthier physical activity behaviours were more likely to join in and stay engaged. It may also reflect the effect of the app and/or the wearable fitness tracker on participants' physical activity behaviours.

These differences in the characteristics of the final group affect how we understand the study results, which we will now explain in more detail.

7.3 Does offering financial incentives improve PA and diet?

The pooled intervention did not significantly impact physical activity outcomes. This was true for both moderate and vigorous physical activity (MVPA) and daily step counts. The direction of the descriptive effects aligned with our hypotheses, though these differences were small. The average substantive difference in MVPA and steps were +1.1 min/day, and +42 steps/day respectively for those in the pooled reward arms compared to control, though both were non-significant, so may have occurred by chance.

However, under sensitivity analysis, offering an incentive statistically significantly increased both MVPA and steps, with substantive effect sizes increasing to 1.9 minutes per day and 256 steps per day respectively.

The intervention had positive effects on some dietary outcomes. Analysis of dietary primary outcomes revealed a small positive significant relationship between the intervention and the consumption of fruit and vegetables, with the pooled reward arms consuming 21g more on average per day than those in control. It also led to a small statistically significant increase in fibre intake (treatment effect: 0.4g). The pilot also assessed the impact of financial incentives on the consumption of saturated fat and free sugars. While no comparable effects were found for these outcomes and

descriptive effects were small, they were in the expected direction (reductions in saturated fat and sugar intake), in line with our hypotheses.

The intervention did not significantly impact broader PA and diet-related factors, including energy expenditure, energy intake, and self-reported weight.

Overall, a comparison of effect sizes across our primary outcomes for physical activity and diet suggest that the pooled intervention was more successful at improving dietary outcomes than those for physical activity. Those who were closer to guidelines for diet and physical activity at the point of randomisation were less likely to churn out of the pilot (regardless of the trial arm they were assigned to) and more likely to use the app and earn points. This indicates greater engagement among those already motivated to maintain good health. The lack of difference between trial arms may support findings from interviews conducted with participants, who report that the wearable fitness tracker and/or app alone encouraged healthier behaviour.

However, the results observed for dietary outcomes may also be a function of the composition of the sample. Notably, people who ate more fruit and vegetables (at baseline) tended to complete fewer physical activity challenges than their counterparts, but were more likely to select and complete diet challenges. Since this group was overrepresented in the pilot sample, and also tended to earn more points, the pilot would have been better able to capture changes in these outcomes.

Further, more diet challenges were available than PA challenges, and diet challenges were selected twice as often as PA challenges overall.

7.3.1 Physical activity: steps

Some participants reported improvements to physical activity, such as increased step count and engagement or re-engagement in exercise. Many interview participants reported ongoing attempts to increase their daily step count, such as choosing to walk rather than drive. However, this qualitative evidence failed to translate into a substantial effect across the population in the study. The descriptive effect on steps per day in the pooled reward arms (+42) is smaller than that seen in the 'Carrot Rewards' app evaluation, which provided very small immediate rewards

for participants who increased their daily step count (Mitchell et al. 2018). Participants in this quasi-experimental study could earn points which could be redeemed for rewards like groceries, cinema tickets, or petrol. On average, they walked 116 more steps per day at study week 12 than at baseline, which was described as a 'modest' effect size. The effect size observed in this pilot is also considerably lower than the mean daily step count increase of 1,579 observed during the Singapore National Steps Challenge (Yao et al. 2022). In this study, a pre-post design was used to measure the effects of an intervention that similarly offered points for physical activity, which could be converted into financial rewards, as well as tickets for 'prize draws'.

There are several potential explanations for the smaller descriptive effects observed in this pilot. The pilot design is experimental, and isolates the impact of the financial incentives, unlike the previous studies (both pre-post designs) in which the descriptive effects noted cannot be attributed to the incentives alone, due to the study designs.

In this pilot, the control arm was an 'active control', who received both the wearable fitness tracker and app at baseline, before randomisation took place. On the assumption that this in itself could impact behaviour, it is possible that all trial arms (control arm and reward arms) responded by increasing their daily steps over and above their usual level of daily activity during baseline. Assessing the impact of the availability of the app itself, or of providing a wearable fitness tracker, was outside the scope of this evaluation. However, we acknowledge the possibility that, if these technologies were an important factor for an increase in physical activity, the addition of financial incentives may have had little room to further increase participants' motivation.

It is notable that the average number of steps per day, both at baseline and at month five, is considerably higher than the UK average of 5,444 (Althoff et al. 2022). This is compatible with a positive self-selection of participants in the pilot - participants who do not churn out from the pilot being more likely to have higher motivation or baseline physical activity to begin with (see Table 11 in Section 6.3. 'Mechanisms of Impact' for evidence that those who were closer to guidelines for physical activity at the point of randomisation had lower levels of churn out during the pilot). It is also compatible with some of the themes that emerged from the qualitative IPE. Participants noted the appeal of both the wearable fitness tracker

and the rewards, with particularly participants from areas of higher deprivation suggesting that the receipt of the wearable fitness tracker itself was a motivating factor for signing up to the pilot.

We note however, that the Althoff figure is derived from smartphone data and as such, is unlikely to be affected by selection bias or the Hawthorne effect (i.e. the notion that we change our behaviour in response to being measured). On the latter, our cohort included ~71% of participants ordering the Better Health: Rewards tracker (with the remaining participants recording steps on FitBits, or on Garmin, GoogleFit, or HealthKit watches). As such, we can assume we have a large proportion of participants tracking their steps for the first time (with the Better Health: Reward tracker) and therefore may be disproportionately affected by the Hawthorne effect.

Finally, limited impact and a higher overall step-count may be a function of the patterns of churn-out and engagement observed during the pilot. Those still active at M5 progressively used their wearable fitness trackers less, and the rate of completion for PA challenges fell much more steeply over the lifetime of the pilot than for diet challenges. The sample remaining at M5 was also closer to guidelines for physical activity (measured at the point of randomisation).The implication of this kind of self-selection could be that the effect of the incentives was diluted; for a group that were already more motivated and active, existing levels of physical activity may have been harder to move. The incentives themselves may not have been sufficiently motivating for this group, a finding reflected in the qualitative IPE. Participants who described themselves as 'competitive' or 'disciplined' reflected that it was the element of competition, either with themself or with others, that motivated behaviour change. This type of user (the 'competitive participants' described in the qualitative personas) was also more likely to be closer to recommended physical activity level at the point of randomisation.

While the descriptive effects detected in this pilot were small, they were both positive and compatible with other findings from our sensitivity and subgroup analyses. Findings from our sensitivity analysis generated a larger, statistically significant increase of 256 steps per day in the pooled reward arms. This increase brings the average daily step count for the pooled reward arms to 6883, 86% of the daily steps recommended by the CMO.

The existing evidence base for the link between step count and mortality is

encouraging, and suggests relatively small effects may still be beneficial. A recent meta-analysis shows that any increase in steps has an inverse effect on all-cause mortality and cardiovascular mortality above the cut-off points of 3,967 and 2,337 steps/day, respectively. In this study, a 1000-step increase correlated with a significant reduction of all-cause mortality of 15%, and a 500-step increase correlated with a reduced risk of CV mortality of 7% (Banach et al. 2023). While the observed effects in our primary analyses do not meet these thresholds, the substantive effect of the intervention became both statistically significant, and increased in magnitude to 256 steps per day under sensitivity analysis planned ahead of collecting the final data. This analysis included participants' who did not meet the original threshold for wearable time. As a result, the composition of this sample was more balanced across several demographic characteristics. It included more men, more participants who were younger, and a greater number of those who were more income-deprived, suggesting the results of this analysis may be more generalisable to the wider UK population. **In combination, these analyses suggest that this intervention could have a small, but meaningful, impact on participants' step count.**

A recent meta-analysis of 15 studies examining the link between step count and mortality suggests a curvilinear relationship; in other words, that the risk of mortality decreases as step count increases, but only up to a point. The point at which the benefits plateau however is quite high and varies by age, with younger people needing to take between 8000-10,000 steps before the benefits begin to level off (Banach et al. 2023). The same study generated two other findings that are relevant to the findings of this pilot; first, that people who took more steps had a 50–60% lower risk of dying compared to those who took the fewest steps; and second, that the most significant drop in mortality risk was seen in people who increased their steps from a low starting point (Paluch et al. 2022). This suggests that not only is there a benefit associated with any initial improvement in step count, but that increasing step count is particularly beneficial among individuals who are more inactive, a characteristic of nearly two thirds of the sample recruited for this pilot.

As noted, under sensitivity analysis, the substantive effect observed on steps increased significantly. These analyses also reduced the average number of steps substantially, from 7,612 to 6,412. While this was in part a function of the increase in the size of the sample (reducing the average step count overall), it does not

account for the increase we observe in effect size.

One possible explanation is linked to participants' use of the wearable fitness tracker. Including participants with a shorter wear time may have captured those who were more likely to wear the wearable fitness tracker when they knew they were likely to earn rewards, for instance before they set out for a long walk or run. This reflects the fact that participants reported they were motivated by the wearable fitness tracker, and that its ability to accurately record their physical activity was a key factor in their engagement with the pilot. Qualitative IPE findings suggest that the wearable fitness tracker acted both as an incentive to sign up to the pilot and a factor in attrition, with participants reporting that issues with syncing, difficulty recording exercise, and inaccurate step counts were reasons for dropping out of the pilot.

7.3.2 Physical activity: MVPA

The effect of financial incentives on MVPA minutes per day was also small. Descriptive findings suggest a non-significant increase of just over 1 minute per day for those in the intervention group. The average number of MVPA minutes per day across all trial arms was 15.4 minutes, substantially lower than the 48.1 minute average observed in other studies using accelerometer data (Rowlands et al. 2021). While the magnitude of the effect nearly doubles to 1.9 minutes following sensitivity analysis, and becomes statistically significant, it is still smaller than that observed in other trials. The active control and self-selection effects discussed earlier may have been a factor, with a more motivated sample potentially diluting the impact of the incentives. It is also important to note MPVA data may not have been counted if participants forgot to activate the relevant mode on their device, which increases the likelihood that the effect size we find is a conservative estimate of the true impact of the intervention on this outcome.

However, as with step count, small effects are nonetheless meaningful. Under sensitivity analysis, the average MVPA for participants in the pooled reward arms increased by a statistically significant 16.4 minutes. A recent analysis of four prospective cohort studies showed that accumulating just 22 minutes of MVPA per day eliminated the association between sedentary time and mortality, suggesting that even small increases in MVPA such as the ones observed in this pilot can have beneficial health impacts (Sagelv et al. 2023). Similarly, the 2018 Physical Activity Guidelines Advisory Committee Scientific Report suggests there is no lower threshold

for the benefits to risk of mortality associated with MVPA, and that an increase of 30 minutes alone can reduce this risk by 15-17% (Office of Disease Prevention and Health Promotion 2018).

7.3.3 Diet

As noted, the intervention had small positive effects on some dietary outcomes, with the pooled reward arms consuming 21g more fruit and vegetables on average than those in control. It also led to a small statistically significant increase in fibre intake (treatment effect: 0.4g). No comparable effects were found on the consumption of saturated fat and free sugars, though descriptive effects were in the expected direction (reductions), in line with our hypotheses. Qualitative findings from participants' accounts indicated a range of positive dietary changes in response to the intervention, including healthier food choices, increased awareness of portion sizes, and nutritional content.

As discussed, analysis of challenge data suggests that diet challenges were selected more often. While there were five times as many diet challenges as PA challenges (which is likely to account in part for their popularity) participants were also more engaged with challenges linked to diet, which were less likely to be abandoned than PA challenges. As a result, they were also more likely to improve on relevant outcomes. Indeed, the diet challenge completed most frequently by participants was designed to increase intake of fruit and vegetables, which is also where we observed the most substantial and significant impacts.

The effect we find on fruit and vegetable consumption is similar to that found in a small study examining the effect of financial incentives to encourage exploration of healthy dietary choices, which found that participants in the incentivised exploration group increased their salad consumption by three salads per week on average (SD = ±1.7) compared with two salads per week on average (SD = ±1.8) in the control group (Shavit, Roth, and Teodorescu 2021). Similarly, a quasi-experimental study found that fruit and vegetable intake increased as a result of receiving vouchers at farmers markets (1.4 servings per 1,000 kcal, $p < .001$) and supermarkets (0.8 servings per 1,000 kcal, $p = .02$; Purnell et al. 2014).

While the effects we find are small, existing evidence suggests even small improvements to diet can be beneficial to health. The average fruit and vegetable

intake for those in the pooled reward arms increased to 187g. This is just shy of 50% of the 5 A Day minimum recommended intake, at 400g. Finally, interpretation of diet outcomes is limited by the possibility that these data were under-reported. In comparison to the average intake recorded by the NDNS, the average energy intake of the pilot sample was very low (1207 kcal/d).

However, in the context of the objectives of the study, the pilot has successfully demonstrated that financial incentives have some potential to encourage people to at least move closer to their recommended intake.

7.4 How does the impact vary over time?

While the impact of financial incentives on PA and dietary outcomes varied over time, findings generated by these analyses were inconclusive. While MPVA effects were most pronounced at three months, the effect on step count declined with time. Diet-related effects showed either a relatively stable or slightly increasing trend, with stronger descriptive effects observed at three months compared to one month. These inconsistencies are likely to reflect the impact of differential drop out over time, as the composition of the sample differed at each point of analysis. Given this, we cannot draw any firm conclusions from these data.

7.5 What is the optimal financial incentive level?

The size of the financial incentive offered appears to increase the magnitude of the effect on behaviour on dietary outcomes, while this did not appear to be the case for PA outcomes. Overall, participants in the medium and high reward arms were more likely to complete challenges than those in the low reward arm, or those in the control arm. Larger incentives appeared to lead to bigger impacts for dietary outcomes. While the direction of the descriptive effects we see for sugar intake and fibre are as anticipated (reductions and increases respectively in line with the level of incentives offered), a positive, significant relationship is observed between all incentive levels for fruit and vegetable intake, alongside a significant negative relationship between high rewards and saturated fat intake. These translate to relatively small substantive effects; the difference for fruit and vegetables in the high

reward arm is equivalent to an increase on control of 16g, 18g and 29g per day in the low, medium and high reward arms respectively. The reduction in saturated fat seen in the high reward arm is equivalent to a 0.4pp reduction in intake compared to control.

This may be linked in part to the kind of diet challenges selected and completed by participants. The greater the incentive level participants were allocated to, the more diet challenges they selected and completed, and this pattern was also stronger for harder challenges.

The effect of incentive levels on PA outcomes is less clear. While the descriptive findings suggest all pooled reward arms performed marginally better than control, this effect is greater for those allocated to the medium reward arm, with a positive significant effect detected on step count equivalent to a substantive increase of 617 steps per day compared to control. This effect was found to be consistent across nearly all timepoints and subgroups, and it appears to largely account for the effect (albeit non-significant) we see in the pooled treatment analysis (increase of 42 steps per day). It also exceeds the 500 step increase required to reduce the risk of CV mortality by 7% (Banach et al. 2023).

It is possible that the findings are linked to the way PA challenges were rewarded for participants. At the point of randomisation, all groups' baseline step count was greater than 7000, with the lowest baseline recorded for the high reward group at 7303 steps per day. While increased effort generally leads to greater points for challenges, the app did not encourage participants to exceed 8000 steps per day once this goal had been reached. Given participants in each group were already approaching this target, the challenge design may have de-incentivised participants to boost their steps further. It may also have encouraged them to substitute PA challenges for diet challenges, where they were better able to earn points and rewards due to having a wider range of options for these challenges.

The highest step count at baseline was found in the medium incentive group, who averaged 7517 steps per day at the point of randomisation. This may indicate that this group had higher intrinsic motivation than other groups. In this case, it is possible that the positive statistically significant effect found for this group is a reflection of this, or that their high level of motivation made this group more likely to respond to the intervention. Lastly, it is also possible that this result is due to chance. In the

absence of other indicators, as well as the non-significant results of the primary analysis, this finding on its own does not provide robust evidence for the medium level of incentivisation being optimal.

The substantial significant effect we see on daily step count in the medium reward arm is consistent with the pattern of engagement in terms of participants' preference for PA challenges, up to a point. As with the pattern we see for diet challenges, the incentive structure was positively correlated with PA challenge selection; participants in the control and low reward arms of the trial were less likely to both select and complete PA challenges than those in the medium or high reward arms, regardless of the challenge level.

Overall, it appears that the incentive structure did motivate people to continue participating, as churn-out was less likely in high and medium reward arms. Participants in the high reward arm were less likely to churn-out of the pilot, and more likely to engage with the app. The pilot identified distinct patterns of engagement with the intervention that indicated the incentive structure did, as expected, have an immediate effect on whether people used the app, and for how long. While user engagement was relatively high across pilot trial arms (at 61% on average), participants in the high reward arm engaged with the app more consistently, and accumulated more points.

The pilot retention rate followed a similar pattern, with retention highest in the high reward arm, followed by the medium, low and control trial arms of the pilot (69%, 65%, 60% and 54% respectively). As anticipated, this was particularly the case for participants from more deprived areas, who were less likely to churn out if assigned to the high reward arm than all other trial arms of the pilot.

7.6 Did all participants respond equally to financial incentives?

Subgroup analysis by level of deprivation and level of inactivity suggest the intervention was particularly effective for more deprived participants, and those who began with a lower baseline level of health.

The majority of physical activity and dietary effects appeared to be concentrated among older participants, specifically those aged 41 and above. For this group,

financial incentives resulted in a series of statistically significant effects; an increase of 1.5 minutes per day in MVPA, a 0.4g increase in fibre consumption, and a reduction of 0.4 percentage points in saturated fat intake. While younger participants (aged below 41) achieved a significant increase in their fruit and vegetable intake (15g), their counterparts managed a 26g increase. We note that this may be in part a reflection of the fact the pilot attracted a sample that was more likely to be aged 41 or over, leaving this group disproportionately represented in the pilot.

Some differential effects were seen by level of deprivation. Financial incentives were more effective in increasing fruit and vegetable intake among participants who were less deprived, with a significant increase of 25g per day, nearly 10g per day more than more deprived participants.

In contrast, incentives were more successful in promoting physical activity among more deprived participants, who comprised half of the final randomised sample. These differences were non-trivial; more deprived individuals in the pooled reward arms increased their daily moderate and vigorous physical activity (MVPA) by a statistically significant 1.7 minutes per day on average, compared to just 0.7 minutes for the less deprived group. This was despite the fact that the same group completed half as many PA challenges per week than their less deprived counterparts.

A similar trend was observed in daily step count, with the more deprived group showing an increase of 206 steps from a baseline of 7550, while the less deprived group saw a decrease of 47 steps per day from a baseline of 7865, though both were descriptive effects.

This pattern of effects could be driven by a number of factors. Over the course of the pilot, those who were the most income-deprived redeemed rewards of cumulatively greater value per person than those outside this category, despite earning fewer points per person, suggesting a higher motivation to convert points into rewards than those who were less income-deprived. Consistent with this, retention rates across different trial arms of the pilot indicate that participants from more deprived areas were also less likely to drop out when assigned to the high reward arm. Themes that emerged from the qualitative findings linked the motivation for participants from areas of high deprivation to the cost of living crisis

and rising food costs, a finding consistent with the selection of pilot rewards observed in the pilot, with the vast majority (83.7%) being 'Grocery Gift Cards'. For participants in this group, and in the context of the challenges they face, incentives were likely a highly motivating element of the pilot. Qualitative findings reflect the fact that participants in the low reward arm, particularly those who were less deprived, felt the rewards offered were not worth the effort required, so were less likely to be motivated by them.

This may also explain why our findings suggest more deprived groups were more responsive to physical activity incentives. Where changes to physical activity are relatively cheap or costless, changes to diet are not. For participants whose primary concerns are the affordability of food and utilities, making these changes may have been less feasible or desirable. For participants with a lower level of deprivation, this is less likely to have been a barrier.

The impact of financial incentives on physical activity was also more pronounced among participants who were initially inactive, performing less than 30 minutes of MVPA during the baseline week. In this subgroup, financial incentives led to a small significant increase of 2.1 minutes per day in MVPA, compared to a null effect observed for active users. Given this is equivalent to an increase of 14 minutes per week, it is a meaningful improvement on their baseline level of activity and as previously noted, likely to yield health benefits given that the most significant drop in mortality risk has been observed in those who increased their steps from a low starting point (Paluch et al. 2022).

7.7 Conclusions

Overall, our findings were mixed. While positive, statistically significant effects were found for two dietary outcomes, and sensitivity analysis of PA outcomes also yielded significant positive effects, observed effect sizes were small. A possible explanation for this limited impact could be the motivational influence of the wearable technology alone and the financial payments for providing data, overshadowing the additional impact of financial incentives. Another factor could be self-selection bias, with more motivated and active individuals over-represented in the pilot, and

already inclined towards health-conscious behaviours, diluting the apparent effect of the incentives.

However, the pilot successfully reached individuals whose behaviours did not meet official guidelines for activity and diet healthfulness, and while effect sizes were small, their wider pattern suggests incentives, at least for some outcomes, are effective at bringing people closer to guidelines for their diet and physical activity.

The results may also suggest that while financial incentives act as a nudge towards healthier behaviour, the magnitude and nature of these incentives, along with individual differences in how people value rewards, play a crucial role. The modest impact on physical activity and diet might indicate that the incentives offered were not sufficiently appealing, or were not structured in a way that effectively tapped into behavioural motivations.

While it was not possible during this pilot to quantitatively examine participants' levels of motivation, qualitative findings suggest that participants found the app and incentives helpful as a tool to change their behaviour. They suggest improvements in exercise and diet were mutually reinforcing, with those who observed initial progress in mental health and weight loss reflecting that this encouraged them to improve their diet and physical activity. IPE findings suggested that participants in households where more than one person participated in the pilot were also more likely to earn points, supporting the idea that the social element of the intervention acted as a driver for engagement. Fully engaged participants also reported feeling better, happier and in improved mental health, a finding reflected in the quantitative impact evaluation.

While changes observed are small, the findings do suggest that financial incentives can be particularly effective in motivating specific groups, such as more deprived individuals, previously inactive users, and older adults, to make positive changes in their physical activity and dietary habits. These differential effects hint at the broader influence of social determinants on health behaviours. This raises questions about accessibility, and suggests a need to better understand the barriers to healthier lifestyle that might be more pronounced for some groups; while overall those with a third-level degree were more likely to earn points than those without, the incentive structure appeared to play an important role in motivating and retaining participants, particularly those who were more deprived. This suggests that

interventions like financial incentives may be easier to navigate for some groups, they may also play an important role in reducing health inequalities for groups whose barriers to good health are intrinsically linked to financial challenges.

7.8 Recommendations

The rich qualitative and quantitative data gathered over the course of the pilot, in particular from interviews with participants, enables us to make a series of recommendations for future consideration across several key areas. We believe that each of the recommendations presented here is practicable at least in part, and that each if implemented would increase the likelihood of financial incentives successfully being used to encourage healthy behaviour in the future. Interviews conducted as part of qualitative IPE work encouraged participants to reflect on how they would improve the intervention, and whether there are specific elements that they would change. Meanwhile, findings and observations from both the impact evaluation and quantitative IPE work lead us to suggest some broader overall recommendations. We present both sets of recommendations as follows:

User-led recommendations *[specific, based on user interviews]*

1. Reach and initial engagement
2. Continued user engagement
3. Maximising impact
4. Implementation and scaling

Other recommendations *[broad, based on quantitative analysis]*

5. Other recommendations

7.8.1 Reach and initial engagement

Recommendations for increasing the reach of the intervention and early engagement with the platform

Given the success of the pilot intervention's reach, it is crucial for any future financial incentives scheme to reflect the core aspects that led to the success of the current approach, taking a **varied approach** to marketing and channels of reach, using both digital and traditional formats, including on the ground presence, and maintaining a **local focus** in materials and communications. In practice, this would involve engaging local stakeholders in the design and adaptation of marketing materials for

each area. While the pilot employed a significant budget to ensure sufficient reach, including paid social media marketing, a future intervention could **leverage networks** to spread awareness of the Financial Incentives Scheme with a lower budget requirement. This would be tailored to the local area, but could involve networks of schools, community groups, or religious organisations to spread awareness of the intervention. Using the NHS as a channel of reach, such as GPs recommending patients to access the Financial Incentives Scheme, would also ensure that the intervention is best able to reach those who would benefit from it. Due to ethical concerns, it was not possible to incorporate NHS referral into the pilot.

The importance of word of mouth recommendations and positive peer review could be enhanced by offering an **initial 'refer a friend' reward** for early engagers, where participants could receive points when a referral code they share is used. To encourage initial engagement in the intervention, marketing and communications should continue to be clear about what rewards are available through the intervention, and consider where this could be made more explicit, such as specifying the reward option of supermarket shopping vouchers. While marketing and communications mentioned rewards, some interview participants were unclear about what this would entail, suggesting that greater clarity could be provided, such as naming specific reward partners in marketing messaging. Some interview participants suggested that offering a **small immediate reward for signing up** would provide an initial impetus for potential participants to engage. While the payment offer for baseline survey completion represents an early incentive, some non-engager interview participants felt that an immediate reward would have encouraged them to sign up. This would leverage the ***reciprocity effect***, which suggests that people tend to feel obligated to do 'a favour' after receiving 'a favour'. In this case, receiving the reward could prompt participants to engage with the sign-up process.

To overcome concerns that prevented some participants from engaging in the intervention, marketing messaging could use **case studies** of pilot participants, such as testimonies from people who used the app. Case studies could be used as short videos used in digital marketing. These could overcome potential participants' concerns about shaming or pressure by emphasising the non-judgemental tone of the intervention, and highlight the simplicity of the app, its manageable time requirement, and the achievability of challenges. This could leverage the ***messenger***

effect by using relatable and credible messengers.

Finally, given the aim of the intervention is primarily to improve the health of those in more deprived and inactive groups, we suggest that future studies limit eligibility to those in more deprived groups, to ensure the user base for these interventions is a better reflection of need. In line with this, based on the pattern of attrition we see in the current pilot, developments could be made to better retain participants who are less likely to stay engaged, including men and those from non-white ethnic groups. These cohorts were underrepresented at every point during the pilot, relative to their size in the Wolverhampton population. More sophisticated targeting for interventions like these are likely to yield larger impacts where their value is greatest, increasing the cost efficiency of any further work in this area. Future iterations of this pilot/intervention should consider over recruiting these cohorts, or fine tuning the user experience in these early phases of the user journey to avoid drop out. Relatedly, it may be possible to target residents with higher risk and lower activity levels by enrolling them in specific incentive programmes offering more extensive and relevant incentives, and / or increasing awareness of existing incentives through focused campaigns.

7.8.2 Continued user engagement

Recommendations to improve sustained user engagement

User churn out between randomisation and the end of the study was relatively similar across all demographics. 57.9% of those who were randomised into one of the study conditions were still active users of the app at the M5 point (week 23). Importantly, the primary determinant of churn out appears to be the study arm that the user was assigned to (i.e. the levels of incentives they faced), rather than any demographic characteristic that a user possessed. However, it is worth noting that those who were less healthy (i.e. with lower recorded baseline levels of MVPA and/or fruit and vegetable consumption) were very active for slightly shorter period of time within the app than their healthier peers, while those who were younger were slightly more likely to churn out than older participants. This suggests that more could be done to encourage **younger** and **less healthy** participants to remain engaged for longer.

Participants in qualitative interviews suggested that the app should introduce **more reminders**, and should allow participants **to select the specific timing and frequency**

of reminders. While some participants requested more reminders, others disagreed with the proposal, suggesting that they might ignore too many reminders. Incorporating individual choices into timing of reminders would allow participants to tailor the app reminder schedule to suit their requirements. Given that the app included the option to amend the frequency of notifications, promoting greater awareness of this aspect of the app's functionality would be beneficial, as well as introducing the option for participants to schedule reminders for certain times. Some participants preferred push notifications, while others preferred reminders via email. Providing the option to choose how to receive reminders would encourage buy-in.

Some participants also reflected that they would have liked **more practical suggestions for how to achieve challenges**, such as recipe ideas, or recommendations for food options to select on the go or when eating out if participants are not always able to cook their meals. This would ***make it easy*** for participants to complete challenges, and also introduces the possibility of partnerships with restaurants to promote certain healthy options on the menu, or partnering with food content creators to develop healthy recipes. Participants who were discouraged by failing challenges proposed that the intervention could include an **option to extend the challenge by a day** in order to complete it, rather than failing. 'Extending the challenge' could be a reward that participants could purchase with a small number of points.

Many participants reflected that introducing an **optional social aspect** to the intervention, either within the app or through online forums would encourage sustained engagement with the intervention. This could comprise online fora, and message boards, either within the app or as private groups on existing social media platforms. These spaces could represent an environment for participants to share challenges and seek encouragement from others. It is important that this should be optional, as some participants felt that this would be off-putting to them, and might make them feel judged.

Recommendations for app features and functionality

To prevent potential churn-out participants from disengaging with the intervention, the intervention could prompt participants who have not opened the app in a while with a ***timely*** reward offer to prompt re-engagement. This offer could be time-limited or use scarcity cues to promote uptake. Specific ***timely*** **wearable fitness**

tracker-related reminders delivered through the app reminding participants in the evening to charge their wearable fitness tracker overnight, or reminding participants in the morning to wear their wearable fitness tracker would reduce the impact of the barrier of participants forgetting to charge or wear their wearable fitness tracker. ***Timely reminders*** at **mealtimes** for participants to take photos of food when required for challenges would minimise the likelihood of participants forgetting to take photos of food. **Positive messaging and encouragement,** leveraging positive framing and gain-frame messaging would encourage participants to remain engaged with the intervention through positive reinforcement when participants achieve goals, or encouragement to continue if participants have not completed challenges.

Ensuring that participants can easily access customer support would reduce the likelihood of drop out due to technical issues. The app could include a **chatbot function** for faster access to support than via email, or **send timely prompts for participants to seek support** if their activity suggests that they might be experiencing issues. Additionally, introducing **automatic wearable fitness tracker syncing** would reduce the burden on participants to sync their devices.

Recommendations for improving engagement with challenges

Introducing a **wider range of challenge types and challenge difficulties**, including more exercise challenge options, would reduce the likelihood of drop out due to boredom with repeated challenges, or concern about challenge difficulty. Some participants reflected that they would have liked a greater range of exercise challenges. Offering participants the option to **intensify and extend challenges**, within the week and across multiple weeks, would encourage habit formation and promote the ***opportunity*** for behaviour change. Consolidating habits could be encouraged by offering higher point availability for selecting the same challenge at a repeatedly higher intensity. This would also increase the likelihood of longer-term changes.

This could also incorporate a greater degree of **personalisation of challenge availability** based on participants' baseline diet and levels of physical activity. Asking participants to **plan** when they intend to complete elements of challenges would increase the likelihood of action, through promoting ***implementation intention***. This could involve participants selecting which days they plan to make food swaps or exercise, and could be supported with timely reminders.

7.8.3 Maximising impact

Some intervention participants felt that offering the **option for feedback** on participants' food recording, particularly on portion sizes could increase participants' knowledge of appropriate diets. This could involve a pop-up after participants report their diet via Intake24 or upload photos of meals to indicate whether the portion is 'above average', 'average' or 'below average'. In addition to promoting engagement with the app, the introduction of an **optional social aspect** could increase participants' ***motivation*** for behaviour change. A social element could take a more competitive or supportive format, and could be for participants' peers, or allow connections to other anonymous participants. Moreover, as well as increasing engagement, **positive and encouraging messaging** would provide additional ***motivation*** for participants to pursue the desired activities. The recommendations of incorporating a social aspect and positive messaging have been explored further in Section 7.8.2.

Recommendations for tailoring the intervention to maximise impact

While most participants viewed the app as adequate for its function, many suggested that the design could be improved. This could include the option for participants to customise aspects of the app to promote a sense of user ownership. The intervention could incorporate more variation and **personalisation** in how participants engage with the app. This could involve initial screening to identify participants as specific personas, then targeting app design accordingly. For example, this could incorporate:

- An optional leader board based on points accumulated or step count for 'Competitive' personas
- More support and information for 'Learning a lot' personas
- More encouragement or indications of achievements so far for 'Repeated attempts' personas - this could leverage the ***endowed progress effect*** *(a situation in which people are more motivated to complete a task or achieve a goal if they believe they have already made some progress towards it)*.

The intervention should allow participants to **select the level of complexity for information** and challenge tips, to provide the option for more complex information provision. Increased knowledge was a valuable mechanism of impact for some

participants, yet many reported that the information provided was straightforward and basic. Offering all participants the opportunity to engage with information at their desired level of complexity would enhance participants' ***capability*** to improve their diet and physical activity. Additionally, allowing participants to select more complex and more novel information may reduce the likelihood of disengagement due to boredom with familiar information. Some participants also suggested that information should be presented differently in general, using more visuals or videos rather than text. This could be incorporated by presenting information pages and challenge tips as videos rather than through a text-based format, or by employing more photos, diagrams, and visuals when presenting information.

The differences we found in engagement and impact across demographic groups hints at the influence of population characteristics and social factors on health behaviours. This highlights the importance of tailoring the intervention to address barriers to healthier lifestyles that might be more pronounced for some groups, and particularly the less affluent communities the intervention is designed to help.

7.8.4 Implementation and scaling

A scaled intervention could incorporate a wider range of challenges, including challenges focused on alcohol consumption, and challenges about sleep. As the wearable fitness tracker already records sleep, this would be feasible to monitor. A scaled intervention could also incorporate streamlined diet surveys to record dietary intake, as there may no longer be a requirement to use Intake24. Participants reflected that a less laborious process for recording food intake would be simpler and quicker to use.

Recommendations for improving rewards

Some participants felt that reward redemption should be designed to allow participants to withdraw all points that they have accrued, rather than limiting them to multiples of £5. This could include the option to donate 'remainder' points to charity. The intervention could also allow participants to build up to greater rewards across multiple weeks, encouraging longer-term engagement. While most participants expressed satisfaction with the rewards available, some suggested that the intervention should include a wider range, such as a greater range of sports clothes retailers.

Supermarket reward partners recommended that creating a bespoke reward card within the supermarket, rather than individual voucher codes would simplify the process of reward delivery, and minimise the additional customer support requirement of working with a third party provider. However, this would require greater upfront resource commitment by supermarket reward partners. As part of scaling the intervention, we recommend conducting further research into what food is purchased with vouchers, and whether greater restrictions to limit vouchers to the purchase of only healthy food would be feasible.

Recommendations to further minimise fraud

While fraud detection processes during this pilot were strong, using NHS logins would further minimise the risk of fraud by allowing identity verification.

7.8.5 Recommendations for future research

The incentive structure developed for this intervention requires further consideration in the next iteration of its development. The current pilot did not identify an optimal incentive level for the encouragement of healthy behaviours, and there is great value in identifying the point at which any benefit generated from financial incentives may plateau. The influence of wearable technology as an incentive in itself, as reported by participants in the pilot, suggests a potential area of focus. This technology may be impactful in its own right, acting as a continuous reminder or motivator for healthy behaviour, independent of financial rewards.

As described in section 6.2.2 above, we find a small but statistically significant positive impact of having another app user in a user's household on points earned. Behavioural science has found similar social and peer effects across a range of different domains, and as social creatures, our behaviours are influenced by our peers (e.g. Chung et al. 2017; Mema et al. 2022). The design of the incentive structure could potentially tap into peer influences by providing incentives for team-based performance of physical activities in particular. With team-based incentives, people are likely to pay attention to their teammates' exercise behaviours and might face social pressure to conform - or exert social pressure on their teammates. They might also wish to avoid feeling guilty for letting their teammates down, or choose to exercise more often if they see their teammates doing so. All of this suggests exploring the potential to allow individuals in such contexts to form teams, and provide team incentives on top of individual incentives.

The pilot also raises questions about the sustainability of behaviour change induced by financial incentives. This pilot was not designed to determine whether once the incentives are removed, individuals maintain their healthier behaviours, or revert to former ones. Understanding the longevity of the impact is an important factor in assessing the real effectiveness of these interventions.

8. Limitations

This section discusses the limitations to this study in terms of external validity and internal validity of the findings. It also covers unforeseen challenges to the pilot.

8.1 Generalisability (external validity) of the findings

Considerations about the external validity of the pilot (the generalisability of the findings) hinge on four points: **location and time; context; and user base.**

The pilot was conducted in a selected location - the local authority of Wolverhampton. Wolverhampton was chosen among a number of LAs expressing interest to join the pilot for its large population size. The fact that a third of residents are classed as being physically inactive and below national average for fruit and vegetable consumption were also contributing factors. There aren't obvious reasons why the Wolverhampton adult population recruited for this pilot would respond differently to financial incentives, compared to residents in other local authorities. However, it is worth flagging that piloting in other areas may lead to a different result, as different local areas carry different local histories and experiences. Furthermore, it is possible that local authorities with a more active, healthier, or wealthier population may respond differently to financial incentives.

The pilot ran between February to October with the majority of participants undertaking their challenges in the Spring/Summer months. There is the possibility that the way participants react to physical activities and dietary challenges is influenced by seasonality - especially outside weather conditions and temperature.

Last, the pilot assesses the effectiveness of offering financial incentives in the specific context of the Better Health: Rewards App. It is possible that specific features of the user journey could have enhanced or inhibited the effect of offering financial incentives. It is also possible that specific features of this evaluation (e.g. the fact that participants received payment for providing their data, or the fact that all participants could receive a free wearable fitness tracker) could intersect with the effect of the intervention - most likely diluting the salience of being offered a financial incentive. This may also be particularly relevant for those assigned to the low reward

arm. These participants were able to earn more through providing data than participating in challenges. Given this, the lowest threshold for incentives may need to be considered in future studies where there is a risk that the incentive to provide data is greater than the incentives offered to change behaviours related to diet and physical activity.

The findings apply to the population of participants who provided data for the analysis. These participants were disproportionately female and older, compared to the target population. This means that reruns of the pilot with a different user composition may find different results.

8.2 Robustness (internal validity) of the findings

Four main aspects of the study should be considered when assessing the internal validity of the study: data collection for PA outcomes; data collection for diet outcomes; attrition; and biases in the qualitative research.

8.2.1 Data collection for PA outcomes

Outcome data for physical activity were collected via a wearable fitness tracker to be synced by the participants with the Better Health: Rewards app. Participants could choose whether to wear the device or not. Approximately half of the participants who wore their wearable fitness tracker did so for most of the day (45% of all participants had an average wear time of more than 18 hours per wear day). Roughly 11% of participants wore their wearable fitness tracker on average for less than 6 hours per day. This proportion was higher than expected, based on preliminary data provided by HUL at the protocol drafting stage. Furthermore, during the pilot we learned that in certain cases PA data could also be captured through the user's mobile phone. It is not possible to distinguish between data collected through phones or data collected through smart wearable fitness trackers. Consequently, wearable time was not an accurate indicator of data validity, and filtering out observations based on this metric likely excluded many 'valid' observations. The sensitivity analysis presented in Section 5.3 shows that including days with less than 6 hours of wearable time substantially increases sample size, and results in a larger and more significant treatment effect.

The incidence of 'anomaly' PA reads in this study was compatible with other studies for the two primary outcomes, however there is some evidence that commercial wearable fitness trackers (Fitbit, Apple, Garmin, Better Health: Rewards) work better in measuring steps than MVPA, compared to research wearable fitness trackers (such as Axivity or ActiGraph) (Mair et al. (2021), even though the topic of reliability of commercial wearable fitness trackers has not been studied in depth so far. No users in this study used a research grade wearable fitness tracker. More details are available in Appendix K.

The analysis controls for the wearable fitness tracker used, effectively controlling for average differences in measurement errors or average differences between the type of people who use certain measurement devices. In addition, the filtering rules used in the analysis remove invalid / extreme reads. It is also worth noting that differences in the functionality of the trackers may have played a wider role in how participants used them, thus impacting wear time and data collection. So, although it can have a negative impact on sample size, it does not negatively impact the internal validity of the pilot.

However, if the wearable fitness trackers are inaccurate this can add noise to our analysis.

On measuring MVPA more specifically, for the Better Health: Rewards tracker, MVPA is only counted if the user turns on the exercise mode on their wearable fitness tracker. This means that some MVPA activity may have not been captured. It is also possible that more conscientious participants were more proactive to switch on the exercise mode and record MVPA data. While most participants provided 3 or 4 valid readings, it is possible that participants did some MVPA in the days during which they did not wear the device.
Last, it is worth noting that this study adopts a well-accepted but narrow definition for MVPA. Systematic reviews of MVPA levels across different populations have found that studies using MVPA generally define the cut-off for MVPA based on either heart rate (Hollis et al. 2017) or movements detected by a wearable fitness tracker (Steene-Johannessen et al. 2020). Our study defines the cut-off for MVPA using both elements, whereby an exercise must satisfy two conditions: Heart rates are "sufficiently elevated", i.e. hitting at least 50 % of the maximal heart rate; AND "sufficient movements" detected by the wearable fitness tracker. Examples of

activities not counted as “exercise” include brisk walking with low heart rate, gardening, yoga.

8.2.2 Data collection for diet outcomes

It is worth noting that many studies struggle to test the effectiveness of providing incentives to reward healthy diet behaviours due to challenges with outcome measurement. Consumption behaviours are often performed in private, and unlike physical activity behaviours which can be tracked using objective measures such as pedometers, tracking individuals' food and drink consumption in a naturalistic setting relies heavily on self-reported data.

Dietary outcomes were collected via a dietary recall questionnaire called Intake24. As described in the TP, based on previous studies, the user experience of completing Intake24 is generally positive and suggests that the majority of people find the system user-friendly, enjoyable to use, and felt it accurately captured their diet. However, this was not universally the case in this study - while some found it interesting, others reported that the process was laborious and long-winded. Despite having been validated in official studies, as all dietary recalls, Intake24 relies heavily on participants' memory and honesty, which can lead to underreporting or overreporting of food intake. This recall bias can significantly impact the accuracy of the data collected. Additionally, Intake24, while user-friendly, may not comprehensively capture all foods consumed, particularly those that are less common or have complex ingredients. Cultural and regional variations in food types and preparation methods can also pose challenges in accurately recording dietary intake.

8.2.3 Data collection for motivation

Surveys measuring participants' motivation were optional, and due to an implementation error, participants were not prompted to complete this survey at the month 5 data collection point. As a result, we did not have sufficient data to conduct any meaningful analysis on this outcome.

8.2.4 Differential attrition across trial arms

Differential attrition across pilot trial arms is a significant concern in RCTs as it can introduce bias and undermine the validity of the study results. When the rates of

dropout or loss to follow-up differ between intervention and control trial arms, it may lead to unequal representation and hence, distort the comparative effectiveness of the intervention. This imbalance can arise from various factors, such as differing levels of side effects, engagement, or satisfaction with the treatment. As expected, we experienced differential attrition. Attrition in the control arm was 14% higher than the pooled reward arms for PA outcomes, and 21% higher for diet outcomes. In Section 4.1.5, it is mentioned that to address this issue, the analysis employed inverse probability weighting (IPW). IPW involves assigning weights to the study participants based on their probability of remaining in the study, calculated using baseline characteristics and other relevant factors. By doing so, IPW helps to rebalance the trial arms, making them more comparable in terms of user characteristics and the likelihood of staying in the pilot. This statistical technique effectively adjusts for the attrition biases, providing a more accurate estimate of the intervention's impact by accounting for the differing dropout rates and ensuring that the analysis reflects the initial random assignment of participants to each pilot trial arm. However, this methodology has its own limitations, including model dependence, possible extreme weights and assumption of random attrition conditional on observable covariates. It is possible that attrition is related to unmeasured variables - to that extent, the IPW adjustment may not be sufficient to adjust for differential attrition across pilot trial arms.

8.2.5 Possible bias in qualitative research

Through our interviews and focus groups, we sought to better understand the range and diversity of experiences of intervention participants (both fully engaged participants and churn-out participants), intervention non-recipients, delivery partners, reward partners, and CWC. We aimed to sample intervention participants and non-engagers across a range of ages, ethnicities, levels of deprivation, and baseline BMI levels. We also sought to sample churn-out participants across a range of churn-out stages. However, despite using a purposive sampling strategy, we may not have captured the full diversity of perspectives, as our sample size was limited. For example, we did not speak to participants who dropped out during the baseline phase, so any consideration of barriers to engagement during this period is likely to be understated.

Additionally, while we aimed to include representation from all pilot trial arms across interview participants, we were not able to recruit churn-out participants from the

high reward arm. Insights from churn-out participants therefore derive only from the control arm, low, and medium reward arm participants. Moreover, some focus groups did not contain representatives from all pooled reward arms due to last minute drop out and non-attendance. For example, the fully engaged focus group included only high and low reward arm participants, while the churn-out focus group was not attended by the control arm participant. As a result, while the focus group generated valuable insights from participants across the full spectrum of engagement, some discussions This means that the group discussion may not consider the impact of different pooled reward arms in full, particularly the impact of being in the control arm.

Despite these limitations, the qualitative approach provides valuable in-depth insights into the experiences and perceptions of participants, offering a nuanced understanding of the Financial Incentives Scheme's implementation and impact.

8.2.6 Research bias

Researcher bias describes the inherent subjectivity brought by researchers to qualitative work, which can influence data collection, interpretation, and analysis. Researcher bias in qualitative data analysis was mitigated by employing peer debriefing: one researcher developed codes, which were validated by a second reviewer. The two researchers met to discuss any discrepancies and agreed on a code-book. Verbatim user quotations and case examples were used to provide evidence and exemplify the theme(s) discussed in the paragraph before the quotation. Any quotations used were selected by the qualitative researchers who conducted the data analysis based on how well they exemplified the theme(s) discussed.

9. Appendix

Appendix A: Eligibility Criteria

Below is a copy of the eligibility/screening criteria interface as it appeared to potential app users:

Is this app for you?

This app is only for people aged 18 and over who live in Wolverhampton. The app provides general health information to encourage a healthy lifestyle, based on UK Government diet and physical activity guidelines.

If you have any health concerns about changing your diet or physical activity, consult your local GP or health professional before starting this programme. Find more information to help you in deciding whether to use the app in our FAQs.

This app may not be suitable for people who have or have previously had an eating disorder, or any other health concern which might stop you from changing your diet or physical activity. If you are not sure whether the app is for you, you should consult your GP or health professional before starting this programme.

I understand and meet the requirements above.

++

If **yes** □ proceed

If **no** □ Unfortunately you do not meet the criteria for participating in this programme. Please consult your healthcare provider for alternative pathways to meet your health and wellbeing objectives.

Appendix B: Detailed information about outcome variables

Table B1: Detailed information about outcome variables

Outcome	Type	Technical Information
MVPA (minutes per day)	PA - Primary	This metric is directly produced by the wearable fitness tracker. It provides a daily read, is positive, and continuous. MVPA is calculated by summing the minutes spent in both moderate and vigorous physical activity, identified by the device's accelerometers and the user's heart rate. Activities are classified as "vigorous" when the heart rate reaches cardio or peak zones, and "moderate" when within fat-burning zones and accompanied by sufficient movement. Users of the Better Health: Rewards tracker need to activate an 'exercise mode' for accurate MVPA tracking.
Steps (steps per day)	PA - Primary	Steps are measured through the wearable fitness tracker or user's phone, recording the total number of steps taken each day. This metric provides a continuous, daily measure of physical activity that is directly incentivized by the app through specific challenges. It is considered a reliable indicator of overall physical activity levels and a proxy for sedentary behaviour.
Fruit and Vegetable Intake (grams per day)	Diet - Primary	The intake of fruits and vegetables is quantified in grams per day based on user inputs through the Intake24 questionnaire. This online research tool facilitates the self-reporting of all foods and drinks consumed over a 24-hour period, from which the intake of fruit and vegetables is calculated.
Fibre Intake (grams per	Diet - Primary	Fibre intake is measured in grams per day, derived from dietary data collected via the Intake24

day)		questionnaire.
Free Sugar Intake (% of daily food energy intake)	Diet - Primary	Free sugar intake is assessed as the percentage of total food energy intake derived from free sugars, based on dietary data collected through Intake24.
Saturated Fat Intake (% of daily food energy intake)	Diet - Primary	Saturated fat intake is measured both as a percentage of total food energy intake derived from saturated fats, using data from the Intake24 questionnaire.
Energy Expenditure (kcal per day)	PA - Secondary	Energy expenditure is defined as the total calories burned from physical activity each day, measured via the wearable fitness tracker. This secondary outcome complements primary physical activity measures by providing a broader assessment of overall energy expenditure, including both moderate, vigorous, and light-intensity activities.
Energy Intake (kcal per day)	Diet - Secondary	Energy intake is calculated in kilocalories per day, based on 24-hour dietary recall data collected through Intake24. This measure provides a comprehensive overview of the total energy consumed, allowing for assessments of dietary adequacy, energy balance, and the risk of over- or undernutrition.

Healthy Eating Score	Diet - Secondary	The Healthy Eating Score is constructed based on dietary intake data collected through the Intake24 questionnaire. Following the methodology outlined by Scheelbeek et al. 2020, it assesses participants' diets against specific nutritional criteria. The scoring system allows for a range from 0 (least healthy) to 7 (most healthy), with points awarded as follows: 1. .Fruit and Vegetables: ≥ 400g/day. 2. Red and Processed Meat: ≤ 70g/day. 3. Free Sugars: ≤ 30g/day. 4. Saturated Fat: ≤ 30g/day for males, ≤ 20g/day for females. 5. Fibre: ≥ 30g/day. 6. Total Fat: For males, ≤ 97g (18-64), ≤ 91g (65-74), ≤ 89g (75+); for females, ≤ 78g (18-64), ≤ 74g (65-74), ≤ 72g (75+). 7. Salt: ≤ 6g/day. Each metric met by a participant scored 1 point, with a maximum of 7. Per the evaluation plan, we excluded two metrics used by the Scheelbeek et al. methodology: oily fish and other fish consumption. We did this as these are expressed on a weekly basis, and we only worked with 24 hours dietary recalls.
Weight (kilograms)	Secondary	Weight is self-reported by participants at baseline and each measurement point through the study app. Participants are encouraged to weigh themselves for accurate data entry. This measure allows for monitoring changes in body weight over time.

Mental Well-being	Exploratory	The WHO Mental Well-being score is constructed using the WHO-5 Well-being Index, a validated questionnaire consisting of five statements related to general feelings of well-being over the past two weeks. Respondents rate each statement on a 6-point scale from 0 ("at no time") to 5 ("all of the time"). The total raw score, ranging from 0 to 25, is then multiplied by 4 to produce a final score between 0 (worst possible well-being) and 100 (best possible well-being). This score allows for the assessment of mental well-being at baseline and subsequent data collection points, offering insights into the intervention's broader health impacts. The specific instructions are worded as follows: "Please indicate for each of the five statements which is closest to how you have been feeling over the last two weeks. Notice that higher numbers mean better well-being. Example: If you have felt cheerful and in good spirits more than half of the time during the last two weeks, put a tick in the box with the number 3 in the upper right corner." The specific questions are reported in Table B2.
Sleep Duration (hours per night)	Exploratory	Sleep duration is measured in hours per day, recorded by the wearable fitness tracker for participants willing to wear it at night. This exploratory outcome assesses the potential impact of physical activity and dietary changes on sleep quality and duration.

Table B2: WHO (Five) Well-Being Index

Over the last two weeks:	All the time	Most of the time	More than half of the time	Less than half of the time	Some of the time	At no time
1. I have felt cheerful and in good spirits	5	4	3	2	1	0
2. I have felt calm and relaxed	5	4	3	2	1	0
3. I have felt active and vigorous	5	4	3	2	1	0
4. I woke up feeling fresh and rested	5	4	3	2	1	0
5. My daily life has been filled with things that interest me	5	4	3	2	1	0

Appendix C: Balance checks

Table C1: Proportion of participants by demographic characteristics at baseline and at the analysis stage (PA and diet), split by intervention arm

	Baseline				M5 (PA)				M5 (Diet)			
Variable	**Control**	**Low reward**	**Medium reward**	**High reward**	**Control**	**Low reward**	**Medium reward**	**High reward**	**Control**	**Low reward**	**Medium reward**	**High reward**
Ethnicity (%)												
Asian / Asian British	18%	16%	19%	17%	19%	18%	21%	16%	18%	18%	20%	18%
Black / African / Caribbean / Black British	7%	6%	8%	7%	7%	5%	8%	8%	7%	5%	7%	7%
White	68%	72%	64%	69%	68%	73%	64%	69%	68%	71%	63%	68%
Mixed / Multiple ethnic groups	3%	3%	5%	4%	3%	3%	4%	4%	3%	3%	5%	4%
Other ethnic group	1%	1%	2%	1%	1%	1%	1%	1%	1%	1%	2%	1%
Private	2%	2%	2%	2%	2%	2%	2%	2%	2%	3%	2%	2%
Gender (%)												
Female	67%	67%	66%	67%	70%	67%	69%	70%	68%	68%	67%	67%
Male	32%	33%	33%	33%	30%	33%	30%	30%	32%	31%	32%	33%
Other	0.3%	0.3%	0.7%	0.5%	0.0%		0.7%		0.1%	0.1%	0.5%	0.1%
Age group												
Above median (41 years)	49%	50%	47%	50%	65%	64%	61%	63%	54%	54%	50%	54%

	Baseline				M5 (PA)				M5 (Diet)			
Variable	**Control**	**Low reward**	**Medium reward**	**High reward**	**Control**	**Low reward**	**Medium reward**	**High reward**	**Control**	**Low reward**	**Medium reward**	**High reward**
Median or below	51%	50%	53%	50%	35%	36%	39%	37%	46%	46%	50%	46%
Deprivation index												
Least deprived (IMD decile >2)	49%	49%	50%	48%	55%	57%	56%	50%	51%	53%	51%	47%
most deprived (IMD decile ≤2)	49%	49%	49%	51%	44%	42%	43%	49%	47%	46%	47%	51%
unknown	1.5%	1.5%	1.2%	1.4%	1%	1%	1%	2%	1.5%	0.9%	1.4%	2.0%
Wearable Fitness Tracker												
Better Health: Rewards	72%	68%	73%	71%	63%	59%	69%	67%	70%	66%	73%	69%
Fitbit	9%	11%	9%	8%	16%	18%	13%	11%	11%	12%	9%	8%
Garmin	2%	3%	2%	2%	4%	4%	2%	4%	2%	3%	1%	3%
GoogleFit	3%	3%	2%	3%	1%	1%	1%	1%	3%	3%	3%	3%
HealthKit	14%	16%	14%	15%	16%	18%	15%	17%	14%	16%	14%	17%

Table C2: Mean of continuous variables at baseline and at the analysis stage (PA and diet), split by intervention arm

	Baseline				Active at M5 (PA)				Active at M5 (Diet)			
Variable	**Control**	**Low reward**	**Medium reward**	**High reward**	**Control**	**Low reward**	**Medium reward**	**High reward**	**Control**	**Low reward**	**Medium reward**	**High reward**
Deprivation	3.5	3.4	3.5	3.4	3.8	3.8	3.8	3.4	3.6	3.6	3.6	3.3

index												
Baseline fibre intake (in g)	10.8	10.8	10.6	10.8	11.4	11.4	11.1	11.2	12.6	12.4	12.3	12.7
Baseline free sugars intake (as % on energy intake)	11.1%	10.7%	11.3%	11.1%	10.1%	10.3%	11.0%	10.4 %	10.9%	10.8 %	10.7%	11.1%
Baseline fruit and vegetable intake (in g)	148	154	150	151	169	169	168	165	151	155	158	154
Baseline saturated fat intake (as % of energy intake)	11.5%	11.5%	11.4%	11.4%	11.6%	11.7%	11.7%	11.3 %	11.6%	11.4 %	11.5%	11.5%
Baseline MVPA (average min / day)	11.5	13.4	13.1	11.4	15.7	17.7	16.0	14.7	13.0	15.3	11.5	11.9
Baseline steps (average steps / day)	7434	7496	7517	7303	7439	7491	7627	7472	7365	7585	7589	7156

Appendix D: Additional tables on reach and engagement

Table D1: Characteristics of participants randomised and providing consent for data usage, disaggregated by subgroup categories

Full sample	**N = 12,767**	**100%**
Ethnicity category	**n**	**% of full sample**
White	8,712	68%
Asian or Asian British	2,270	18%
Black, African, Caribbean, or Black British	874	7%
Mixed or Multiple Ethnic Backgrounds	459	4%
Other ethnic group	171	1%
No data	281	2%
Gender category	**n**	**% of full sample**
Female	8,569	67%
Male	4,153	33%
Other	45	0.4%
Deprivation (IMD) category	**n**	**% of full sample**
Most deprived *(IMD score of 1 or 2)*	6,298	49%
Other	6,287	49%
No data	182	1%
Age category	**n**	**% of full sample**
Aged 45 or over	5,382	42%
FruitVeg (baseline) category	**n**	**% of full sample**
High FruitVeg consumption at baseline	2,618	21%
Low FruitVeg consumption at baseline	9,991	78%
No data	158	1%
MVPA (baseline) category	**n**	**% of full sample**

High MVPA minutes at baseline	3,952	31%
Low MVPA minutes at baseline	8,815	69%
Study arm	**n**	**% of full sample**
Control	7,791	61%
Low reward	1,900	15%
Medium reward	1,532	12%
High reward	1,544	12%

Notes: Due to rounding, not all percentage figures within subgroups add to 100%.

Table D2: Characteristics of participants still active at M5, disaggregated by subgroup categories

Full sample	**N = 7,387**	**100%**
Ethnicity category	**n**	**% of full sample**
White	4,967	67%
Asian or Asian British	1,362	18%
Black, African, Caribbean, or Black British	535	7%
Mixed or Multiple Ethnic Backgrounds	256	3%
Other ethnic group	101	1%
No data	166	2%
Gender category	**n**	**% of full sample**
Female	5,024	68%
Male	2,347	32%
Other	16	0.2%
Deprivation (IMD) category	**n**	**% of full sample**
Most deprived *(IMD score of 1 or 2)*	3,536	48%

Other	3,745	51%
No data	106	1%
Age category	**n**	**% of full sample**
Aged 45 or over	3,409	46%
FruitVeg (baseline) category	**n**	**% of full sample**
High FruitVeg consumption at baseline	1,575	21%
Low FruitVeg consumption at baseline	3,949	53%
No data	1,863	25%
MVPA (baseline) category	**n**	**% of full sample**
High MVPA minutes at baseline	2,454	33%
Low MVPA minutes at baseline	4,933	67%
Study arm	**n**	**% of full sample**
Control	4,194	57%
Low reward	1,146	16%
Medium reward	989	13%
High reward	1,058	14%

Notes: Due to rounding, not all percentage figures within subgroups add to 100%.

Table D3: Attrition at M0, M1, M3 and M5, split by intervention arm

		Participants with valid data					Attrition rate			
	N at randomisation	**At M0**	**At M1**	**At M3**	**At M5**		**At M0**	**At M1**	**At M3**	**At M5**
MVPA										
High reward	1,544	1,344	971	698	595	**High reward**	13%	37%	55%	61%

Medium reward	1,532	1,375	912	610	561	**Medium reward**	10%	40%	60%	63%
Low reward	1,900	1,677	1,096	731	666	**Low reward**	12%	42%	62%	65%
Control arm	7,791	6,970	4,211	2,419	2,165	**Control arm**	11%	46%	69%	72%
Total	*12,767*	*11,366*	*7,190*	*4,458*	*3,987*	***Total***	*11%*	*44%*	*65%*	*69%*
Diet										
High reward	1,544	1523	975	935	905	**High reward**	1%	37%	39%	41%
Medium reward	1,532	1512	905	906	835	**Medium reward**	1%	41%	41%	45%
Low reward	1,901	1875	1031	1010	952	**Low reward**	1%	46%	47%	50%
Control arm	7,792	7699	4018	3616	3457	**Control arm**	1%	48%	54%	56%
Total	*12,769*	*12609*	*6929*	*6467*	*6149*	***Total***	*1%*	*46%*	*49%*	*52%*

Notes: Attrition is defined as the % of randomised participants who have not provided valid observations at a given time point. M0 is the baseline measurement before randomisation.The attrition at this time point represents participants who were randomised despite providing 'invalid' data.

Appendix E: IPW methodology

The IPW works by modelling the probability of successful retention at 1 month, 3 months and 5 months separately, using participants' baseline observables data and re-weighting observations for participants that were retained. In this way, the final reweighted data are re-balanced in terms of baseline observables across different trial arms.

To do this, we firstly ran a regression model with response as the binary outcome variable, for each of the outcome variables. The general equation for the response model for MVPA and steps is as follows:

$$Y_i \sim bernoulli(p_i)\ ;\ logit(p_i) = \alpha + \beta_{low} Treatment_{low,i} + \beta_{medium} Treatment_{medium,i} + \beta_{high} Treatment_{high,i} + \gamma_X X_{ij} + u_i$$

where:

- y_i is a dummy variable, taking value 1 for if individual *i* is included in the regression for the relevant outcome variable, and 0 if they have dropped out;
- $Treatment_{x,j}$ is a dummy variable taking value 1 for households allocated to reward arm x, 0 otherwise;
- X_{ij} are individual level and family level covariates that are *known at baseline*. This includes household size, deprivation index score, level of education, BMI at baseline, ethnicity, gender, age at baseline, device used, mean baseline steps, and mean baseline MVPA.
- u_i is an idiosyncratic error term, where $u_i \sim N(0, \sigma^2_u)\ for\ all\ i$.

As attrition is different for different outcome variables, we ran this regression for each outcome variable separately. We used the same equation for the each outcome, but with small differences in the covariates included in X_{ij}. For example, for the diet outcome variables we didn't include the mean baseline steps and MVPA. Instead, we included the number of baseline Intake24 submissions, fruit & vegetable intake, fibre intake, free sugar intake as % of total energy intake, saturated fat intake as % of total energy intake, total energy intake, and healthy eating score (all at baseline) as covariates.

We then used these models to predict whether the user would have a valid response for the relevant outcome variable. This prediction model outputs a probability between 0 and 1 for each user. The inverse of these probabilities were then used as weights in the final regression.

Appendix F: IPE Research topics and methods

Table F1: Research topics and questions for qualitative analysis (IPE)

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 was the intervention perceived by recipients, and non-recipients of the financial incentives?
	1.6	What were the barriers and facilitators to the intervention's reach?
Engagement What factors affected the engagement of participants with the intervention?	2.3	In what ways did participants engage with the app after signing up?
	2.4	What were participants' experiences and perspectives of the intervention?
	2.5	What were the barriers and facilitators to engagement?
Mechanisms of impact Through what mechanisms did the intervention affect behaviour change?	3.2	What barriers and facilitators - both contextual and individual - affected the extent to which the intervention changed behaviours for recipients and non-recipients of the financial incentives?
	3.3	How did features of the Financial Incentives Scheme affect (or not) the extent to which the intervention changed 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?

Since each population had 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.

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.

Sampling for recipients and non-recipients

Using purposive sampling we selected participants based on two primary characteristics: their level of engagement with the Financial Incentives Scheme (fully engaged, churn-out (post randomisation) or non-recipient) and the reward arm (control and level of rewards).

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

Table F2: Sampling frame for qualitative data collection

Sample:38	**14 Fully engaged intervention recipients:** • 12 interviews with intervention recipients (3 per reward arm) • 1 focus group with intervention recipients (2 participants - 1 high arm, 1 low arm)

	12 Churn-out intervention recipients: • 9 Interviews with intervention recipients (2 low arm, 4 medium arm, 3 control arm) • 1 focus group of intervention recipients (3 participants - 1 low arm, 2 medium arm) **12 Non-recipients:** • 9 Interviews • 1 focus group (3 participants)	
Primary criteria		**Interviewed**
Financial Incentives Scheme participation	Fully engaged	14
	Churn-out	12
	Non-recipients (recipient targeted by HUL to join the Financial Incentives Scheme but who did not sign up)	12
Intervention Arm	Control (no financial incentive)	6
	Low incentive	7
	Medium incentive	9
	High incentive	4
	Non-recipients (recipient targeted by HUL to join the Financial Incentives Scheme but who did not	12

	sign up)	
Secondary criteria (desirable)		**Recruited**
Baseline BMI	BMI ≥ 25	Fully engaged: 7 Churn-outs: 9 Non-recipients: 6
	BMI < 25	Fully engaged: 7 Churn-outs: 3 Non-recipients: 4
	Missing BMI data	Non-recipients: 2
Gender	Male	Fully engaged: 7 Churn-outs: 4 Non-recipients: 2
	Female	Fully engaged: 7 Churn-outs: 8 Non-recipients: 10
	Non-binary	Fully engaged: 0 Churn-outs: 0 Non-recipients: 0
Ethnicity	White	Fully engaged: 10 Churn-outs: 9 Non-recipients: 2
	Black	Fully engaged: 1 Churn-outs: 2 Non-recipients: 4
	Asian	Fully engaged: 3 Churn-outs: 1 Non-recipients: 1
	Mixed	Fully engaged: 0

		Churn-outs: 0 Non-recipients: 2
	Other	Fully engaged: 0 Churn-outs: 0 Non-recipients: 3
Deprivation	Index of Multiple Deprivation (IMD) ≤ 2	Fully engaged: 4 Churn-outs: 8 Non-recipients: 7
	Index of Multiple Deprivation (IMD) > 2	Fully engaged: 10 Churn-outs: 4 Non-recipients: 5

The sample size was based on an estimate of the number of participants required for the IPE to achieve a diversity of views and experiences, triangulation of findings and a comprehensive understanding of the barriers or facilitators to a behaviour under the Theoretical Domains Framework (TDF).

Recruitment of recipients and non-recipients

Intervention participants

The recruitment of intervention recipients (both fully engaged and churn-outs) for interviews and focus groups was led by HUL.

In order to encourage and enable participation, each interview user was provided with a £30 gift card by BIT at the end of the interview. Those that participated in a focus group were thanked with a £50 gift card. Provided a user attended the interview, they would receive the gift card regardless of whether or not they completed their interview. Participants were sent their £30 or £50 Tango e-gift card via email from a member of the BIT team. This is a virtual gift card that allows the recipient to spend a set monetary value on a variety of retailers, such as Amazon, Currys, PC World, Tesco, Cineworld or John Lewis.HUL was 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 selected participants who consented to be contacted by email and phone, and stratified participants into cohorts by the Financial Incentives Scheme participation level
2. For each cohort and participation level, participants' email addresses were randomly split into "focus group" (10%) and "incentive interview" (90%).
3. HUL sent an email inviting recipients (both fully engaged and churn-outs) to express an interest in participating in focus groups or interviews which included:
    - A sign-up form.
    - An information sheet that outlined the purpose of the interviews or focus groups, the topics that would be discussed and how their data would be protected. We did not need to collect written consent from interview participants as we collected and audio recorded verbal consent at the beginning of the interview.
4. Respondents were selected on a first-response basis providing they fulfilled sampling requirements set out by BIT:
    - Primary criteria ("hard" requirement): trial arms (including the control arm) and scheme participation level.
    - Secondary criteria ("soft" requirements): age group, gender, deprivation, ethnicity, and BMI (at sign-up) category.
5. Having been screened based on sampling characteristics and eligibility, HUL assigned a unique study code ID to participants which was used to link two separate, secure spreadsheets, containing:
    - Contact information for booking interviews and demographic information.
    - Interview data.
6. HUL contacted each new user to request available time slots for interviews/focus groups. Available times were shared with BIT researchers through a secure spreadsheet. BIT researchers then sent the user a Google

calendar invite including a video-conferencing link (and an alternative telephone number to dial in) from their BIT email account.

7. If the candidate did not respond to BIT, the next candidate from step 4 was selected, where possible and appropriate.

Response rates

HUL contacted 4092 fully engaged and churn-out participants for interviews. 94 candidates replied, leading to 21 conducted interviews.

HUL contacted 450 fully engaged and churn-out participants for focus group participation. 25 candidates replied, leading to 5 focus group participants.

Non-recipients (pure control)

The recruitment of intervention non-recipients (i.e., individuals that lived in Wolverhampton and were therefore exposed to the marketing and recruitment activities encouraging sign up to the Financial Incentives Scheme, but chose not to) was led by an external recruitment agency, Acumen following the recruitment process below:

1. BIT worked with DHSC and Acumen to identify local channels through which to recruit.
2. Acumen contacted individuals not involved in the pilot via an invitation email including:
    - A sign up form
    - An information sheet that outlined the purpose of the interviews or focus groups, the topics that would be discussed and how their data would be protected.
3. Having been screened based on sampling characteristics and eligibility, Acumen assigned a unique study code ID to participants which was used to link two separate, secure spreadsheets, containing:
    - Contact information for booking interviews.
    - Demographic information

4. BIT provided Acumen with a list of interview and focus group slots which Acumen shared with participants to select a convenient time. BIT then sent out a Google calendar invitation (including a telephone option) to participants and BIT researchers on the selected slot together with an email that also contained the information sheet.

Data collection method

Interviews for intervention recipients and non-recipients

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

The interviews with intervention recipients addressed all four research topics: reach, engagement, mechanisms of impact, and implementation. In particular, the interviews explored the interviewees' reasons for participating in the intervention, the ways in which the intervention 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, their interviews explored the reasons why the individual did not choose to participate in the intervention, including a thorough examination of their perceived acceptability of the intervention.

The interviews employed the TDF (Cane, O'Connor, and Michie 2012) to explore the capabilities, motivation and opportunities of participants in changing 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 provided sufficient time to explore each of these domains of behaviour, and, in doing so, helped identify barriers and facilitators to the Financial Incentives Scheme.

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

Focus groups with intervention recipients

Focus groups were conducted to complement the in-depth interviews by enabling an opportunity for discussion and the exchange of ideas between participants, helping participants to further develop their own ideas and tease out some nuances in their experiences and perspectives.

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

Focus groups lasted up to 90 minutes. Both focus groups and interviews were conducted by video call or telephone, and were audio recorded and transcribed.

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).

Sampling for Delivery Partners HUL

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

Table F3 below provides an overview of the primary criteria and achieved sample for this population.

Table F3: Sampling frame for Delivery Partners (HUL) Focus Group

Sample: 4		**4 x Semi-structured interviews**
Primary criteria		**Target**
Role/Job Title	UX and customer support	2
	Design and implementation	1
	Marketing	1

Recruitment of delivery partner

HUL appointed a recruitment lead who worked with BIT to recruit staff for individual interviews. The recruitment process was carried out as follows:

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

Data collection methods for delivery partners

Four individual interviews were conducted with HUL staff members to understand their experiences of implementing the intervention. Given HUL's role, these interviews provided 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.

Interviews lasted up to 90 minutes. They were conducted by video call or telephone and were audio recorded and transcribed.

Data collection with the reward partners

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

Sampling for reward partners

Using purposive sampling, we selected 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.

Table F4 below provides an overview of the primary criteria and achieved sample for this population.

Table F4: Sampling frame for reward partners

Sample: 7		**6 participants:** • 4 interviews with national supermarkets (3 individual, 1 paired) • 2 interviews with national gyms
Primary criteria		**Interviewed**
Type of partner	National supermarkets	5
	National gyms	2

Recruitment of reward partners

The recruitment of incentive providers for focus groups and interviews was led by HUL. They appointed an internal recruitment lead who worked with BIT to recruit participants. The recruitment approach was:

1. HUL's recruitment lead reached out to reward partners to invite them to express an interest in participating in focus groups. The email gave a summary of what focus groups would entail, and included:
    - A sign-up form
    - An information sheet that outlined the purpose of the interviews or focus groups, the topics that would be discussed and how their data would be protected.
2. Having screened participants based on the sampling and eligibility criteria, HUL assigned each user a unique study ID. This code and a secure spreadsheet with contact and sampling information was shared with BIT.
3. BIT researchers sent new participants a link to a Google Form through which they booked an interview slot using their user ID. A copy of the information sheet was also included.
4. The researcher then sent the user a Google calendar invite including a video-conferencing link (and an alternative telephone number to dial in) from their BIT email account.

Individual interviews with reward partners

Individual interviews were conducted with reward partners - both corporate and local - to understand their experiences of implementing the Financial Incentives Scheme. The interviews were as follows:

- **4 interviews** with national supermarkets
- **2 interviews** with national gyms

Individual interviews were used in place of focus groups to enable reward partners to speak freely and ensure commercial confidentiality, as reward partners expressed a reluctance to discuss company priorities and processes in a call with rival

organisations. One paired interview was conducted with two participants from the same organisation.

Interviews provided insights into the challenges of implementing the intervention, barriers and facilitators to the scalability and sustainability of the intervention, and important recommendations for sustaining engagement with the intervention following the pilot.

Interviews lasted up to 60 minutes. They were conducted by video call or telephone and were audio recorded and transcribed.

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).

Sampling for CWC

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

Table F5 below provides an overview of the primary criteria and achieved sample for this population.

Table F5: Sampling frame for City of Wolverhampton Council

Sample: 7		**7 participants:** • 1 focus groups with City of Wolverhampton Council Leadership • 1 focus groups with City of Wolverhampton Council frontline working level staff
Primary criteria		**Sample**
Type of partner	Frontline working staff	4
	Leadership	3

Recruitment of CWC

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

1. DHSC reached out to the CWC staff on their user list, informing them of the evaluation aim, focus group aims and BIT's role.
2. BIT reached out to these staff members via email. The email gave a summary of what focus groups would entail, and included:
    - A sign-up form
    - An information sheet that outlined the purpose of the interviews or focus groups, the topics that would be discussed and how their data would be protected.
3. BIT assigned a unique study ID to each user. This code was recorded in a secure spreadsheet with contact details.

4. BIT researchers sent participants a link to a Google Form through which they could book a focus group slot using their user ID. A copy of the information sheet was also included.

5. The researcher then sent the user a Google calendar invite, including a video-conferencing link (and an alternative telephone number to dial in) from their BIT email account.

Focus groups with CWC

Focus groups were conducted with members of staff at CWC - both in leadership positions and frontline positions - to understand their experiences of supporting the implementation of the scheme. The focus groups were 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 allowed comparison and contrast of perspectives and experiences of employees involved in a variety of different roles relevant to the delivery of the intervention. They provided 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 lasted up to 90 minutes. They were conducted by video call or telephone and were audio recorded and transcribed.

Data collection timelines

For recipients (both fully-engaged and Churn-outs), fieldwork was conducted at the conclusion of the intervention. This allowed a holistic assessment of their experience and ensured that the qualitative research does not affect the impact evaluation.

For non-recipients, interviews took place during the intervention's delivery so that they remembered their reasons for not signing up. This only happened after the recruitment window for the intervention had closed.

Reward partner and delivery partner interviews were 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.

Risks

Table F6 outlines the potential risks that were associated with the data collection process along with the mitigation measures BIT had in place. This risk register was periodically reviewed and updated during the final stages of the design and delivery of the evaluation.

Table F6: List of risks and mitigations

Risk Type	Risk	Mitigation
Methodological	Meeting recruitment targets for qualitative evaluation.	We employed multiple recruitment strategies simultaneously. If and when unable to reach targets, we relaxed our sampling criteria. We worked flexibly around participants' schedules to enable participation. We emphasised to people that their data would remain confidential, anonymous, and would be presented in aggregate form, and any personal information will be removed from reports, slides or other deliverables.
	There could have been vulnerable participants, such as people with eating disorders.	To ensure that participants were fully informed and empowered when taking part in this research, we sent participants an information sheet before the interview. This information sheet contained: an overview of topics that would be covered in the interviews, the research topics, how the interview

		responses would be used, and the research instruments in a way that was understandable to people participating in the research. The information sheet also contained information about where participants could access support, based on the DHSC guidelines. Researchers made themselves available to answer questions from participants via the telephone or email. This allowed people to make informed decisions about participating in the research and sharing their experiences.
	Special category data were being collected (such as ethnicity). This personal information was vulnerable to increased levels of harm and stress if there was a data breach or misuse of the data, over and above what could be caused by the release of less sensitive categories of data.	Security controls as outlined in the data protection and data security checklist were in place and reviewed periodically. Permissions and personnel involvement were reviewed regularly to ensure access was only granted to the minimum number of people that need it. The risk of a data breach could never be completely eliminated but the security controls and organisational procedures resulted in an acceptable level of risk given the personal data in question.
	Difficulties obtaining consent from people	Consent was always provided verbally at the beginning of the interview to the BIT researchers.

		Accessible information sheets i.e. the materials were sent electronically, posted in person or read aloud verbally by researchers and/or coaches. Consent was granular, and participants could consent to some forms of data collection and processing and not others, if they wished to.
	Guaranteeing confidentiality when a safeguarding issue was disclosed	In line with BIT's internal safeguarding procedures, participants were provided with an information sheet outlining sources of support available to participants (e.g., mental health). Further, they were informed at the start of the interview that, whilst the information shared would remain anonymous, confidentiality could be broken if something they say raises concerns about their safety or someone else's. In this scenario, BIT would have shared these details with the Wolverhampton City Council safeguarding lead who would escalate it according to their safeguarding policies.
	If a user became distressed	Specific mitigations for interviews included 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 also provided the option for participants to turn off their video when

		conducting a video interview. If videos remained on or data collection was conducted face to face, interviewers would be mindful of body language that indicated discomfort with the research. We also signposted participants to resources provided by HUL.

Appendix G: Qualitative data analysis procedure

Thematic analysis was carried out across the interview and focus group data, following a three-stage process:

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

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

2. **Data were coded by themes that responded to the research questions**: these themes were identified both deductively and inductively, using constructs from the literature when supported by evidence, and creating new constructs where it did not. When appropriate, we used the TDF to help categorise the themes to inform findings and adopted a contextualist method that took into account the individual perspective, as well as the social context.

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

4. The predetermined topics of the interview guide were 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 be identified from the data).

Appendix H: Full regression outputs

For all regression tables in this table, stars indicate the **unadjusted** p-values: +: p<0.1, *:p<0.05, **:p<0.01, ***: p<0.001. The figures in brackets indicate the 95% confidence intervals.

Table H1: Physical Activity M5 (95% CI in brackets)

	MVPA	Steps
(Intercept)	4.79* (0.73 - 8.84)	7809*** (7051 - 8567)
Treatment: any rewards	1.125+ (-0.049 - 2.299)	42 (-170 - 254)
Day of week: Monday[1]	2.85*** (1.73 - 3.96)	937*** (754 - 1121)
Day of week: Tuesday	3.17*** (2.06 - 4.28)	1213*** (1030 - 1396)
Day of week: Wednesday	1.85*** (0.76 - 2.95)	1320*** (1140 - 1500)
Day of week: Thursday	1.99*** (0.90 - 3.09)	1233*** (1054 - 1412)
Day of week: Friday	1.49** (0.40 - 2.58)	1175*** (998 - 1353)
Day of week: Saturday	-0.057 (-1.134 - 1.020)	523*** (347 - 699)
Week: 29[2]	-1 (-3 - 1)	-270 (-598 - 57)
Week: 30	-0.54 (-2.80 - 1.72)	-56 (-428 - 316)
Week: 31	-0.55 (-2.94 - 1.83)	-169 (-564 - 227)
Week: 32	0.083 (-2.355 - 2.522)	-35 (-443 - 372)
Week: 33	-0.23 (-2.76 - 2.30)	55 (-372 - 482)
Week: 34	-0.35 (-3.03 - 2.32)	-237 (-693 - 220)
Week: 35	-2.56+ (-5.48 - 0.36)	-514* (-1014 - -15)
Week: 36	-2.4 (-5.9 - 1.0)	-100 (-699 - 498)
Week: 37	-3.65+ (-7.85 - 0.55)	-270 (-1006 - 465)
Week: 38	-11.0* (-20.2 - -1.9)	-606 (-2178 - 965)

Age	0.127*** (0.083 - 0.172)	-11.8** (-20.0 - -3.5)
Sex: Male[3]	0.69 (-0.53 - 1.91)	508*** (286 - 729)
Sex: Other	-6.6 (-18.8 - 5.6)	-2523+ (-5257 - 211)
Ethnicity: Black / African / Caribbean / Black British[4]	0.88 (-1.75 - 3.52)	-135 (-624 - 353)
Ethnicity: Mixed / Multiple Ethnic Groups	2.75 (-0.71 - 6.22)	-154 (-776 - 469)
Ethnicity: Other Ethnic Group	-0.52 (-6.12 - 5.07)	-671 (-1786 - 444)
Ethnicity: Private	1.3 (-2.8 - 5.4)	421 (-336 - 1178)
Ethnicity: White	-0.014 (-1.597 - 1.569)	-197 (-484 - 89)
Education: No Qualification[5]	-1.78+ (-3.70 - 0.13)	228 (-126 - 581)
Education: Other Qualification	-1.338* (-2.654 - -0.022)	173 (-64 - 410)
Education: Prefer Not To Say	-1.29 (-3.43 - 0.85)	560** (159 - 961)
Education: Unknown	0.75 (-3.38 - 4.88)	53 (-752 - 857)
BMI	-0.196*** (-0.266 - -0.125)	-73.5*** (-87.0 - -60.1)
Device: Error[6]	1.5 (-7.8 - 10.7)	-397 (-3134 - 2340)
Device: Fitbit	16.25*** (14.57 - 17.94)	1881*** (1594 - 2167)
Device: Garmin	9.9*** (6.7 - 13.0)	1298*** (782 - 1813)
Device: Google Fit	4.2 (-3.0 - 11.4)	1724** (469 - 2978)
Device: HealthKit	22.07*** (20.42 - 23.72)	479** (181 - 778)
Deprivation	0.2394+ (-0.0081 - 0.4870)	38.3+ (-6.3 - 82.9)
Household size: 2[7]	0.18 (-1.14 - 1.50)	-82 (-319 - 156)
Household size: 3	-0.22 (-2.60 - 2.16)	-107 (-536 - 322)
Household size: 4	-1.5 (-6.6 - 3.7)	-556 (-1445 - 334)
Household size: 5	-1.1 (-11.2 - 9.0)	-1432 (-3410 - 547)

Household size: 16	-6.9 (-28.3 - 14.4)	-401 (-5258 - 4455)
Household size: 8	-0.72 (-26.96 - 25.52)	1066 (-3770 - 5901)
Baseline	0.0835*** (0.0649 - 0.1020)	0.1397*** (0.1218 - 0.1576)
Missing Baseline	9.7*** (7.6 - 11.9)	1112*** (939 - 1285)
N (observations)	21227	20874
N (participants)	3987	3934
[1] *reference level for day of week: Sunday*		
[2] *reference level for week: 28*		
[3] *reference level for sex: Female*		
[4] *reference level ethnicity: Asian / Asian British*		
[5] *reference level for education: Degree or above*		
[6] *reference level for device: Better Health: Rewards tracker*		
[7] *reference level for household size: 1*		

Table H2: Diet outcomes M5 (95% CI in brackets)

	Saturated Fat	**Fibre**	**Fruit and Vegetables**	**Sugars**
(Intercept)	0.08346*** (0.07475 - 0.09216)	3.138*** (2.142 - 4.134)	-14.4 (-43.04 - 14.25)	0.06548*** (0.05147 - 0.0795)
Treatment	-0.001032 (-0.003039 - 0.0009749)	0.3548** (0.1192 - 0.5904)	20.79*** (13.8 - 27.79)	-0.002133 (-0.005474 - 0.001209)
Weekday	-0.0005323 (-0.003421 - 0.002356)	-0.002064 (-0.34 - 0.3359)	11.56* (1.67 - 21.45)	-0.005321* (-0.01006 - -0.0005818)
Week: 34[1]	-0.003407 (-0.007891 - 0.001077)	-0.2931 (-0.8176 - 0.2314)	-5.38 (-20.75 - 9.995)	0.004855 (-0.002497 - 0.01221)

Week: 28	-0.002712 (-0.006121 - 0.0006973)	-0.1004 (-0.5004 - 0.2996)	-3.901 (-15.63 - 7.83)	-0.004137 (-0.009755 - 0.00148)
Week: 32	-0.001251 (-0.004963 - 0.002461)	-0.06643 (-0.5005 - 0.3676)	1.677 (-11.08 - 14.43)	-0.005027 (-0.01114 - 0.00109)
Week: 29	-0.002711 (-0.006031 - 0.0006082)	-0.2206 (-0.6086 - 0.1674)	4.196 (-7.193 - 15.58)	-0.007217** (-0.01268 - -0.001756)
Week: 37	-0.01107* (-0.02094 - -0.001205)	-0.5792 (-1.755 - 0.5966)	4.236 (-29.7 - 38.17)	-0.006783 (-0.02326 - 0.009696)
Week: 31	-0.000556 (-0.004076 - 0.002964)	-0.1868 (-0.5975 - 0.2238)	-1.046 (-13.08 - 10.99)	-0.003596 (-0.00936 - 0.002168)
Week: 35	-0.001365 (-0.006741 - 0.004011)	-0.2118 (-0.8389 - 0.4152)	3.444 (-14.95 - 21.84)	0.001031 (-0.007765 - 0.009826)
Week: 27	0.001375 (-0.004746 - 0.007496)	0.004992 (-0.7108 - 0.7208)	-6.959 (-27.88 - 13.97)	-0.0009725 (-0.01104 - 0.009099)
Week: 33	-0.0002758 (-0.004301 - 0.003749)	-0.2987 (-0.7681 - 0.1708)	-8.851 (-22.67 - 4.971)	-0.004465 (-0.01106 - 0.00213)
Week: 36	-0.008232* (-0.01509 - -0.001375)	-0.4525 (-1.266 - 0.361)	-6.239 (-29.97 - 17.49)	-0.002014 (-0.01338 - 0.009354)
Week: 38	-0.008017 (-0.03317 - 0.01714)	-1.206 (-4.155 - 1.743)	35.39 (-52.09 - 122.9)	0.003433 (-0.03829 - 0.04515)
Submissions	-0.001328 (-0.004259 - 0.001603)	0.4366* (0.09415 - 0.779)	23.93*** (13.96 - 33.9)	0.00321 (-0.001586 - 0.008006)

Baseline	0.2352*** (0.2092 - 0.2612)	0.3423*** (0.3204 - 0.3641)	0.4041*** (0.3783 - 0.4299)	0.3187*** (0.2971 - 0.3403)
Weekday baseline	0.0006852 (-0.002459 - 0.00383)	0.3932* (0.02299 - 0.7634)	5.609 (-5.15 - 16.37)	-0.001411 (-0.006591 - 0.003769)
Missing baseline	23.38*** (20.78 - 25.97)	77.98*** (40.96 - 115)	844.8 (-228.4 - 1918)	31.46*** (29.25 - 33.66)
Age	-0.0001018** (-0.0001743 - -0.00002921)	0.05182*** (0.04331 - 0.06033)	1.703*** (1.451 - 1.956)	-0.0002527*** (-0.0003731 - -0.0001324)
Sex: Male[2]	-0.00003333 (-0.002036 - 0.00197)	0.5443*** (0.3098 - 0.7788)	-15.06*** (-21.83 - -8.282)	0.00116 (-0.002095 - 0.004415)
Sex: Other	-0.0099 (-0.02904 - 0.009239)	-0.8652 (-3.238 - 1.508)	-54.12 (-122 - 13.76)	0.001984 (-0.03213 - 0.0361)
Ethnicity: Black / African / Caribbean / Black British[3]	0.0009264 (-0.003545 - 0.005397)	0.257 (-0.2686 - 0.7826)	10.49 (-5.091 - 26.08)	0.0202*** (0.01278 - 0.02762)
Ethnicity: Mixed / Multiple Ethnic Groups	0.009532*** (0.004004 - 0.01506)	0.3717 (-0.2843 - 1.028)	-6.448 (-25.49 - 12.59)	0.003957 (-0.00519 - 0.0131)
Ethnicity: Other Ethnic Group	0.009348* (0.0005587 - 0.01814)	0.3397 (-0.6917 - 1.371)	-11.44 (-41.68 - 18.81)	0.001844 (-0.01254 - 0.01623)
Ethnicity: Private	0.001722 (-0.00538 - 0.008824)	-0.06115 (-0.8842 - 0.7619)	13.52 (-10.37 - 37.42)	0.003858 (-0.007699 - 0.01542)
Ethnicity: White	0.008969*** (0.006238 - 0.0117)	0.04237 (-0.2768 - 0.3616)	2.8 (-6.612 - 12.21)	0.006241** (0.001742 - 0.01074)

Education: No Qualification[4]	-0.002885+ (-0.006081 - 0.0003111)	-0.8918*** (-1.266 - -0.5176)	-26.91*** (-37.87 - -15.94)	-0.002643 (-0.00789 - 0.002604)
Education: Other Qualification	-0.0006022 (-0.002831 - 0.001626)	-0.6238*** (-0.885 - -0.3625)	-16.77*** (-24.45 - -9.095)	-0.002262 (-0.005929 - 0.001405)
Education: Prefer Not To Say	-0.0009956 (-0.004503 - 0.002512)	-0.7468*** (-1.158 - -0.3357)	-20.42*** (-32.45 - -8.4)	-0.001949 (-0.007722 - 0.003825)
Education: Unknown	0.002134 (-0.005314 - 0.009581)	0.4509 (-0.4013 - 1.303)	5.45 (-19.9 - 30.8)	-0.005077 (-0.017 - 0.006842)
BMI	-0.00004655 (-0.0001649 - 0.00007182)	-0.01382+ (-0.0277 - 0.00006246)	-0.1315 (-0.5409 - 0.2778)	-0.0001896+ (-0.0003856 - 0.000006275)
Deprivation	0.0002972 (-0.000132 - 0.0007264)	0.09375*** (0.04322 - 0.1443)	1.103 (-0.3963 - 2.602)	-0.0004752 (-0.001189 - 0.0002386)
Household size: 2[5]	-0.00004836 (-0.002302 - 0.002205)	-0.09881 (-0.3635 - 0.1658)	-6.082 (-13.97 - 1.803)	-0.004196* (-0.007952 - -0.0004399)
Household size: 3	-0.001335 (-0.005399 - 0.002728)	-0.1616 (-0.6409 - 0.3178)	-0.9015 (-15.3 - 13.5)	-0.001824 (-0.008683 - 0.005035)
Household size: 4	0.002598 (-0.005855 - 0.01105)	-0.6424 (-1.646 - 0.3611)	-20.7 (-51.09 - 9.698)	-0.004656 (-0.01908 - 0.009771)
Household size: 5	0.001839 (-0.0156 - 0.01928)	-1.25 (-3.314 - 0.8143)	26.36 (-37 - 89.73)	-0.007921 (-0.03828 - 0.02244)
Household size: 16	-0.03193 (-0.07816 - 0.0143)	-1.825 (-7.536 - 3.885)	-68.15 (-242.7 - 106.4)	0.03066 (-0.05302 - 0.1143)

Household size: 8	0.00003476 (-0.04476 - 0.04483)	-0.05766 (-5.46 - 5.345)	57.53 (-113 - 228)	0.004686 (-0.07505 - 0.08443)
N	6073	6083	6149	6121
[1] *reference level for week: 26*				
[2] *reference level for sex: Female*				
[3] *reference level ethnicity: Asian / Asian British*				
[4] *reference level for education: Degree or above*				
[5] *reference level for household size: 1*				

Table H3: Physical Activity M3 (95% CI in brackets)

	MVPA	**Steps**
(Intercept)	11.2*** (6.9 - 15.5)	9239*** (8520 - 9959)
Treatment: any reward	2.78*** (1.58 - 3.98)	88 (-105 - 280)
Day of week: Monday[1]	3.78*** (2.69 - 4.88)	1229*** (1055 - 1402)
Day of week: Tuesday	3.99*** (2.90 - 5.07)	1172*** (999 - 1344)
Day of week: Wednesday	3.78*** (2.70 - 4.86)	1274*** (1103 - 1446)
Day of week: Thursday	2.31*** (1.23 - 3.38)	1310*** (1139 - 1481)
Day of week: Friday	1.31* (0.24 - 2.38)	1102*** (932 - 1272)
Day of week: Saturday	0.79 (-0.28 - 1.86)	1050*** (880 - 1219)
Week: 21[2]	-1.87+ (-3.94 - 0.21)	31 (-291 - 354)
Week: 22	-3.08** (-5.40 - -0.75)	-515** (-875 - -156)
Week: 23	-1.79 (-4.24 - 0.66)	-202 (-582 - 177)
Week: 24	-3.8** (-6.4 - -1.3)	-624** (-1015 - -233)

Week: 25	-4.6*** (-7.2 - -2.0)	-713*** (-1120 - -306)
Week: 26	-2.792* (-5.561 - -0.023)	-527* (-961 - -93)
Week: 27	-4.6** (-7.6 - -1.6)	-639** (-1111 - -168)
Week: 28	-4.9** (-8.3 - -1.5)	-1077*** (-1616 - -539)
Week: 29	-5.3* (-9.4 - -1.2)	-1134*** (-1793 - -475)
Week: 30	-4.7 (-13.2 - 3.7)	-959 (-2300 - 382)
Age	0.130*** (0.083 - 0.177)	-17.6*** (-25.2 - -10.0)
Sex: Male[3]	2.65*** (1.40 - 3.90)	424*** (219 - 629)
Sex: Other	-1.4 (-13.2 - 10.3)	312 (-1744 - 2369)
Ethnicity: Black / African / Caribbean / Black British[4]	2.7216+ (-0.0072 - 5.4504)	-123 (-568 - 322)
Ethnicity: Mixed / Multiple Ethnic Groups	2.2 (-1.4 - 5.8)	204 (-385 - 793)
Ethnicity: Other Ethnic Group	0.52 (-5.21 - 6.25)	-550 (-1524 - 425)
Ethnicity: Private	4.00+ (-0.29 - 8.29)	582+ (-111 - 1274)
Ethnicity: White	1.21 (-0.42 - 2.84)	14 (-247 - 276)
Education: No Qualification[5]	-1.87+ (-3.85 - 0.12)	444** (115 - 774)
Education: Other Qualification	-0.76 (-2.11 - 0.58)	163 (-52 - 379)
Education: Prefer Not To Say	-1.30 (-3.51 - 0.91)	441* (71 - 812)
Education: Unknown	1.6 (-2.7 - 5.8)	-198 (-922 - 526)
BMI	-0.339*** (-0.420 - -0.258)	-92.1*** (-105.5 - -78.7)
Device: Error[6]	7.7 (-4.9 - 20.3)	1035 (-1295 - 3365)
Device: Fitbit	13.85*** (12.13 - 15.56)	1505*** (1245 - 1764)
Device: Garmin	9.7*** (6.4 - 13.0)	807*** (331 - 1283)
Device: Google Fit	-3.7 (-11.8 - 4.4)	1041 (-459 - 2541)

Device: HealthKit	20.3*** (18.6 - 22.1)	118 (-165 - 400)
Deprivation	0.28* (0.03 - 0.53)	17 (-23 - 57)
Household size: 2[7]	0.21 (-1.14 - 1.57)	-160 (-375 - 55)
Household size: 3	-0.51 (-2.97 - 1.95)	-134 (-528 - 260)
Household size: 4	2.8 (-2.3 - 7.9)	-733+ (-1520 - 54)
Household size: 5	23.8*** (12.5 - 35.1)	-228 (-2138 - 1682)
Household size: 16	-6.7 (-29.7 - 16.2)	-989 (-4532 - 2553)
Baseline	0.0908*** (0.0724 - 0.1092)	0.1714*** (0.1546 - 0.1883)
Missing Baseline	10.1*** (7.9 - 12.2)	1418*** (1256 - 1581)
N (observations)	24710	24389
N (participants)	4458	4437
[1] *reference level for day of week: Sunday*		
[2] *reference level for week: 20*		
[3] *reference level for sex: Female*		
[4] *reference level ethnicity: Asian / Asian British*		
[5] *reference level for education: Degree or above*		
[6] *reference level for device: Better Health: Rewards tracker*		
[7] *reference level for household size: 1*		

Table H4: Diet outcomes M3 (95% CI in brackets)

	Saturated Fat	**Fibre**	**Fruit and Vegetables**	**Sugars**
(Intercept)	0.08115*** (0.07251 - 0.0898)	2.692*** (1.679 - 3.705)	-8.86 (-38.05 - 20.32)	0.07657*** (0.06241 - 0.09073)
Treatment	-0.00113 (-0.003071 - 0.0008116)	0.4299*** (0.1964 - 0.6635)	21.01*** (14.1 - 27.92)	-0.002134 (-0.005408 - 0.001139)

Weekday	-0.001168 (-0.003975 - 0.001638)	0.2974+ (-0.03984 - 0.6346)	13.05* (3.064 - 23.03)	-0.005429* (-0.01015 - -0.000713)
Week: 28	0.00007658 (-0.004427 - 0.00458)	-0.1438 (-0.6843 - 0.3966)	-9.137 (-25.09 - 6.819)	-0.005401 (-0.01294 - 0.002142)
Week: 22	0.001499 (-0.001822 - 0.004819)	0.1587 (-0.2394 - 0.5568)	-10.77+ (-22.6 - 1.062)	-0.002376 (-0.007957 - 0.003206)
Week: 26	0.003964* (0.0002853 - 0.007643)	0.04344 (-0.3983 - 0.4852)	-6.974 (-20.06 - 6.108)	-0.001352 (-0.007525 - 0.004821)
Week: 23	0.0005398 (-0.002786 - 0.003865)	0.1778 (-0.2215 - 0.5771)	-3.112 (-14.96 - 8.735)	-0.005149+ (-0.01075 - 0.00045)
Week: 29	-0.000901 (-0.006191 - 0.004389)	-0.05422 (-0.6945 - 0.5861)	-3.648 (-22.41 - 15.11)	-0.008738+ (-0.01766 - 0.0001826)
Week: 25	0.0007431 (-0.002754 - 0.00424)	-0.1851 (-0.6034 - 0.2332)	-6.757 (-19.15 - 5.639)	-0.0005619 (-0.00643 - 0.005306)
Week: 21	0.001127 (-0.003193 - 0.005447)	-0.03688 (-0.5552 - 0.4814)	-5.255 (-20.67 - 10.16)	-0.000486 (-0.007767 - 0.006796)
Week: 27	0.002908 (-0.001177 - 0.006993)	-0.1651 (-0.6558 - 0.3257)	-18.86* (-33.43 - -4.286)	0.0002348 (-0.006659 - 0.007128)
Week: 31	0.006011 (-0.006203 - 0.01823)	-0.003295 (-1.43 - 1.424)	-12.33 (-54.83 - 30.18)	0.01269 (-0.007382 - 0.03276)
Week: 30	-0.0007806 (-0.00801 - 0.006448)	-0.03665 (-0.9024 - 0.8291)	-27.62* (-53.31 - -1.935)	-0.0006898 (-0.01285 - 0.01147)

Week: 32	-0.01347 (-0.03797 - 0.01104)	0.4587 (-2.501 - 3.418)	3.758 (-84.09 - 91.6)	-0.03469 (-0.07609 - 0.00671)
Submissions	-0.0004321 (-0.003225 - 0.002361)	0.3277+ (-0.008578 - 0.6639)	14.02** (4.234 - 23.8)	-0.0002642 (-0.004936 - 0.004408)
Baseline	0.2204*** (0.1953 - 0.2456)	0.4009*** (0.3792 - 0.4226)	0.459*** (0.4333 - 0.4847)	0.3307*** (0.3095 - 0.3519)
Weekday baseline	0.003766* (0.0007263 - 0.006806)	0.2561 (-0.1098 - 0.6221)	-3.077 (-13.82 - 7.662)	0.001077 (-0.004045 - 0.0062)
Missing baseline	22.22*** (19.71 - 24.74)	70.02*** (33.46 - 106.6)	-21.41 (-1093 - 1050)	32.89*** (30.72 - 35.06)
Age	-0.00009163* (-0.0001615 - -0.0000218)	0.04615*** (0.03769 - 0.05462)	1.768*** (1.515 - 2.021)	-0.0002878*** (-0.0004068 - -0.0001688)
Sex: Male	-0.001602 (-0.003536 - 0.0003315)	0.4879*** (0.2549 - 0.7208)	-11.37** (-18.25 - -4.486)	-0.001642 (-0.004893 - 0.001609)
Sex: Other	0.002917 (-0.01743 - 0.02326)	-1.801 (-4.065 - 0.4626)	-81.47* (-148.7 - -14.2)	0.003135 (-0.03098 - 0.03725)
Ethnicity: Black / African / Caribbean / Black British	-0.0004723 (-0.004817 - 0.003872)	0.1575 (-0.3657 - 0.6808)	23.39** (7.944 - 38.83)	0.009741** (0.002426 - 0.01706)
Ethnicity: Mixed / Multiple Ethnic Groups	0.006433* (0.001024 - 0.01184)	-0.1102 (-0.7592 - 0.5389)	14.92 (-4.339 - 34.18)	0.003348 (-0.005771 - 0.01247)
Ethnicity: Other Ethnic Group	0.002214 (-0.006127 - 0.01055)	-0.852+ (-1.856 - 0.1523)	-4.823 (-34.86 - 25.21)	0.003416 (-0.01059 - 0.01742)

Ethnicity: Private	0.002325 (-0.004455 - 0.009106)	0.2961 (-0.5186 - 1.111)	21.06+ (-2.665 - 44.78)	-0.0005126 (-0.01182 - 0.0108)
Ethnicity: White	0.009098*** (0.00644 - 0.01176)	-0.2389 (-0.5567 - 0.07893)	2.66 (-6.698 - 12.02)	0.004625* (0.0001885 - 0.009062)
Education: No Qualification	-0.0041** (-0.007195 - -0.001005)	-0.8075*** (-1.179 - -0.4364)	-30.82*** (-41.79 - -19.85)	-0.003057 (-0.008255 - 0.00214)
Education: Other Qualification	-0.002939** (-0.005097 - -0.0007817)	-0.4573*** (-0.7174 - -0.1972)	-13.04*** (-20.76 - -5.331)	-0.001109 (-0.004743 - 0.002525)
Education: Prefer Not To Say	-0.0009821 (-0.004374 - 0.00241)	-0.5029* (-0.9122 - -0.09348)	-19.47** (-31.51 - -7.429)	0.001209 (-0.004491 - 0.006909)
Education: Unknown	0.0003995 (-0.006699 - 0.007498)	0.05968 (-0.8001 - 0.9194)	-2.702 (-27.89 - 22.49)	-0.005058 (-0.01704 - 0.006924)
BMI	-0.00006927 (-0.0001857 - 0.00004711)	-0.01125 (-0.02512 - 0.002615)	0.02211 (-0.389 - 0.4332)	-0.0002743** (-0.0004705 - -0.00007822)
Deprivation	0.0001269 (-0.0002886 - 0.0005423)	0.107*** (0.05685 - 0.1572)	3.327*** (1.841 - 4.812)	-0.0006788+ (-0.001377 - 0.00001978)
Household size: 2	-0.0000373 (-0.002219 - 0.002144)	0.2584+ (-0.0039 - 0.5207)	3.81 (-3.952 - 11.57)	-0.0008971 (-0.004566 - 0.002772)
Household size: 3	-0.002063 (-0.006061 - 0.001935)	0.01867 (-0.465 - 0.5023)	0.5654 (-13.65 - 14.78)	-0.001178 (-0.007935 - 0.005579)
Household size: 4	0.005559 (-0.002818 - 0.01394)	0.1441 (-0.8587 - 1.147)	-18.07 (-47.67 - 11.53)	-0.002898 (-0.01696 - 0.01117)

Household size: 5	-0.009501 (-0.02682 - 0.007815)	0.8347 (-1.185 - 2.854)	72.21* (12.41 - 132)	-0.02793+ (-0.05641 - 0.0005537)
Household size: 16	-0.005287 (-0.05021 - 0.03964)	0.6858 (-4.77 - 6.142)	51.59 (-108.5 - 211.7)	-0.03631 (-0.1123 - 0.0397)
Household size: 8	-0.0328 (-0.08053 - 0.01493)	2.326 (-3.465 - 8.116)	83.96 (-93.72 - 261.6)	-0.001744 (-0.08247 - 0.07898)
N	6382	6397	6467	6430

Table H5: Physical Activity M1 (95% CI in brackets)

	MVPA	**Steps**
(Intercept)	35.3* (8.3 - 62.2)	14696*** (10557 - 18835)
Treatment: any reward	0.897* (0.015 - 1.778)	151.0+ (-2.4 - 304.4)
Day of week: Monday[1]	3.01*** (2.17 - 3.86)	840*** (707 - 972)
Day of week: Tuesday	3.47*** (2.63 - 4.31)	1332*** (1201 - 1464)
Day of week: Wednesday	2.16*** (1.33 - 2.99)	1134*** (1004 - 1264)
Day of week: Thursday	2.26*** (1.43 - 3.08)	1261*** (1131 - 1390)
Day of week: Friday	0.51 (-0.31 - 1.34)	1202*** (1073 - 1331)
Day of week: Saturday	0.48 (-0.34 - 1.31)	943*** (815 - 1072)
Week: 12[2]	-27.02* (-53.81 - -0.22)	-6729** (-10836 - -2622)
Week: 13	-28.9* (-55.7 - -2.1)	-6656** (-10767 - -2544)
Week: 14	-28.5* (-55.3 - -1.7)	-6394** (-10508 - -2280)
Week: 15	-30.9* (-57.7 - -4.1)	-7045*** (-11160 - -2930)
Week: 16	-29.0* (-55.8 - -2.1)	-6315** (-10430 - -2199)

Week: 17	-29.4* (-56.2 - -2.6)	-6247** (-10363 - -2131)
Week: 18	-30.1* (-56.9 - -3.3)	-6421** (-10538 - -2303)
Week: 19	-30* (-57 - -3)	-6690** (-10810 - -2570)
Week: 20	-33.2* (-60.0 - -6.3)	-6384** (-10509 - -2259)
Week: 21	-32.0* (-58.9 - -5.1)	-6473** (-10608 - -2338)
Week: 22	-29.9* (-57.4 - -2.4)	-6576** (-10806 - -2346)
Age	0.148*** (0.115 - 0.180)	-16.4*** (-22.0 - -10.7)
Sex: Male[3]	0.867+ (-0.032 - 1.765)	372*** (213 - 530)
Sex: Other	2.5 (-5.1 - 10.2)	-837 (-2288 - 613)
Ethnicity: Black / African / Caribbean / Black British[4]	0.51 (-1.45 - 2.48)	-7.2 (-356.6 - 342.2)
Ethnicity: Mixed / Multiple Ethnic Groups	3.9** (1.4 - 6.5)	398+ (-66 - 862)
Ethnicity: Other Ethnic Group	2.5 (-1.5 - 6.5)	-557 (-1303 - 188)
Ethnicity: Private	2.38 (-0.68 - 5.44)	426 (-114 - 966)
Ethnicity: White	0.60 (-0.59 - 1.79)	234* (27 - 441)
Education: No Qualification[5]	-0.82 (-2.23 - 0.60)	505*** (252 - 759)
Education: Other Qualification	-0.42 (-1.41 - 0.56)	240** (69 - 410)
Education: Prefer Not To Say	-0.76 (-2.34 - 0.81)	636*** (355 - 917)
Education: Unknown	-0.58 (-3.59 - 2.44)	192 (-373 - 758)
BMI	-0.236*** (-0.290 - -0.181)	-87*** (-97 - -77)
Device: Error[6]	-2.9 (-9.3 - 3.6)	778 (-313 - 1868)
Device: Fitbit	11.7*** (10.3 - 13.0)	1238*** (1014 - 1462)
Device: Garmin	10.9*** (8.2 - 13.5)	939*** (513 - 1365)
Device: Google Fit	-1.6 (-6.4 - 3.3)	-44 (-957 - 869)

Device: HealthKit	20.15*** (18.86 - 21.43)	11 (-210 - 231)
Deprivation	0.237* (0.052 - 0.423)	22 (-10 - 54)
Household size: 2[7]	0.60 (-0.38 - 1.59)	-0.5 (-171.4 - 170.4)
Household size: 3	0.99 (-0.83 - 2.81)	-185 (-505 - 135)
Household size: 4	0.49 (-3.41 - 4.40)	-908** (-1570 - -246)
Household size: 5	3.7 (-4.4 - 11.9)	-210 (-1801 - 1382)
Household size: 16	-2.3 (-21.5 - 16.8)	-235 (-3812 - 3343)
Household size: 8	-5.5 (-27.9 - 16.9)	1314 (-2323 - 4951)
Baseline	0.1272*** (0.1132 - 0.1413)	0.1791*** (0.1664 - 0.1918)
Missing Baseline	14.41*** (12.76 - 16.05)	1406*** (1282 - 1530)
N (observations)	41201	40624
N (participants)	7190	7142
[1] *reference level for day of week: Sunday*		
[2] *reference level for week: 11*		
[3] *reference level for sex: Female*		
[4] *reference level ethnicity: Asian / Asian British*		
[5] *reference level for education: Degree or above*		
[6] *reference level for device: Better Health: Rewards tracker*		
[7] *reference level for household size: 1*		

Table H6: Diet outcomes M1 (95% CI in brackets)

	Saturated Fat	**Fibre**	**Fruit and Vegetables**	**Sugars**
(Intercept)	0.0813*** (0.0039)	3.71*** (0.46)	-9.9 (12.8)	0.0514*** (0.0063)
Treatment	-0.00088 (0.00095)	0.15 (0.12)	8.2* (3.3)	-0.0011 (0.0016)
Weekday	-0.00012	0.40*	14.9** (4.8)	0.00057

	(0.00137)	(0.17)		(0.00229)
Week: 13	-0.0045** (0.0018)	0.56** (0.21)	6.2 (6.1)	-0.0036 (0.0029)
Week: 19	-0.0038+ (0.0021)	0.0091 (0.2509)	-1.3 (7.2)	-0.0057+ (0.0034)
Week: 14	-0.0038* (0.0017)	0.093 (0.202)	1.7 (5.8)	-0.0044 (0.0028)
Week: 16	-0.0051** (0.0017)	0.22 (0.20)	-2.7 (5.8)	-0.0044 (0.0028)
Week: 22	-0.0092+ (0.0055)	-0.62 (0.66)	3.7 (18.9)	-0.0012 (0.0091)
Week: 17	-0.0046** (0.0017)	0.15 (0.20)	-1.6 (5.8)	-0.0037 (0.0028)
Week: 21	-0.0062* (0.0030)	0.13 (0.36)	-2 (10)	-0.0052 (0.0050)
Week: 20	-0.0048+ (0.0025)	0.077 (0.309)	6.4 (8.8)	-0.0013 (0.0042)
Week: 18	-0.0064*** (0.0019)	0.21 (0.23)	6.1 (6.5)	-0.0059+ (0.0031)
Week: 12	-0.0053 (0.0034)	0.066 (0.416)	5.1 (12.0)	-0.0134* (0.0058)
Week: 23	0.0087 (0.0189)	1.4 (2.0)	-46 (65)	-0.010 (0.028)
Submissions	-0.0010 (0.0011)	0.31* (0.14)	21.2*** (3.9)	0.0046* (0.0019)
Baseline	0.281*** (0.012)	0.375*** (0.011)	0.453*** (0.012)	0.38*** (0.01)
Weekday baseline	0.0017 (0.0015)	-0.26 (0.18)	-4.6 (5.2)	0.0034 (0.0025)
Missing baseline	28.0*** (1.2)	17 (18)	-127 (515)	38*** (1)
Age	-0.000156*** (0.000034)	0.0404*** (0.0041)	1.52*** (0.12)	-0.000237*** (0.000057)
Sex: Male	-0.00194*	0.14 (0.12)	-20.0*** (3.3)	0.0016

	(0.00096)			(0.0016)
Sex: Other	0.0094 (0.0087)	-1.2 (1.0)	-22 (30)	0.0028 (0.0149)
Ethnicity: Black / African / Caribbean / Black British	-0.0002 (0.0021)	0.55* (0.26)	22.5** (7.5)	0.0158*** (0.0036)
Ethnicity: Mixed / Multiple Ethnic Groups	0.0093*** (0.0027)	0.014 (0.329)	-3.5 (9.3)	0.0071 (0.0045)
Ethnicity: Other Ethnic Group	-0.0015 (0.0041)	-0.77 (0.49)	8.9 (14.1)	0.0024 (0.0068)
Ethnicity: Private	0.0013 (0.0033)	-0.92* (0.40)	18 (11)	0.0095+ (0.0056)
Ethnicity: White	0.0071*** (0.0013)	-0.012 (0.158)	11.1* (4.5)	0.0033 (0.0022)
Education: No Qualification	-0.0009 (0.0015)	-1.13*** (0.18)	-33.8*** (5.2)	0.00098 (0.00252)
Education: Other Qualification	-0.0010 (0.0011)	-0.46*** (0.13)	-11.5** (3.7)	-0.0029 (0.0018)
Education: Prefer Not To Say	-0.0034* (0.0016)	-0.9*** (0.2)	-18.5** (5.7)	0.00022 (0.00275)
Education: Unknown	-0.00071 (0.00346)	-1.12** (0.43)	-28* (12)	0.0088 (0.0058)
BMI	-0.000026 (0.000056)	-0.0071 (0.0069)	0.084 (0.198)	-0.000150 (0.000094)
Deprivation	0.00043* (0.00020)	0.113*** (0.025)	1.92** (0.71)	-0.00014 (0.00034)
Household size: 2	0.00055 (0.00106)	0.055 (0.131)	-2.6 (3.7)	-0.0023 (0.0018)
Household size: 3	0.0026 (0.0019)	-0.36 (0.24)	-9.8 (6.8)	-0.0020 (0.0033)
Household size: 4	0.0039 (0.0041)	-0.52 (0.51)	-13 (15)	0.0034 (0.0070)
Household size: 5	0.0075	0.35 (1.02)	43 (29)	-0.0022

	(0.0080)			(0.0139)
Household size: 16	0.012 (0.023)	1.8 (3.1)	-2.1 (86.9)	-0.026 (0.041)
Household size: 8	-0.012 (0.021)	-0.48 (2.84)	-52 (83)	-0.014 (0.038)
N	6843	6857	6929	6897

Table H7: Physical Activity by incentive level M5 (95% CI in brackets)

	MVPA	**Steps**
(Intercept)	4.79* (0.73 - 8.85)	7796*** (7040 - 8552)
Treatment arm: Low reward[1]	0.80 (-0.85 - 2.46)	-117 (-412 - 177)
Treatment arm: Medium reward	1.59+ (-0.22 - 3.41)	617*** (296 - 937)
Treatment arm: High reward	1.07 (-0.69 - 2.82)	-294+ (-601 - 12)
Day of week: Monday[2]	2.85*** (1.73 - 3.96)	937*** (753 - 1120)
Day of week: Tuesday	3.17*** (2.06 - 4.28)	1213*** (1030 - 1396)
Day of week: Wednesday	1.85*** (0.75 - 2.95)	1319*** (1139 - 1499)
Day of week: Thursday	1.99*** (0.90 - 3.09)	1232*** (1053 - 1412)
Day of week: Friday	1.49** (0.40 - 2.58)	1176*** (998 - 1354)
Day of week: Saturday	-0.057 (-1.135 - 1.020)	522*** (346 - 699)
Week: 29[3]	-1.06 (-3.08 - 0.96)	-290+ (-619 - 39)
Week: 30	-0.58 (-2.85 - 1.69)	-74 (-447 - 298)
Week: 31	-0.59 (-2.98 - 1.80)	-186 (-581 - 210)
Week: 32	0.049 (-2.393 - 2.491)	-53 (-460 - 355)
Week: 33	-0.25 (-2.78 - 2.28)	50 (-377 - 477)
Week: 34	-0.38 (-3.06 - 2.31)	-240 (-696 - 216)
Week: 35	-2.58+ (-5.50 - 0.34)	-516* (-1014 - -17)
Week: 36	-2.5 (-5.9 - 1.0)	-100 (-697 - 498)
Week: 37	-3.67+ (-7.88 - 0.53)	-269 (-1004 - 465)
Week: 38	-11.1* (-20.2 - -1.9)	-607 (-2178 - 964)
Age	0.128*** (0.083 - 0.173)	-11.1** (-19.3 - -2.8)
Sex: Male[4]	0.70 (-0.52 - 1.92)	506*** (285 - 727)

Sex: Other	-6.9 (-19.1 - 5.3)	-2965* (-5697 - -233)
Ethnicity: Black / African / Caribbean / Black British[5]	0.87 (-1.77 - 3.51)	-137 (-625 - 350)
Ethnicity: Mixed / Multiple Ethnic Groups	2.74 (-0.73 - 6.21)	-155 (-776 - 465)
Ethnicity: Other Ethnic Group	-0.58 (-6.18 - 5.02)	-742 (-1854 - 371)
Ethnicity: Private	1.3 (-2.8 - 5.4)	418 (-337 - 1172)
Ethnicity: White	-0.0084 (-1.5918 - 1.5751)	-191 (-476 - 95)
Education: No Qualification[6]	-1.78+ (-3.70 - 0.13)	225 (-128 - 578)
Education: Other Qualification	-1.342* (-2.658 - -0.026)	176 (-60 - 412)
Education: Prefer Not To Say	-1.29 (-3.43 - 0.85)	572** (172 - 972)
Education: Unknown	0.69 (-3.44 - 4.82)	-3.3 (-806.2 - 799.7)
BMI	-0.196*** (-0.266 - -0.125)	-73.6*** (-87.1 - -60.2)
Device: Error[7]	1.5 (-7.7 - 10.7)	-389 (-3125 - 2347)
Device: Fitbit	16.28*** (14.59 - 17.97)	1890*** (1604 - 2177)
Device: Garmin	9.9*** (6.8 - 13.1)	1360*** (846 - 1875)
Device: Google Fit	4.3 (-2.9 - 11.5)	1779** (525 - 3033)
Device: HealthKit	22.09*** (20.45 - 23.74)	498** (200 - 796)
Deprivation	0.2380+ (-0.0098 - 0.4859)	34 (-10 - 79)
Household size: 2[8]	0.17 (-1.15 - 1.49)	-91 (-327 - 145)
Household size: 3	-0.22 (-2.60 - 2.16)	-113 (-540 - 315)
Household size: 4	-1.5 (-6.7 - 3.6)	-662 (-1549 - 225)
Household size: 5	-1.4 (-11.5 - 8.7)	-1830+ (-3808 - 148)
Household size: 16	-7.4 (-28.8 - 14.0)	-974 (-5817 - 3869)
Household size: 8	-0.69 (-26.93 - 25.56)	1110 (-3707 - 5927)
Baseline	0.0834*** (0.0648 - 0.1020)	0.1400*** (0.1221 - 0.1579)
Missing Baseline	9.7*** (7.5 - 11.9)	1116*** (943 - 1289)
N (observations)	21227	20874

N (participants)	3987	3934
[1] *reference level for day of treatment arm: Control*		
[2] *reference level for day of week: Sunday*		
[3] *reference level for week: 28*		
[4] *reference level for sex: Female*		
[5] *reference level ethnicity: Asian / Asian British*		
[6] *reference level for education: Degree or above*		
[7] *reference level for device: Better Health: Rewards tracker*		
[8] *reference level for household size: 1*		

Table H8: Diet primary outcomes by incentive level at M5 (95% CI in brackets)

	Saturated Fat	**Fibre**	**Fruit and Vegetables**	**Sugars**
(Intercept)	0.08329*** (0.07459 - 0.09199)	3.147*** (2.151 - 4.143)	-13.94 (-42.58 - 14.71)	0.06546*** (0.05144 - 0.07948)
Treatment arm: High reward	-0.004438** (-0.007448 - -0.001429)	0.4991** (0.1458 - 0.8523)	28.88*** (18.43 - 39.33)	-0.002118 (-0.007116 - 0.00288)
Treatment arm: Low reward	0.001368 (-0.001473 - 0.004209)	0.2141 (-0.1199 - 0.5482)	16.33** (6.415 - 26.25)	-0.002744 (-0.00748 - 0.001993)
Treatment arm: Medium reward	-0.0003699 (-0.003438 - 0.002698)	0.3731* (0.0133 - 0.7328)	17.65** (6.966 - 28.33)	-0.001417 (-0.006528 - 0.003694)
Weekday	-0.0006582 (-0.003546 - 0.00223)	0.004175 (-0.3339 - 0.3422)	11.86* (1.968 - 21.75)	-0.005302* (-0.01004 - -0.0005607)
Week: 34	-0.003427 (-0.007909 -	-0.2895 (-0.8142 -	-5.448 (-20.82 -	0.004889 (-0.002465 -

	0.001055)	0.2351)	9.928)	0.01224)
Week: 28	-0.002542 (-0.00595 - 0.0008669)	-0.1076 (-0.5078 - 0.2926)	-4.262 (-16 - 7.472)	-0.004137 (-0.009757 - 0.001484)
Week: 32	-0.001167 (-0.004878 - 0.002544)	-0.06787 (-0.5021 - 0.3663)	1.472 (-11.28 - 14.23)	-0.004999 (-0.01112 - 0.001121)
Week: 29	-0.002665 (-0.005982 - 0.0006533)	-0.2215 (-0.6095 - 0.1665)	4.09 (-7.298 - 15.48)	-0.007205** (-0.01267 - -0.001743)
Week: 37	-0.01115* (-0.02101 - -0.001286)	-0.568 (-1.745 - 0.6085)	4.268 (-29.68 - 38.22)	-0.006681 (-0.02317 - 0.009806)
Week: 31	-0.0005194 (-0.004038 - 0.002999)	-0.1875 (-0.5982 - 0.2233)	-1.142 (-13.18 - 10.89)	-0.003588 (-0.009353 - 0.002178)
Week: 35	-0.001256 (-0.00663 - 0.004117)	-0.2141 (-0.8413 - 0.413)	3.236 (-15.16 - 21.63)	0.001039 (-0.007758 - 0.009836)
Week: 27	0.000936 (-0.0052 - 0.007072)	0.03181 (-0.6859 - 0.7495)	-6.312 (-27.29 - 14.67)	-0.0008228 (-0.01092 - 0.009278)
Week: 33	-0.0002102 (-0.004234 - 0.003813)	-0.3009 (-0.7705 - 0.1686)	-9.021 (-22.84 - 4.801)	-0.004453 (-0.01105 - 0.002143)
Week: 36	-0.008253* (-0.01511 - -0.001401)	-0.4557 (-1.269 - 0.3579)	-6.263 (-29.99 - 17.46)	-0.002018 (-0.01339 - 0.009352)
Week: 38	-0.007664 (-0.0328 - 0.01748)	-1.213 (-4.163 - 1.737)	34.27 (-53.2 - 121.7)	0.003522 (-0.0382 - 0.04525)
Submissions	-0.001234 (-0.004164 - 0.001695)	0.4305* (0.08798 - 0.7731)	23.73*** (13.76 - 33.7)	0.003188 (-0.00161 - 0.007985)

Baseline	0.2351*** (0.2091 - 0.2611)	0.3421*** (0.3202 - 0.3639)	0.4041*** (0.3783 - 0.4299)	0.3187*** (0.2971 - 0.3403)
Weekday baseline	0.0007501 (-0.002393 - 0.003893)	0.3902* (0.01987 - 0.7604)	5.466 (-5.292 - 16.22)	-0.001404 (-0.006585 - 0.003778)
Missing baseline	23.37*** (20.78 - 25.97)	77.65*** (40.63 - 114.7)	830.7 (-242.4 - 1904)	31.46*** (29.26 - 33.66)
Age	-0.0001016** (-0.0001742 - -0.00002909)	0.05188*** (0.04336 - 0.06039)	1.702*** (1.45 - 1.954)	-0.0002522*** (-0.0003726 - -0.0001317)
Sex: Male	-0.00001628 (-0.002018 - 0.001986)	0.5436*** (0.3091 - 0.7782)	-15.09*** (-21.86 - -8.311)	0.001157 (-0.0021 - 0.004414)
Sex: Other	-0.01002 (-0.02914 - 0.009109)	-0.862 (-3.236 - 1.512)	-53.66 (-121.5 - 14.22)	0.001828 (-0.03229 - 0.03594)
Ethnicity: Black / African / Caribbean / Black British	0.0009902 (-0.003478 - 0.005458)	0.2513 (-0.2744 - 0.777)	10.34 (-5.243 - 25.92)	0.02017*** (0.01275 - 0.02759)
Ethnicity: Mixed / Multiple Ethnic Groups	0.00959*** (0.004065 - 0.01512)	0.3659 (-0.2902 - 1.022)	-6.514 (-25.55 - 12.53)	0.003915 (-0.005233 - 0.01306)
Ethnicity: Other Ethnic Group	0.009384* (0.0005993 - 0.01817)	0.3308 (-0.7008 - 1.362)	-11.62 (-41.86 - 18.63)	0.0018 (-0.01258 - 0.01618)
Ethnicity: Private	0.001589 (-0.005508 - 0.008686)	-0.05667 (-0.8797 - 0.7664)	13.76 (-10.12 - 37.65)	0.003875 (-0.007683 - 0.01543)
Ethnicity: White	0.008977*** (0.006248 - 0.01171)	0.04274 (-0.2765 - 0.3619)	2.781 (-6.628 - 12.19)	0.006253** (0.001755 - 0.01075)

Education: No Qualification	-0.002916+ (-0.006111 - 0.0002779)	-0.8906*** (-1.265 - -0.5163)	-26.95*** (-37.91 - -15.99)	-0.002634 (-0.007883 - 0.002615)
Education: Other Qualification	-0.0005811 (-0.002808 - 0.001646)	-0.6247*** (-0.886 - -0.3634)	-16.86*** (-24.53 - -9.18)	-0.002259 (-0.005927 - 0.001408)
Education: Prefer Not To Say	-0.0009174 (-0.004423 - 0.002588)	-0.7502*** (-1.161 - -0.339)	-20.63*** (-32.66 - -8.607)	-0.001938 (-0.007713 - 0.003838)
Education: Unknown	0.002021 (-0.005423 - 0.009464)	0.4529 (-0.3995 - 1.305)	5.728 (-19.62 - 31.08)	-0.005149 (-0.01707 - 0.006773)
BMI	-0.00004617 (-0.0001645 - 0.00007212)	-0.01384+ (-0.02772 - 0.00004354)	-0.1335 (-0.5428 - 0.2757)	-0.0001895+ (-0.0003854 - 0.000006461)
Deprivation	0.0002796 (-0.0001494 - 0.0007087)	0.09447*** (0.04391 - 0.145)	1.146 (-0.3528 - 2.646)	-0.0004764 (-0.00119 - 0.0002374)
Household size: 2	-0.00001884 (-0.002271 - 0.002233)	-0.09956 (-0.3642 - 0.1651)	-6.154 (-14.04 - 1.729)	-0.004195* (-0.00795 - -0.0004405)
Household size: 3	-0.001357 (-0.005416 - 0.002703)	-0.1626 (-0.642 - 0.3168)	-0.8135 (-15.21 - 13.58)	-0.001844 (-0.008697 - 0.00501)
Household size: 4	0.002852 (-0.005593 - 0.0113)	-0.6576 (-1.661 - 0.3463)	-21.07 (-51.45 - 9.319)	-0.004728 (-0.01914 - 0.009684)
Household size: 5	0.001535 (-0.01591 - 0.01898)	-1.256 (-3.324 - 0.8114)	27.76 (-35.65 - 91.17)	-0.008258 (-0.03861 - 0.0221)
Household size: 16	-0.03252 (-0.07872 - 0.01368)	-1.838 (-7.555 - 3.878)	-65.15 (-239.7 - 109.4)	0.02996 (-0.0536 - 0.1135)

Household size: 8	-0.000003925 (-0.04471 - 0.0447)	-0.05194 (-5.454 - 5.35)	57.72 (-112.6 - 228)	0.004715 (-0.0748 - 0.08423)
N	6073	6083	6149	6121

Table H9: Secondary outcome: Energy expenditure M5 (95% CI in brackets)

	Energy expenditure
(Intercept)	1063.1*** (1016.8 - 1109.4)
Treatment: any reward	-2.29 (-16.03 - 11.45)
Day of week: Monday	20.74*** (13.05 - 28.44)
Day of week: Tuesday	24.22*** (16.63 - 31.80)
Day of week: Wednesday	23.71*** (16.29 - 31.12)
Day of week: Thursday	31.05*** (23.73 - 38.37)
Day of week: Friday	30.71*** (23.49 - 37.93)
Day of week: Saturday	14.31*** (7.19 - 21.42)
Week: 29	-3.30 (-18.07 - 11.46)
Week: 30	-5.26 (-22.85 - 12.33)
Week: 31	-7.99 (-27.61 - 11.62)
Week: 32	-5.62 (-26.58 - 15.34)
Week: 33	-9.87 (-32.25 - 12.51)
Week: 34	-7.38 (-31.64 - 16.87)
Week: 35	-9.14 (-35.90 - 17.62)
Week: 36	-20.1 (-51.9 - 11.6)
Week: 37	-12.1 (-51.0 - 26.9)
Week: 38	-38.5 (-109.8 - 32.8)
Age	-5.40*** (-5.93 - -4.87)
Sex: Male	409.71*** (395.24 - 424.17)
Sex: Other	330.4+ (-18.5 - 679.3)
Ethnicity: Black / African / Caribbean	21.38 (-9.85 - 52.61)

/ Black British	
Ethnicity: Mixed / Multiple Ethnic Groups	58.7** (18.5 - 98.9)
Ethnicity: Other Ethnic Group	-4.4 (-73.7 - 64.8)
Ethnicity: Private	15.4 (-33.3 - 64.1)
Ethnicity: White	62.70*** (44.16 - 81.24)
Education: No Qualification	-23.224* (-45.918 - -0.531)
Education: Other Qualification	1.19 (-14.17 - 16.55)
Education: Prefer Not To Say	-7.98 (-33.69 - 17.73)
Education: Unknown	-24.1 (-75.0 - 26.8)
BMI	19.10*** (18.23 - 19.96)
Device: Error	73.02* (7.84 - 138.20)
Device: Fitbit	889.13*** (870.03 - 908.22)
Device: Garmin	672.4*** (637.7 - 707.1)
Device: Google Fit	174.2*** (114.6 - 233.9)
Device: HealthKit	-9.40 (-27.99 - 9.18)
Deprivation	-0.0711 (-2.9716 - 2.8294)
Household size: 2	-13.04+ (-28.46 - 2.39)
Household size: 3	1.1 (-26.5 - 28.7)
Household size: 4	-7.32 (-64.72 - 50.09)
Household size: 5	17.3 (-109.0 - 143.7)
Household size: 16	-106 (-409 - 198)
Household size: 8	158 (-143 - 460)
Baseline	0.02979*** (0.02016 - 0.03943)
Missing baseline	61.78*** (43.30 - 80.26)
N (observations)	21038
N (participants)	3973

Table H10: Secondary outcomes (Weight and Diet, M5) (95% CI in brackets)

	Weight	**Healthy eating**	**Energy Intake**

		score	
(Intercept)	11.14 (-4.883 - 27.17)	3.204*** (3.003 - 3.404)	668.5*** (560.5 - 776.5)
Treatment	-0.2566 (-0.7554 - 0.2422)	0.04017+ (-0.003868 - 0.0842)	1.716 (-23.67 - 27.1)
Weekday	-0.213 (-0.7388 - 0.3128)	0.05077 (-0.01227 - 0.1138)	-29.14 (-65.53 - 7.249)
Week: 27	-1.911 (-17.92 - 14.1)	0.007123 (-0.1277 - 0.1419)	-25.47 (-103.1 - 52.12)
Week: 28	-2.332 (-18.29 - 13.63)	0.06862+ (-0.006056 - 0.1433)	-34.72 (-77.76 - 8.312)
Week: 29	-2.427 (-18.39 - 13.53)	0.08231* (0.009639 - 0.155)	-60.14** (-102 - -18.29)
Week: 30	-2.75 (-18.71 - 13.21)		
Week: 31	-2.902 (-18.87 - 13.06)	0.04335 (-0.03347 - 0.1202)	-15.19 (-59.48 - 29.1)
Week: 32	-2.853 (-18.82 - 13.11)	0.05972 (-0.0215 - 0.1409)	-19.37 (-66.22 - 27.48)
Week: 33	-2.247 (-18.22 - 13.73)	0.04353 (-0.04409 - 0.1311)	-34.35 (-84.93 - 16.22)
Week: 34	-1.235 (-17.21 - 14.74)	0.06554 (-0.03227 - 0.1633)	-51.91+ (-108.3 - 4.479)
Week: 35	-1.562 (-17.56 - 14.44)	0.01262 (-0.1044 - 0.1296)	17.8 (-49.79 - 85.39)
Week: 36	-0.5484 (-16.57 - 15.47)	0.0624 (-0.08718 - 0.212)	10.83 (-76.35 - 98.01)
Week: 37	-1.494 (-17.59 - 14.6)	0.1748 (-0.041 - 0.3907)	-39.94 (-165.8 - 85.97)
Week: 38	-1.976 (-18.28 - 14.32)	0.1827 (-0.3715 - 0.7369)	11.9 (-305.9 - 329.7)

Week: 39	-1.44 (-18.26 - 15.38)		
Week: 40	-1.637 (-18.26 - 14.99)		
Week: 41	-1.824 (-18.4 - 14.75)		
Age	-0.03088** (-0.04929 - -0.01247)	0.003878*** (0.002289 - 0.005466)	1.448** (0.5365 - 2.359)
Sex: Male	1.375*** (0.7427 - 2.008)	-0.04443* (-0.0881 - -0.0007473)	122*** (96.42 - 147.6)
Sex: Other	0.1474 (-4.097 - 4.392)	-0.9473*** (-1.36 - -0.5348)	62.47 (-193 - 318)
Ethnicity: Black / African / Caribbean / Black British	0.5413 (-0.5801 - 1.663)	-0.1624** (-0.26 - -0.06487)	75.88** (19.51 - 132.3)
Ethnicity: Mixed / Multiple Ethnic Groups	0.7207 (-0.701 - 2.142)	-0.1766** (-0.2976 - -0.05556)	101** (31.08 - 171)
Ethnicity: Other Ethnic Group	0.01081 (-2.114 - 2.135)	-0.06601 (-0.2579 - 0.1259)	44.57 (-67.45 - 156.6)
Ethnicity: Private	0.8462 (-0.874 - 2.566)	-0.1314+ (-0.2832 - 0.02032)	75.45 (-15.2 - 166.1)
Ethnicity: White	0.753* (0.07003 - 1.436)	-0.1712*** (-0.2308 - -0.1116)	66.97*** (32.65 - 101.3)
Education: No Qualification	0.9999* (0.1992 - 1.801)	0.04258 (-0.02684 - 0.112)	-104.5*** (-144.7 - -64.33)
Education: Other Qualification	0.1979 (-0.361 - 0.7569)	0.05672* (0.007956 - 0.1055)	-76.79*** (-104.9 - -48.65)
Education: Prefer Not To Say	-0.2596 (-1.145 - 0.6255)	0.1002* (0.02371 - 0.1768)	-118.2*** (-162.4 - -73.97)

Education: Unknown	-2.238* (-4.153 - -0.3236)	0.02128 (-0.137 - 0.1796)	27.89 (-64.12 - 119.9)
BMI	-0.1567*** (-0.235 - -0.07842)	0.0001119 (-0.002496 - 0.00272)	-1.203 (-2.705 - 0.2981)
Deprivation	0.01978 (-0.08677 - 0.1263)	-0.003294 (-0.01269 - 0.006105)	9.307*** (3.862 - 14.75)
Household size: 2	0.1361 (-0.4238 - 0.6959)	0.01278 (-0.03651 - 0.06206)	-21.86 (-50.38 - 6.665)
Household size: 3	-0.09819 (-1.1 - 0.904)	0.0585 (-0.03053 - 0.1475)	-28.14 (-79.59 - 23.31)
Household size: 4	1.099 (-1.018 - 3.216)	-0.06529 (-0.251 - 0.1204)	-45.21 (-153.1 - 62.63)
Household size: 5	-1.019 (-5.381 - 3.342)	0.1694 (-0.2115 - 0.5504)	-172.9 (-394.8 - 48.91)
Household size: 16	-1.022 (-11.25 - 9.204)	0.007455 (-1.017 - 1.032)	-178.2 (-791.1 - 434.7)
Household size: 8	-0.5425 (-10.2 - 9.115)	0.1966 (-0.7966 - 1.19)	-84.54 (-663.8 - 494.7)
Baseline	0.9286*** (0.9012 - 0.956)	0.2517*** (0.2309 - 0.2726)	0.3018*** (0.2826 - 0.3211)
Submissions		-0.008785 (-0.07236 - 0.05479)	16.54 (-20.26 - 53.34)
Weekday baseline		-0.01135 (-0.07964 - 0.05694)	21.57 (-18.37 - 61.52)
Missing baseline		24.96*** (17.88 - 32.03)	2666 (-1319 - 6652)
N	4670	6168	6083

Table H11: Physical Activity Sensitivity Analysis M5 (95% CI in brackets)

	MVPA	Steps
(Intercept)	0.98 (-2.20 - 4.17)	5341*** (4683 - 5999)
Treatment: any reward	1.93*** (1.00 - 2.85)	256** (71 - 442)
Day of week: Monday[1]	2.97*** (2.12 - 3.82)	943*** (786 - 1100)
Day of week: Tuesday	3.22*** (2.37 - 4.06)	1193*** (1037 - 1349)
Day of week: Wednesday	2.42*** (1.59 - 3.26)	1387*** (1233 - 1540)
Day of week: Thursday	2.14*** (1.32 - 2.97)	1278*** (1125 - 1431)
Day of week: Friday	1.69*** (0.86 - 2.51)	1144*** (991 - 1296)
Day of week: Saturday	0.59 (-0.23 - 1.41)	456*** (305 - 607)
Week: 29[2]	-1.16 (-2.75 - 0.43)	7 (-276 - 289)
Week: 30	-1.13 (-2.92 - 0.66)	-19 (-339 - 302)
Week: 31	-1.13 (-3.02 - 0.76)	-90 (-431 - 252)
Week: 32	-0.53 (-2.45 - 1.40)	66 (-286 - 417)
Week: 33	-1.54 (-3.54 - 0.45)	-147 (-514 - 220)
Week: 34	-1.82+ (-3.92 - 0.29)	-133 (-523 - 257)
Week: 35	-3.5** (-5.7 - -1.2)	-654** (-1078 - -230)
Week: 36	-4.1** (-6.7 - -1.5)	-163 (-655 - 329)
Week: 37	-5.6*** (-8.7 - -2.5)	-620* (-1215 - -26)
Week: 38	-11.9*** (-18.2 - -5.6)	-862 (-2041 - 317)
Age	0.178*** (0.143 - 0.213)	19.2*** (12.1 - 26.3)
Sex: Male[3]	0.895+ (-0.081 - 1.871)	420*** (228 - 612)
Sex: Other	-5.0 (-12.9 - 2.9)	-2333* (-4387 - -279)
Ethnicity: Black / African / Caribbean / Black British[4]	1.72 (-0.34 - 3.78)	-352 (-773 - 69)
Ethnicity: Mixed / Multiple Ethnic Groups	1.6 (-1.1 - 4.2)	-570* (-1106 - -35)

Ethnicity: Other Ethnic Group	0.18 (-3.83 - 4.19)	-834+ (-1696 - 28)
Ethnicity: Private	-0.14 (-3.34 - 3.07)	323 (-345 - 992)
Ethnicity: White	0.38 (-0.86 - 1.61)	-233+ (-484 - 17)
Education: No Qualification[5]	-2.60*** (-4.12 - -1.09)	124 (-188 - 437)
Education: Other Qualification	-1.140* (-2.195 - -0.085)	84 (-123 - 290)
Education: Prefer Not To Say	-3.01*** (-4.69 - -1.34)	383* (34 - 733)
Education: Unknown	1.8 (-1.3 - 5.0)	-71 (-767 - 626)
BMI	-0.145*** (-0.201 - -0.090)	-69.3*** (-81.2 - -57.4)
Device: Fitbit[6]	12.04*** (10.63 - 13.45)	2050*** (1787 - 2314)
Device: Garmin	10.1*** (7.6 - 12.7)	1911*** (1440 - 2381)
Device: Google Fit	-0.4 (-2.6 - 1.8)	-662** (-1097 - -226)
Device: HealthKit	9.77*** (8.57 - 10.97)	192 (-45 - 429)
Deprivation	0.313** (0.118 - 0.507)	44.9* (5.9 - 83.9)
Household size: 2[7]	0.66 (-0.38 - 1.69)	-57 (-265 - 151)
Household size: 3	-0.95 (-2.75 - 0.85)	-214 (-582 - 155)
Household size: 4	-2.2 (-6.0 - 1.6)	-705+ (-1494 - 85)
Household size: 5	5.0 (-2.4 - 12.3)	-1989* (-3670 - -308)
Household size: 16	-6.2 (-21.4 - 9.1)	775 (-3496 - 5046)
Household size: 8	-2.9 (-22.2 - 16.5)	230 (-3856 - 4315)
Baseline	0.0763*** (0.0618 - 0.0908)	0.131*** (0.115 - 0.147)
Missing Baseline	8.83*** (7.14 - 10.53)	992*** (841 - 1143)
N (observations)	35527	33351
N (participants)	5645	5480
[1] *reference level for day of week: Sunday*		

[2] *reference level for week: 28*
[3] *reference level for sex: Female*
[4] *reference level ethnicity: Asian / Asian British*
[5] *reference level for education: Degree or above*
[6] *reference level for device: Better Health: Rewards tracker*
[7] *reference level for household size: 1*

Table H12: Exploratory outcome variables (95% CI in brackets)

	Wellbeing	**Sleep**
(Intercept)	25.1*** (19.88 - 30.32)	365.7*** (344.5 - 386.8)
Treatment	2.045*** (0.8746 - 3.215)	-1.938 (-6.782 - 2.906)
Day of week: Monday	0.5971 (-1.522 - 2.716)	-14.74*** (-19.53 - -9.952)
Day of week: Tuesday	-1.448 (-3.548 - 0.6527)	-17.35*** (-22.09 - -12.6)
Day of week: Wednesday	1.504 (-0.5856 - 3.593)	-15.98*** (-20.73 - -11.24)
Day of week: Thursday	-0.4239 (-2.427 - 1.579)	-15.45*** (-20.17 - -10.73)
Day of week: Friday	0.8397 (-1.203 - 2.882)	-17.88*** (-22.57 - -13.19)
Day of week: Saturday	0.3223 (-1.712 - 2.357)	-7.543** (-12.23 - -2.853)
Week: 28	2.717 (-0.8129 - 6.246)	
Week: 29	2.338 (-1.242 - 5.918)	2.948 (-5.181 - 11.08)
Week: 30	1.659 (-1.948 - 5.266)	4.582 (-4.522 - 13.69)
Week: 31	2.272 (-1.406 - 5.951)	1.095 (-8.572 - 10.76)
Week: 32	0.02072 (-3.732 -	-0.2004 (-10.13 -

	3.773)	9.732)
Week: 33	0.09955 (-3.788 - 3.988)	4.296 (-6.068 - 14.66)
Week: 34	1.839 (-2.21 - 5.889)	11.55* (0.549 - 22.55)
Week: 35	2.566 (-1.78 - 6.913)	12.05* (0.03877 - 24.06)
Week: 36	0.4576 (-4.584 - 5.499)	2.838 (-12.16 - 17.84)
Week: 37	-2.544 (-10.47 - 5.384)	5.45 (-13.43 - 24.33)
Week: 38	6.382 (-7.842 - 20.61)	26.47 (-20.91 - 73.84)
Week: 39	-0.9501 (-24.23 - 22.33)	
Age	0.1275*** (0.08308 - 0.1719)	0.02187 (-0.1686 - 0.2124)
Sex: Male	0.09496 (-1.12 - 1.31)	-21.98*** (-27.12 - -16.83)
Sex: Other	-7.845 (-17.63 - 1.937)	43.98 (-11.45 - 99.41)
Ethnicity: Black / African / Caribbean / Black British	1.09 (-1.558 - 3.738)	-24.04*** (-35.64 - -12.44)
Ethnicity: Mixed / Multiple Ethnic Groups	-3.239+ (-6.538 - 0.0603)	8.335 (-7.42 - 24.09)
Ethnicity: Other Ethnic Group	-0.3835 (-5.519 - 4.752)	18.68 (-5.564 - 42.93)
Ethnicity: Private	-4.487* (-8.603 - -0.3701)	13.76 (-4.546 - 32.07)
Ethnicity: White	-4*** (-5.607 - -2.393)	19.23*** (12.57 - 25.89)
Education: No Qualification	-0.8378 (-2.739 - 1.063)	-6.125 (-14.24 - 1.992)
Education: Other Qualification	0.6393 (-0.6721 - 1.951)	-5.088+ (-10.53 - 0.3538)
Education: Prefer Not To Say	-0.7427 (-2.835 - 1.349)	-13.18** (-22.5 - -3.859)
Education: Unknown	-4.39+ (-9.166 -	9.14 (-9.006 - 27.29)

	0.3863)	
BMI	-0.1049** (-0.1775 - -0.03222)	-0.8718*** (-1.185 - -0.5585)
Device: Fitbit	-0.5641 (-2.472 - 1.344)	12.39*** (6.367 - 18.42)
Device: Garmin	1.797 (-1.981 - 5.575)	50.46*** (40.11 - 60.81)
Device: Google Fit	1.225 (-1.955 - 4.406)	-25.75 (-81.36 - 29.86)
Device: HealthKit	0.2683 (-1.374 - 1.911)	17.98*** (8.549 - 27.41)
Deprivation	0.2315+ (-0.01965 - 0.4827)	1.475** (0.4618 - 2.488)
Household size: 2	0.7456 (-0.5742 - 2.065)	5.933* (0.5172 - 11.35)
Household size: 3	2.046+ (-0.298 - 4.389)	-0.8608 (-10.58 - 8.862)
Household size: 4	0.4124 (-4.567 - 5.392)	4.669 (-15.82 - 25.16)
Household size: 5	5.181 (-4.552 - 14.91)	36.55 (-13.15 - 86.26)
Household size: 16	7.682 (-18.7 - 34.06)	80.04 (-15.36 - 175.4)
Household size: 8	-8.209 (-32.39 - 15.97)	40.29 (-51.51 - 132.1)
Baseline	0.585*** (0.5568 - 0.6132)	0.1206*** (0.09273 - 0.1486)
Missing baseline	89.4*** (84.8 - 94)	60.46*** (46.19 - 74.73)
Device: Error		-64.93** (-105.6 - -24.21)
N	4253	12114

Table H13: Overview of treatment effects for all subgroup regressions

Subgroup category	Outcome variable	intervention effect	95% CI	Unadj. p value	N
Age Group - Above median	Steps	84	-176 - 345	0.527	2532

(above 41)					
Age Group - Below median (41 or younger)	Steps	-82	-431 - 267	0.645	1402
Age Group - Above median (above 41)	MVPA	1.51	0.02 - 3	0.048	2554
Age Group - Below median (41 or younger)	MVPA	0.67	-1.21 - 2.55	0.483	1433
Age Group - Above median (above 41)	Fibre	0.397	0.076 - 0.718	0.016	3271
Age Group - Below median (41 or younger)	Fibre	0.291	-0.046 - 0.628	0.09	2812
Age Group - Above median (above 41)	fruitveg	26.24	16.43 - 36.04	<0.001	3287
Age Group - Below median (41 or younger)	fruitveg	15.09	5.54 - 24.64	0.002	2862
Age Group - Above median (above 41)	Sugar	-0.0009	-0.0052 - 0.0035	0.696	3310
Age Group - Below median (41 or younger)	Sugar	-0.0033	-0.0089 - 0.0023	0.246	2811
Age Group - Above median (above 41)	Fat	-0.0035	-0.0063 - -0.0008	0.013	3281
Age Group - Below median (41 or younger)	Fat	0.0006	-0.0024 - 0.0037	0.68	2792
Gender - Female	Steps	175	-70 - 420	0.161	2746
Gender - Male	Steps	-311	-700 - 78	0.117	1183
Gender - Female	MVPA	0.96	-0.43 - 2.35	0.176	2777
Gender - Male	MVPA	1.43	-0.65 - 3.51	0.179	1205
Gender - Female	Fibre	0.394	0.133 - 0.654	0.003	4143
Gender - Male	Fibre	0.325	-0.126 - 0.776	0.162	1932
Gender - Female	fruitveg	20.48	12.23 - 28.73	<0.001	4187
Gender - Male	fruitveg	22.37	10.65 - 34.09	<0.001	1954

Gender - Female	Sugar	-0.0022	-0.0063 - 0.002	0.302	4169
Gender - Male	Sugar	-0.0008	-0.0069 - 0.0053	0.797	1944
Gender - Female	Fat	-0.0015	-0.0039 - 0.0009	0.229	4138
Gender - Male	Fat	-0.0014	-0.005 - 0.0022	0.441	1927
Deprived Group - Less Deprived	Steps	-47	-335 - 240	0.746	2156
Deprived Group - More Deprived	Steps	206	-115 - 527	0.21	1731
Deprived Group - Less Deprived	MVPA	0.66	-1.01 - 2.33	0.438	2172
Deprived Group - More Deprived	MVPA	1.7	0.04 - 3.37	0.046	1768
Deprived Group - Less Deprived	Fibre	0.244	-0.089 - 0.578	0.151	3097
Deprived Group - More Deprived	Fibre	0.406	0.066 - 0.747	0.02	2899
Deprived Group - Less Deprived	fruitveg	25.3	15.31 - 35.29	<0.001	3136
Deprived Group - More Deprived	fruitveg	16.8	6.9 - 26.7	0.001	2923
Deprived Group - Less Deprived	Sugar	0.0007	-0.0039 - 0.0053	0.756	3131
Deprived Group - More Deprived	Sugar	-0.0047	-0.0104 - 0.0009	0.098	2901
Deprived Group - Less Deprived	Fat	-0.0018	-0.0047 - 0.001	0.205	3101
Deprived Group - More Deprived	Fat	-0.0011	-0.0041 - 0.0019	0.464	2885
FV Group - High	Steps	-96	-496 - 305	0.64	969
FV Group - Low	Steps	80	-166 - 326	0.522	2170
FV Group - High	MVPA	1.25	-1.17 - 3.67	0.31	975

FV Group - Low	MVPA	0.95	-0.39 - 2.28	0.165	2197
FV Group - High	Fibre	0.335	-0.201 - 0.87	0.221	1280
FV Group - Low	Fibre	0.362	0.104 - 0.621	0.006	4740
FV Group - High	fruitveg	24.03	5.31 - 42.75	0.012	1287
FV Group - Low	fruitveg	20.54	13.42 - 27.66	<0.001	4804
FV Group - High	Sugar	-0.0034	-0.0097 - 0.003	0.303	1314
FV Group - Low	Sugar	-0.0014	-0.0056 - 0.0028	0.508	4740
FV Group - High	Fat	-0.0005	-0.0047 - 0.0037	0.819	1306
FV Group - Low	Fat	-0.0016	-0.004 - 0.0007	0.173	4703
MVPA Group - Active	Steps	8	-324 - 340	0.962	1606
MVPA Group - Inactive	Steps	79	-206 - 364	0.589	2328
MVPA Group - Active	MVPA	-0.66	-2.71 - 1.39	0.528	1622
MVPA Group - Inactive	MVPA	2.12	0.76 - 3.48	0.003	2365
MVPA Group - Active	Fibre	0.277	-0.126 - 0.68	0.178	2010
MVPA Group - Inactive	Fibre	0.394	0.109 - 0.679	0.007	4073
MVPA Group - Active	fruitveg	26.91	14.73 - 39.09	<0.001	2028
MVPA Group - Inactive	fruitveg	18.36	10.06 - 26.67	<0.001	4121
MVPA Group - Active	Sugar	0.0008	-0.0049 - 0.0065	0.786	2030
MVPA Group - Inactive	Sugar	-0.0028	-0.0073 - 0.0016	0.21	4091
MVPA Group - Active	Fat	-0.0018	-0.0053 - 0.0016	0.301	2012

MVPA Group - Inactive	Fat	-0.0012	-0.0037 - 0.0014	0.372	4061
Ethnicity Group - Asian / Asian British	Steps	446	-100 - 993	0.11	731
Ethnicity Group - Black / African / Caribbean / Black British	Steps	468	-307 - 1244	0.238	254
Ethnicity Group - Mixed / Multiple Ethnic Groups	Steps	-432	-1590 - 725	0.466	132
Ethnicity Group - Mixed / Multiple Ethnic Groups	Steps	-432	-1590 - 725	0.466	132
Ethnicity Group - Private	Steps	345	-1410 - 2099	0.701	83
Ethnicity Group - White	Steps	-82	-331 - 167	0.519	2698
Ethnicity Group - Asian / Asian British	MVPA	0.72	-2.06 - 3.51	0.611	735
Ethnicity Group - Black / African / Caribbean / Black British	MVPA	-1.68	-5.8 - 2.45	0.427	262
Ethnicity Group - Mixed / Multiple Ethnic Groups	MVPA	2.15	-4.65 - 8.94	0.537	133
Ethnicity Group - Mixed / Multiple Ethnic Groups	MVPA	2.15	-4.65 - 8.94	0.537	133
Ethnicity Group - Private	MVPA	4	-5.73 - 13.72	0.423	83
Ethnicity Group - White	MVPA	1.37	-0.06 - 2.79	0.061	2736
Ethnicity Group - Asian / Asian British	Fibre	0.247	-0.376 - 0.869	0.438	1091
Ethnicity Group - Black / African / Caribbean / Black British	Fibre	1	-0.001 - 2.001	0.051	407
Ethnicity Group - Mixed / Multiple Ethnic Groups	Fibre	0.972	-0.327 - 2.271	0.144	217
Ethnicity Group - Private	Fibre	0.399	-1.573 - 2.371	0.693	127

Ethnicity Group - White	Fibre	0.288	0.02 - 0.555	0.035	4162
Ethnicity Group - Asian / Asian British	fruitveg	7.81	-9.39 - 25	0.374	1118
Ethnicity Group - Black / African / Caribbean / Black British	fruitveg	19.39	-8.77 - 47.56	0.178	407
Ethnicity Group - Mixed / Multiple Ethnic Groups	fruitveg	38.48	7.08 - 69.89	0.017	225
Ethnicity Group - Private	fruitveg	5.2	-55.9 - 66.29	0.868	132
Ethnicity Group - White	fruitveg	25.24	16.93 - 33.55	<0.001	4187
Ethnicity Group - Asian / Asian British	Sugar	0.0039	-0.0039 - 0.0117	0.328	1112
Ethnicity Group - Black / African / Caribbean / Black British	Sugar	-0.0235	-0.0402 - -0.0068	0.006	409
Ethnicity Group - Mixed / Multiple Ethnic Groups	Sugar	-0.0178	-0.0377 - 0.002	0.08	223
Ethnicity Group - Private	Sugar	-0.0047	-0.0337 - 0.0242	0.749	129
Ethnicity Group - White	Sugar	-0.0004	-0.0047 - 0.0038	0.84	4167
Ethnicity Group - Asian / Asian British	Fat	0.0032	-0.0018 - 0.0082	0.208	1095
Ethnicity Group - Black / African / Caribbean / Black British	Fat	0.0029	-0.0054 - 0.0111	0.492	405
Ethnicity Group - Mixed / Multiple Ethnic Groups	Fat	0.0075	-0.0039 - 0.0189	0.201	224
Ethnicity Group - Private	Fat	-0.0124	-0.0278 - 0.003	0.117	123
Ethnicity Group - White	Fat	-0.0028	-0.0053 - -0.0004	0.024	4147

Note: Each row in this table provides the treatment effect estimate for the given outcome variable from a regression over only the named subpopulation. Full regression tables for subgroup analysis are available upon request.

Appendix I: Obstacles to user engagement

Obstacles to initial engagement

Some initially engaged participants noted initial **potential concerns about the legitimacy** of the Financial Incentives Scheme, given that it was offering financial rewards. However, these concerns were ultimately overcome by the clear link to the Government, and the use of professional branding across marketing for the intervention.

> *"It sounded a bit dodgy when [person] mentioned it, but because we had a letter from the council as well, it was like, okay, obviously it's above board."* - Fully engaged user, 'Low' reward arm

A further obstacle identified was that **the intervention would be disappointing**, comprising generic advice and disappointing rewards. However, this was overcome when participants downloaded the app and saw the programme contents and the rewards available.

> *"I just thought it was going to be a very, very extremely generic, oh, the government wants you to be healthy, it's just going to tell you to eat vegetables, blah, blah, blah"* - Churn-out user, Control arm

> *"It just said rewards. Which could have literally been anything. As far as I was aware, it could have just been like a free day pass to a library."* - Churn-out user, Control arm

Obstacles to sustained engagement

A widespread theme identified across sample groups was the **complex and laborious process of redeeming rewards**, involving multiple steps. Participants also reported **restrictions in how they were able to redeem rewards,** such as only being able to redeem rewards in multiples of £5, or being unable to use vouchers for online shopping.

> *"It was quite a convoluted process"* - Fully engaged user, Control arm

Participants also reported frustration when the wearable fitness tracker **did not register certain types of exercise**. For example, the wearable fitness tracker could not be worn in water, so could not track swimming, and didn't record certain activities as exercise, such as static gym equipment.

> *"It just kept saying challenge failed because it wasn't registering it with an exercise."* - Fully engaged user, 'Low' reward treatment arm

Participants in the control arm reported confusion about what points led to and why these were not linked to rewards.

Appendix J: Additional Quant IPE findings

Table J1: Average number of PA and diet challenges completed per week, and average n points earned per week, by subgroup (for participants still active at M5)

Subgroup category		PA completed per week	Diet completed per week	Points per week	Pounds (£) redeemed total
Deprivation	**Most Deprived**	0.25	0.57	61	£89.33
	Least Deprived	0.28	0.60	65	£86.52
Age	**Aged>41**	0.30	0.63	69	£88.05
	Aged<41	0.23	0.53	57	£88.04
Baseline F&V	**High**	0.29	0.67	70	£88.95
	Low	0.26	0.56	61	£87.72
Baseline MVPA	**High**	0.24	0.71	68	£89.30
	Low	0.28	0.52	61	£87.45
Ethnicity	**Asian**	0.25	0.55	59	£83.96
	Black	0.23	0.47	51	£81.62
	Mixed	0.26	0.58	63	£100.00

	Other	0.21	0.55	58	£92.31
	White	0.28	0.61	66	£89.07
Gender	**Female**	0.27	0.60	64	£87.46
	Male	0.26	0.55	60	£89.09
	Other	0.28	0.78	86	£114.34

Table J2: Total points and rewards by intervention arm (for participants still active at M5)

			Control	**Low incentives**	**Medium incentives**	**High incentives**	**Total across groups**
For reference: Number of participants			4,194	1,146	989	1,058	7,387
Earned	Total earned via challenge points	Points equivalent	4,586,472	1,752,710	1,993,507	2,396,809	10,729,498
		Pound (£) equivalent	n/a	£8,763.55	£49,837.675	£83,888.315	£142,489.54
	Total £ for providing data		£225,395	£62,975	£55,885	£60,570	£404,825
Redeemed	Total number of rewards redeemed		16,565	3,904	4,945	6,103	31,517

	Total £ equivalent of rewards redeemed (earned via challenge points and reimbursed for providing data)	£217,528.00	£66,785.62	£101,178.1	£139,462.00	£524,953.7

Note: The pound equivalent in fourth row was computed by multiplying the total points earned via challenges times the financial equivalent for those points per arm (0.005 for the low reward arm; 0.025 for the medium reward arm; 0.035 for the high reward arm). Total across groups has been calculated by summing the values for the low, medium and high reward arm

Figure J1: Average number of PA (left panel) and diet (right panel) challenges completed by week, by intervention arm (sample active at M5)

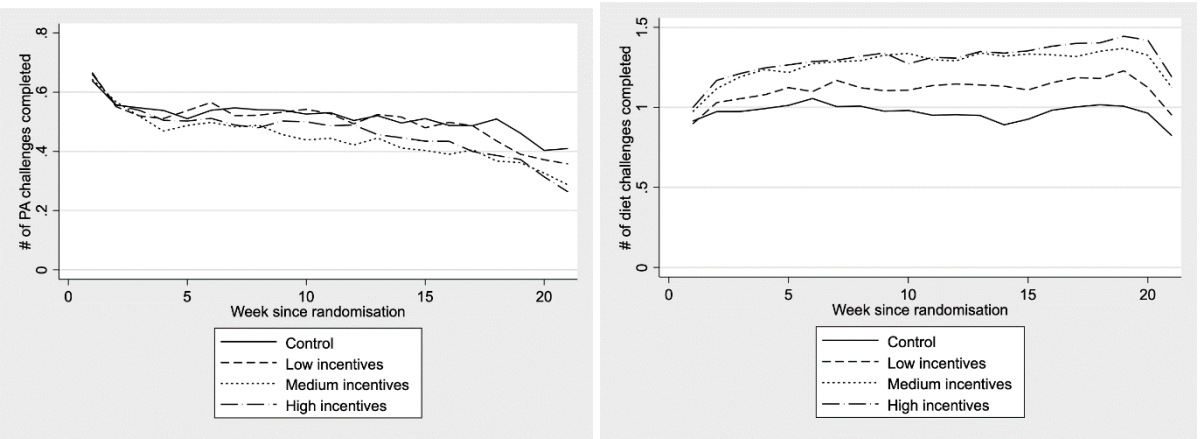


Table J3: Results of regression analysis of user activity and points earned; separate pooled reward arms

	Active (Weeks in which user opened the app, as proportion of trial weeks)	**Cumulative points earned**
(Intercept)	0.596*** (0.568 - 0.624)	482*** (322 - 643)
Treatment arm: Low reward[1]	0.0312*** (0.0178 - 0.0447)	409*** (331 - 487)
Treatment arm: Medium reward	0.0799*** (0.0655 - 0.0943)	927*** (844 - 1011)
Treatment arm: High reward	0.1000*** (0.0862 - 0.1139)	1177*** (1096 - 1257)
Age	0.00437*** (0.00400 - 0.00473)	11.0*** (8.9 - 13.1)
Sex: Male[2]	-0.0248*** (-0.0351 - -0.0144)	-127*** (-187 - -67)
Sex: Other	0.030 (-0.071 - 0.132)	278 (-311 - 867)
Ethnicity: Black / African / Caribbean / Black British[3]	-0.0077 (-0.0287 - 0.0133)	-121+ (-243 - 0)
Ethnicity: Mixed / Multiple Ethnic Groups	-0.014 (-0.042 - 0.014)	141+ (-21 - 302)
Ethnicity: Other Ethnic Group	-0.006 (-0.048 - 0.036)	70 (-174 - 314)
Ethnicity: Private	-0.00051 (-0.03381 - 0.03279)	110 (-83 - 302)
Ethnicity: White	-0.0241*** (-0.0369 -	129*** (55 - 203)

	-0.0113)	
Education: No Qualification[4]	-0.0063 (-0.0222 - 0.0096)	-223*** (-315 - -131)
Education: Other Qualification	0.0021 (-0.0090 - 0.0132)	-80* (-144 - -16)
Education: Prefer Not To Say	-0.0001 (-0.0179 - 0.0177)	-200*** (-303 - -97)
Education: Unknown	-0.010 (-0.054 - 0.034)	-146 (-399 - 107)
BMI	-0.00043 (-0.00105 - 0.00018)	-0.65 (-4.22 - 2.92)
Device: Fitbit[5]	-0.0271*** (-0.0425 - -0.0117)	8.1 (-81.1 - 97.3)
Device: Garmin	-0.0336* (-0.0624 - -0.0047)	-7.7 (-174.4 - 159.0)
Device: Google Fit	-0.0047 (-0.0319 - 0.0226)	175* (18 - 332)
Device: HealthKit	0.0018 (-0.0117 - 0.0154)	-79.7* (-158.2 - -1.3)
Deprivation	0.00097 (-0.00109 - 0.00302)	7.9 (-4.0 - 19.8)
Disability: Prefer not to say[6]	0.0042 (-0.0205 - 0.0289)	69 (-74 - 212)
Disability: Unknown	-0.032 (-0.083 - 0.019)	-428** (-723 - -133)
Disability: Yes	0.0040 (-0.0073 - 0.0152)	0.29 (-64.72 - 65.30)
Household size: 2[7]	0.00024 (-0.01029 - 0.01078)	156*** (96 - 217)
Household size: 3	0.0146+ (-0.0027 - 0.0319)	222*** (122 - 322)

Household size: 4	0.0265 (-0.0065 - 0.0596)	432*** (240 - 623)
Household size: 5	-0.043 (-0.110 - 0.024)	412* (24 - 799)
Household size: 16	-0.0075 (-0.1431 - 0.1282)	-357 (-1141 - 428)
Household size: 8	0.00088 (-0.14406 - 0.14582)	-215 (-1053 - 624)
Baseline MVPA	0.00037*** (0.00017 - 0.00058)	2.69*** (1.50 - 3.89)
Missing Baseline MVPA	0.0346** (0.0086 - 0.0607)	214** (63 - 365)
Baseline F&V	0.000077*** (0.000041 - 0.000114)	0.48*** (0.27 - 0.69)
Missing Baseline F&V	0.033 (-0.011 - 0.078)	157 (-101 - 414)
N (observations)	7385	7385
*95% CI in brackets, +: p<0.1, *:p<0.05, **:p<0.01, ***: p<0.001;*		
[1] *reference level for treatment arm: Control*		
[2] *reference level for sex: Female*		
[3] *reference level ethnicity: Asian / Asian British*		
[4] *reference level for education: Degree or above*		
[5] *reference level for device: Better Health: Rewards tracker*		
[6] *reference level for disability: No*		
[7] *reference level for household size: 1*		

Table J4: Results of regression analysis of user activity and points earned; combined pooled reward arms

	Active	Points earned
(Intercept)	0.598*** (0.570 - 0.626)	498*** (334 - 661)
Treatment[1]: any reward	0.0690*** (0.0595 - 0.0785)	823*** (767 - 879)
Age	0.00433*** (0.00397 - 0.00470)	10.6*** (8.5 - 12.8)
Sex: Male[2]	-0.0249*** (-0.0353 - -0.0145)	-128*** (-189 - -67)
Sex: Other	0.036 (-0.066 - 0.139)	342 (-256 - 941)
Ethnicity: Black / African / Caribbean / Black British[3]	-0.0076 (-0.0287 - 0.0135)	-120.2+ (-243.7 - 3.2)
Ethnicity: Mixed / Multiple Ethnic Groups	-0.013 (-0.041 - 0.015)	156.5+ (-7.4 - 320.3)
Ethnicity: Other Ethnic Group	-0.0047 (-0.0471 - 0.0377)	83 (-165 - 331)
Ethnicity: Private	-0.0016 (-0.0350 - 0.0319)	98 (-98 - 294)
Ethnicity: White	-0.0243*** (-0.0372 - -0.0115)	127*** (52 - 202)
Education: No Qualification[4]	-0.0064 (-0.0224 - 0.0095)	-224*** (-318 - -131)
Education: Other Qualification	0.0028 (-0.0083 - 0.0139)	-72* (-137 - -7)
Education: Prefer Not To Say	0.0012 (-0.0167 - 0.0191)	-185*** (-290 - -80)
Education: Unknown	-0.011 (-0.055 - 0.033)	-160 (-417 - 97)
BMI	-0.00042 (-0.00104 - 0.00020)	-0.45 (-4.08 - 3.18)

Device: Fitbit[5]	-0.0294*** (-0.0448 - -0.0139)	-17 (-107 - 74)
Device: Garmin	-0.0344* (-0.0633 - -0.0055)	-15 (-185 - 154)
Device: Google Fit	-0.0067 (-0.0340 - 0.0206)	153+ (-7 - 313)
Device: HealthKit	0.0013 (-0.0123 - 0.0149)	-84.6* (-164.3 - -4.9)
Deprivation	0.0009 (-0.0012 - 0.0030)	7.0 (-5.1 - 19.1)
Disability: Prefer not to say[6]	0.0036 (-0.0212 - 0.0284)	63 (-82 - 208)
Disability: Unknown	-0.027 (-0.079 - 0.024)	-377* (-677 - -77)
Disability: Yes	0.0038 (-0.0075 - 0.0151)	-2 (-68 - 64)
Household size: 2[7]	0.00022 (-0.01036 - 0.01080)	156*** (94 - 218)
Household size: 3	0.0140 (-0.0033 - 0.0314)	216*** (114 - 317)
Household size: 4	0.0300+ (-0.0032 - 0.0632)	469*** (275 - 663)
Household size: 5	-0.038 (-0.105 - 0.029)	459* (67 - 851)
Household size: 16	0.0017 (-0.1342 - 0.1375)	-271 (-1066 - 524)
Household size: 8	-0.00027 (-0.14582 - 0.14529)	-227 (-1079 - 625)
Baseline MVPA	0.00035** (0.00014 - 0.00056)	2.40*** (1.18 - 3.61)
Missing Baseline MVPA	0.0321* (0.0059 - 0.0583)	186* (33 - 339)
Baseline F&V	0.000078*** (0.000042 - 0.000115)	0.49*** (0.28 - 0.70)
Missing Baseline F&V	0.036 (-0.009 - 0.080)	181 (-80 - 443)
N (observations)	7385	7385

[1] *reference level for treatment arm: Control*
[2] *reference level for sex: Female*
[3] *reference level ethnicity: Asian / Asian British*
[4] *reference level for education: Degree or above*
[5] *reference level for device: Better Health: Rewards tracker*
[6] *reference level for disability: No*
[7] *reference level for household size: 1*

Appendix K: Data collection methods and data quality management

Physical activity outcomes. The PA measures were automatically recorded through the wearable fitness tracker for the full duration of the pilot, but it required participants to sync their device with the app for HUL to collect this data. Participants were reminded to wear and sync the device during crucial data collection weeks/moments. Participants were nudged (via push notifications) to wear and synchronise the wearable fitness tracker during the trial period to improve accuracy of data collection. This was remunerated an additional £5 in each data collection moment.

Dietary outcomes. Diet was measured using Intake24, a retrospective 24 hour dietary recall questionnaire. Completing this questionnaire could take up to 20 minutes. Participants were encouraged to complete two recalls on different days at each data collection point. Participants may provide a recall for a weekday or a weekend day, based on when they completed the questionnaire. As there were four measuring points (baseline, M1, M3, and M5), and each of which participants had two Intake24 surveys to complete, participants could complete eight surveys in total. Participants were remunerated £5 for each of the four surveys at baseline and M1, and £10 for each of the four surveys at M3 and M5.

Measurement time points across the pilot. The TP provided a high-level user journey for participants, from an evaluation point of view, to illustrate when participants were asked to provide data throughout the trial. Broadly speaking, as illustrated in Table K1, there were 4 data collection points during the study:

- the baseline period;
- the 5th week after randomisation for the M1 mark;
- the 13th week after randomisation for the M3 mark;
- the 21st week after randomisation for the M5 mark.

Table K1: 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**: participants are invited to provide their first recall. Participants 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; participants are encouraged to complete the second one within 3 days. The survey will be prominently displayed on the dashboard of the app when participants 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 fitness tracker, 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 fitness tracker, they are encouraged to sync it with the app and provide their PA data as soon as they receive the device. Participants 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:** participants are invited to provide a recall. Participants are informed that the first two recalls they provide in the following two weeks will be remunerated	In the 5th week after randomisation, participants will be nudged to wear the device as long as possible for 7 days and sync the device with the app.

	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 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: participants are invited to provide a recall. Participants are informed that 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*	In the 13th week after randomisation, participants 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 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*

	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).	*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 3 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, participants will be nudged to wear the device for as long as possible for 7 days and sync the device with the app. participants are provided a £10 bonus in total (£5 per sync) to sync their data in the final data collection point. *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 quality management

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

- **During data collection.** Participants were nudged to wear and synchronise the wearable fitness tracker 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 searched for and excluded or replaced outlier values. We were 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 excluded the implausible records and applied further boundary rules (excluding observations that are below the 2.5th or above the 97.5th percentiles) to ensure data fidelity.
    - Upon inspection, we found that the incidence of 'anomaly' PA reads in this study was compatible with other studies for the two primary outcomes. Specifically, we observed that for the two primary outcomes, steps and MVPA, 4.8% and 0.05% fall outside of the reasonable range, respectively. The incidence of these anomalies on steps and MVPA is comparable to what is observed in other studies on PA that rely on wearable fitness tracker data, such as Corder et al. (2020), which finds a comparable incidence of 7.8% of invalid assessment of MPVA reads. On the other hand, about 21% of activity records collected up until M3 fell outside the normal ranges for Cals BMR and Physical Activity Level (PAL).

- As discussed in the limitation section, there is limited available research on the topic of reliability of commercial wearable fitness trackers. Fitbit devices were found to provide a relatively accurate measure of free-living steps (within ±10%, 50% of the time) when compared to research-grade accelerometers (Feehan et al. 2018); Garmin activity wearable fitness trackers are also reported to have good to excellent correlation coefficients and acceptable (<10%) mean absolute percentage errors with respect to step count (Evenson & Spade. 2020); However, one study (Redenius et al. 2019) found that Fitbit Flex overestimated MVPA compared to a research wearable fitness tracker (GT3X+ accelerometer).

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

- **During data collection.** Intake24 included automatic checks and nudges to rule out implausible inputs (e.g. portion size, time gap between meals).
- **Before analysis.** We excluded values below the 1st or above the 99th percentile, by treatment arm.
  - This approach is different to what was pre-specified in the TP and was aimed at rectifying an erroneous attribution of the stated data cleaning criteria to the NDNS approach to data cleaning. This changes was conducted after the evaluation team and DHSC were party to the analysis result.
  - The TP stated that, before excluding specific values based on the rule stated above, we would have excluded administrations satisfying any of the following conditions:
    - 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.

- The TP states that this is the approach that the NDNS takes in cleaning Intake24. However, this is a mistake, as the NDNS uses those criteria to identify suspicious survey entries that are then inspected manually and modified at need, but never excluded.
- Following consultation with BIT and the Trial Steering Committee, the DHSC team instructed BIT to re-run the analysis without those rules mistakenly attributed to the NDNS. The main rationale was that these rules (especially the first 2) were likely to exclude valid surveys filled in by participants from socio-deprived backgrounds, skewing the composition of the sample.

Attrition management

Based on the previous literature and on HUL's experience, we expected high levels of attrition. Attrition (participants not providing the wearable fitness tracker 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 was attrition at the specific measurement points:

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

As discussed in section 4.1.6, we observed differential attrition across trial arms. As the ratio of attrition between the control arm and the pooled intervention arm was greater than 1.1, according to the TP we adjusted for the differential attrition using inverse probability weighting (IPW; Schmidt & Woll 2017), 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 trial arms.

Missing data management

Physical activity outcomes

In the main analysis, we defined a 'valid' read as occurring when the individual wore their wearable fitness tracker for at least 6 hours in the day. As it was possible for reads based on at least 6 hours of wear-time to be erroneous due to issues with the wearable fitness trackers (point 1 below), we replaced 'invalid' reads or erroneous daily reads in the last 7 days of the baseline period and the 1 month, 3 months and 5 months measurement weeks as follows, according to the TP:

1. We excluded reads that are below the 2.5th or above the 97.5th percentile within each combination of period (baseline, 1 month, 3 months, 5 months), reward arm (no / low / medium / high reward) and day of the week, for reads based on at least 6 hours of wear-time.

2. We replaced 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.

3. We imputed 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 only looked at earlier reads from the baseline period as potential replacements. We only did this if a user has at least 4 days of valid baseline data.

Sensitivity analysis to physical activity outcomes

As outlined in the text, the sensitivity analysis to the primary analysis on PA used a different approach to data cleaning.

The sensitivity analysis was carried out because, when inspecting data up until month 3, evidence suggested these filtering rules were potentially filtering out a substantial portion of valid data:

1. Wearable fitness tracker time might not have been a good indicator of data quality. The data were often missing or seemed to not always be accurately recorded. One reason for this is that participants with an Apple phone were able to record steps through their phone, without wearing a wearable fitness tracker.

2. Imputing missing data in addition to invalid data could have substantially increased sample size. When looking at month 3 data, more participants than expected synced data in the two weeks before or after the main measurement week, but not during the measurement week. One reason for this likely is that the Intake24 survey was pushed a week before the activity measurement week. Replacing missing data by observations in the 2 weeks before or after can improve our ability to detect a treatment effect through increasing the sample size.

Last, it is worth noticing that the sensitivity analysis presented in the report was designed and agreed with the funder before M5 data were accessible.

Dietary outcomes

Dietary recalls were provided by participants only at key data collection points (baseline, 1 month, 3 months, 5 months). At these data collection points, participants were encouraged to provide two recalls in the same week. According to the TP, we did not impute dietary outcomes.

Appendix L: Trial information sheet

Consent form

Evaluating an incentives scheme to promote healthy eating and physical activity.

We are inviting you to take part in a study called Better Health: Rewards.

- Before deciding whether to take part, please read this information. This will help you to understand why the study is being done, and what it will involve.
- You can discuss it with family and friends if you want to.
- You do not have to take part if you do not want to.
- Ask us if anything is unclear or if you need more information.

Summary

- We want to know if giving out rewards for physical activity and healthy eating will help people be more active and eat more healthily.
- To take part in the study, you must be a resident of Wolverhampton and you must be aged 18 or above.
- **If you have any health concerns about changing your diet or physical activity, consult your local GP or health professional before starting this programme.** Find more information to help you in deciding whether to use the app in our FAQs.
- **This app may not be suitable for people who have or have previously had an eating disorder, or any other health concern which might stop you from changing your diet or physical activity.** If you are not sure whether the app is for you, you should consult your GP or health professional before starting this programme.
- The study has 4 different groups. These groups are: base reward, low value reward, medium value reward, and high value reward. If you decide to take part, you will be randomly allocated to one of these groups.

- Even if you are in the base reward group, there are benefits. You will receive a free fitness tracker (if required) and have access to the Better Health: Rewards app. You can also earn up to £40 by answering surveys during the study.
- We expect the study to run between 5 and 6 months once you have signed up and have been allocated to a group. During this period, we will ask you to continue using the app and sync your activity tracker. We will also ask you to answer surveys in the app.
- We do not expect any side-effects from taking part in the study. If during the study you have any physical or mental health symptoms, speak with a healthcare professional. You may stop using the app and participating in the trial at any point if your health status changes.

You can stop taking part in the study at any time by sending an email to our team customer support team at betterhealthrewards@customersupport.team. You do not need to tell us why you are stopping.

Why are we doing this research?

Healthy eating and physical activity are important for health. For example, walking, exercising, and eating fruit and vegetables regularly can all contribute to good health.

Getting rewards for healthy behaviours can help people to be healthier. For example, using rewards to help people stop smoking is known to be effective. Evidence has shown that rewards can help incentivise people to improve their physical activity and diets, although evidence is mixed on optimum value of incentive and exact details of incentive scheme design. This research will build on that evidence and show how a scheme can work in a local area.

So, we are inviting Wolverhampton residents to take part in a research study. This is to test if rewards can help people eat more healthily and be more active.

What does taking part involve?

If you agree to take part in this research, you will get free access to the Better Health: Rewards app. You will also get a free fitness tracker, if you do not already own one.

At the start, we will ask you to complete a 'baseline period' for up to 2 weeks. During this time, you can learn to use the app and fitness tracker. We will also ask you to answer questions about your diet and physical activity.

After the baseline period, you will be randomly assigned to 1 of 4 groups:

A. Control group: Access to the app and the fitness tracker.

B. Intervention groups: Access to the app and the fitness tracker. Individuals in these groups can also earn financial rewards. There are 3 different values:

    B.1 low value rewards
    B.2 medium value rewards
    B.3 high value rewards

The study will last between 5 and 6 months once you have been signed up and allocated to a group. We ask that you continue using the app and wearing and syncing your fitness tracker during this time. We also ask that you answer surveys in the app during this time. You can stop being a part of the study at any time. You do not need to give a reason to stop. Please contact a healthcare professional if you have any concerns about your involvement prior to, or during, the scheme.

During or after the study, we may contact you about this research. For example, we might invite you to an interview. In the interview, we may ask you about your experience of using the Better Health: Rewards app.

In the event we do contact to ask you to participate in interviews about the study and your participation in it, we will give you a separate information sheet which will clearly set out the remit and scope of the research and seek your explicit consent to proceed. Participation in any additional research is voluntary and you can decline to take part at any time without giving a reason.

All participants can earn up to £40 in rewards for completing surveys during the baseline period and throughout the course of the study.

A summary of the results of this research will be available to everyone taking part and the results may be submitted for publication in a scientific journal. The data that support the findings of this research may be made available for future academic use. Due to the nature of this research, all data will be pseudonymised for

publications, which means no personally identifiable information will be made available to third parties without your informed consent.

What are the possible benefits of taking part?

All participants will get a free fitness tracker if they do not already own one. Everyone will also have free access to the Better Health: Rewards healthy lifestyle app which provides personalised physical activity and nutrition challenges and content to help with your everyday health. All participants can also earn up to £40 by answering diet questionnaires throughout the programme. Participants can use their points in a safe and secure online store to redeem a wide range of products and gift cards.

The results of this research may help to improve the services and care given to all UK residents. You will be helping science and helping the health of the nation.

What are the possible disadvantages or risks of taking part?

If you have any health concerns about changing your diet or physical activity, consult your local GP or health professional before starting this programme. Find more information to help you in deciding whether to use the app in our FAQs.

This app may not be suitable for people who have or have previously had an eating disorder, or any other health concern which might stop you from changing your diet or physical activity. If you are not sure whether the app is for you, you should consult your GP or health professional before starting this programme.

We do not expect any side-effects from taking part in the study. If during the study you have any physical or mental health symptoms, speak with a healthcare professional. You may stop using the app and participating in the trial at any point if your health status changes.

The information contained in this app is intended for education purposes only and is not intended to replace and is not professional, medical , or healthcare advice, diagnosis or treatment and should not be used for such purposes. You should not use it to prevent, detect, track, manage or to treat any medical condition, disease or injury. Better Health does not provide individual dietary advice. Better Health does not provide clinical care or clinical decision making.

Do I have to take part?

No. It is your choice. If you do not want to take part, that's OK. Your decision will not change the quality of any care you receive from your doctor or healthcare provider, the NHS, or any other Government services. This will not change your ability to use any Government or related services.

If you do decide to take part, you may stop at any time, and you do not need to give a reason. To stop, send an email to our customer support team. The email is: betterhealthrewards@customersupport.team

Who is organising and funding the research?

HeadUp Systems Ltd are conducting this study on behalf of the Department of Health and Social Care. The Behavioural Insights Team, an organisation which is separate and distinct from Government, Better Health, and the Department of Health and Social Care, and which has been appointed by HeadUp Systems Ltd, is sponsoring the study.

The Department of Health and Social Care is paying for the financial rewards, the fitness tracker, the research, and all other costs.

What will happen to information collected about me during the study?

Your data and information are confidential, except where safeguarding or public interest or vital interests of a subject in accordance with the General Data Protection Regulation (GDPR) apply. We will only ask for information that is necessary for research purposes and for the purposes of providing the app and supporting services to you. All data is stored in the UK / EEA and is not transferred outside the EEA.

Before we ask you to share any information, you will be asked if you consent to participate in the research study. If you have provided consent to participate in the research study, we will ask you to review the Better Health: Rewards Privacy Notice which sets out what data will be collected, why and how it will be used. You will be asked to confirm that you have read and understood this information and agree to proceed. You will only be asked to share your data if you have agreed to proceed.

You can withdraw your consent from the study at any time by following the instructions under "*Do I have to take part?*". If you choose to withdraw your consent to participate in the study, please note that we may continue to process some your data in a pseudonymised* form for analysis purposes. This would not include personal information such as your name, email address or other contact information If you withdraw your consent, we will delete your contact information such as your name and email address and you will not be contacted by researchers.

*Pseudonymisation is where we replace, remove or transform information that identifies an individual. This means it cannot identify an individual without the use of additional information.

Further details can be found in the Better Health Rewards: Privacy Notice, but please note:

- Data will be collected by HeadUp Systems Ltd on behalf of the Department for Health and Social Care for the purposes of research and to provide the pilot service to you. All data will be destroyed by HeadUp Systems Ltd after 6 months following the pilot end.
- Data will be shared with the Behavioural Insights Team for research purposes only. Data will be shared with the Behavioural Insights Team in a pseudonymised form for analysis, but some personal information (for example, your name and email address) will shared with the Behavioural Insights Team to enable them to invite some participants to take part in a voluntary interview about their experience of the study. Please note, if you choose to withdraw your consent before the study has ended, your contact information will not be passed onto BIT and you will not be invited to voluntary interviews. Data is destroyed by BIT after 6 months following the pilot end.
- Data will be published in pseudonymised or anonymised form only.
- Some personal data, such as with your name and address, will be shared with named third parties (Allegro Logistics and Power2Motivate), in order for delivery service providers and incentive partners to fulfil fitness tracker orders and incentives, as chosen by you. Data will be destroyed by third parties after 6 months following the pilot end.

- Some personal data, such as your name and email address will be shared with Freshworks for the purposes of HeadUp Systems Ltd providing you with participant support during the course of the pilot. Data will be destroyed after 6 months following the pilot end.
- Your phone number and address will be shared with UK Gov's Notify service and Address.io which are used to send your SMS messages to validate your account and to verify your eligibility to take part in the programme respectively.
- Data will be shared with the Department for Health and Social Care in pseudonymised form to inform further research and analysis. DHSC will destroy the data after 8 years.
- Pseudonymised data may be shared by the Department for Health and Social Care to the Office for National Statistics (ONS) for the purpose of securely hosting the data so that it can be accessed by accredited researchers for further research. The data is deleted if not used for a period of 2 years. If it had previously been accessed for research then it is archived for 5 years from the last date of access to enable analysis to be repeated. It is thereafter deleted. Pseudonymised data means that information shared will exclude your name and other information that directly identifies you. In some circumstances it might still be possible in principle to identify you from the data that remains (for example if there are only a small number of participants similar to you in other ways). We have strict policies around accessing and using pseudonymised data that are designed to prevent this.
- In the unlikely event we receive information suggesting a safety or safeguarding concern, relevant personal information may be shared in order to keep a child or individual at risk safe from neglect or physical, emotional or mental harm, or if it is protecting their physical, mental, or emotional well-being.
- If we suspect fraudulent activity, we may contact you to ask you about your use of the app. Personal information may also be shared with authorities where such data is necessary for the law enforcement authority to fulfil its law enforcement purposes and with The Department for Health and Social Care's

Anti-Fraud team in the event of actual or reasonably suspected fraudulent behaviour.

Reporting safety incidents and safeguarding disclosure

The app cannot make an accurate assessment of a participant's clinical risk, which means it cannot identify safety incidents.

If at any stage during the research you wish to report a safety incident, contact betterhealthrewards@customersupport.team. Reporting safety incidents will inform the research study, we ask you to report any safety incidents to help us make sure the study is safe for everyone. But please note that Better Health: Rewards will not be able to provide medical advice. If you need medical advice you should contact a health professional.

Confidentiality of the reported incident may need to be broken if information is shared or observed which raises concern for a participant's safety or that of someone else.

Thank you for reading this information. Thank you also for considering taking part in this research

You can get more detailed information about this research through the links below. This includes the terms of use, and how we will use your information and protect your privacy. If you would like to complain about this study, please email [removed]

Contact Details:

HeadUp Systems Ltd: [removed]

Department for Health and Social Care: [removed]

Chief Investigator: [removed]

Version No. 4.2 Date 23rd June 2022

Appendix M: References

Althoff, T., Sosič, R., Hicks, J. L., King, A. C., Delp, S. L., & Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature, 547*(7663), 336–339. https://doi.org/10.1038/nature23018

Aune, D., Giovannucci, E., Boffetta, P., Fadnes, L. T., Keum, N., Norat, T., Greenwood, D. C., Riboli, E., Vatten, L. J., & Tonstad, S. (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. https://doi.org/10.1093/ije/dyw319

Banach, M., Lewek, J., Surma, S., Penson, P. E., Sahebkar, A., Martin, S. S., Bajraktari, G., et al. (2023). The association between daily step count and all-cause and cardiovascular mortality: A meta-analysis. *European Journal of Preventive Cardiology*. https://doi.org/10.1093/eurjpc/zwad229

Barte, J. C. M., & Wendel-Vos, G. C. W. (2015). A systematic review of financial incentives for physical activity: The effects on physical activity and related outcomes. *Behavioral Medicine, 43*(2), 79–90. https://doi.org/10.1080/08964289.2015.1074880

Bradley, J., Simpson, E., Poliakov, I., Matthews, J., Olivier, P., Adamson, A., & Foster, E. (2016). Comparison of INTAKE24 (an online 24-h dietary recall tool) with interviewer-led 24-h recall in 11–24 year-olds. *Nutrients, 8*(6), 358. https://doi.org/10.3390/nu8060358

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cane, J., O'Connor, D., & Michie, S. (2012). Validation of the theoretical domains framework for use in behaviour change and implementation research. *Implementation Science, 7*(1). https://doi.org/10.1186/1748-5908-7-37

Chung, S. J., Ersig, A. L., & McCarthy, A. M. (2017). The influence of peers on diet and exercise among adolescents: A systematic review. *Journal of Pediatric Nursing, 36*, 44–56. https://doi.org/10.1016/j.pedn.2017.04.010

Corder, K., Sharp, S. J., Jong, S. T., Foubister, C., Brown, H. E., Wells, E. K., Armitage, S. M., et al. (2020). Effectiveness and cost-effectiveness of the GoActive intervention to increase physical activity among UK adolescents: A cluster randomized controlled trial. *PLOS Medicine, 17*(7), e1003210. https://doi.org/10.1371/journal.pmed.1003210

Department for Digital, Culture, Media and Sport. (2022, November 28). Healthy eating among adults. https://www.ethnicity-facts-figures.service.gov.uk/health/diet-and-exercise/healthy-eating-of-5-a-day-among-adults/latest/

Forde, I., & Zeuner, D. (2009). Financial incentives to promote social mobility. *BMJ, 339*, b3219–19. https://doi.org/10.1136/bmj.b3219

Foster, E., Lee, C., Imamura, F., Hollidge, S. E., Westgate, K. L., Venables, M. C., Poliakov, I., et al. (2019). Validity and reliability of an online self-report 24-h dietary recall method (Intake24): A doubly labelled water study and repeated-measures analysis. *Journal of Nutritional Science, 8*(29). https://doi.org/10.1017/jns.2019.20

Giles, E. L., Robalino, S., McColl, E., Sniehotta, F. F., & Adams, J. (2014). The effectiveness of financial incentives for health behaviour change: Systematic review and meta-analysis. *PLOS ONE,* 9(3), e90347. https://doi.org/10.1371/journal.pone.0090347

Grajek, M., Krupa-Kotara, K., Białek-Dratwa, A., Sobczyk, K., Grot, M., Kowalski, O., & Staśkiewicz, W. (2022). Nutrition and mental health: A review of current knowledge about the impact of diet on mental health. *Frontiers in Nutrition,* 9, 943998. https://doi.org/10.3389/fnut.2022.943998

Green, J., & Thorogood, N. (2004). *Qualitative methods for health research*. Sage Publications.

Haff, N., Patel, M. S., Lim, R., Zhu, J., Troxel, A. B., Asch, D. A., & Volpp, K. G. (2015). The role of behavioral economic incentive design and demographic characteristics in financial incentive-based approaches to changing health behaviors: A meta-analysis. *American Journal of Health Promotion, 29*(5), 314–323. https://doi.org/10.4278/ajhp.140714-lit-333

Hollis, J. L., Sutherland, R., Williams, A. J., Campbell, E., Nathan, N., Wolfenden, L., Morgan, P. J., et al. (2017). A systematic review and meta-analysis of

moderate-to-vigorous physical activity levels in secondary school physical education lessons. *International Journal of Behavioral Nutrition and Physical Activity, 14*(52). https://doi.org/10.1186/s12966-017-0504-0

Hooper, L., Martin, N., Jimoh, O. F., Kirk, C., Foster, E., & Abdelhamid, A. S. (2020). Reduction in saturated fat intake for cardiovascular disease. *Cochrane Database of Systematic Reviews, 8*(August). https://doi.org/10.1002/14651858.cd011737.pub3

Huang, Y., Chen, Z., Chen, B., Li, J., Yuan, X., Li, J., Wang, W., et al. (2023). Dietary sugar consumption and health: Umbrella review. *BMJ, April*, e071609. https://doi.org/10.1136/bmj-2022-071609

Kullgren, J. T., Williams, G. C., Rensicow, K., An, L. C., Rothberg, A., Volpp, K. G., & Heisler, M. (2016). The promise of tailoring incentives for healthy behaviors. *International Journal of Workplace Health Management, 9*(1), 2–16. https://doi.org/10.1108/IJWHM-12-2014-0060

Mair, J. L., Hayes, L. D., Campbell, A. K., & Sculthorpe, N. (2021). Should we use activity tracker data from smartphones and wearables to understand population physical activity patterns? *Journal for the Measurement of Physical Behaviour*, 1–5. https://doi.org/10.1123/jmpb.2021-00

Mema, E., Spain, E. S., Martin, C. K., Hill, J. O., Sayer, R. D., McInvale, H. D., Evans, L. A., et al. (2022). Social influences on physical activity for establishing criteria leading to exercise persistence. *PLOS ONE, 17*(10), e0274259. https://doi.org/10.1371/journal.pone.0274259

Michie, S. (2015). Making psychological theory useful for implementing evidence-based practice: A consensus approach. *Quality and Safety in Health Care, 14*(1), 26–33. https://doi.org/10.1136/qshc.2004.011155

Mitchell, M. S., Goodman, J. M., Alter, D. A., John, L. K., Oh, P. I., Pakosh, M. T., & Faulkner, G. E. (2012). Financial incentives for exercise adherence in adults: Systematic review and meta-analysis. *American Journal of Preventive Medicine, 45*(5). https://doi.org/10.1016/j.amepre.2013.06.017

Mitchell, M. S., Orstad, S. L., Biswas, A., Oh, P. I., Jay, M., Pakosh, M. T., & Faulkner, G. (2019). Financial incentives for physical activity in adults: Systematic review and

meta-analysis. *British Journal of Sports Medicine, 54*(21), bjsports-2019-100633. https://doi.org/10.1136/bjsports-2019-100633

Mitchell, M., White, L., Lau, E., Leahey, T., Adams, M. A., & Faulkner, G. (2018). Evaluating the Carrot Rewards app, a population-level incentive-based intervention promoting step counts across two Canadian provinces: Quasi-experimental study. *JMIR MHealth and UHealth, 6*(9), e178. https://doi.org/10.2196/mhealth.9912

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., et al. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *BMJ, 350*(mar19 6), h1258–58. https://doi.org/10.1136/bmj.h1258

NHS Digital. (2022, December 15). Health survey for England, 2021 part 1. https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021/

Office for National Statistics. (2021a). Wolverhampton Local Authority. 2021 Census Area Profile. https://www.nomisweb.co.uk/sources/census_2021/report?compare=E08000031

Office for National Statistics. (2021b). Exploring local income deprivation. A detailed picture of disparities within English local authorities to a neighbourhood level. https://www.ons.gov.uk/visualisations/dvc1371/#/E08000031

Office for National Statistics. (2023). How life has changed in Wolverhampton: Census 2021. https://www.ons.gov.uk/visualisations/censusareachanges/E08000031/

Office of Disease Prevention and Health Promotion. (2018). *2018 physical activity guidelines advisory committee scientific report*. https://health.gov/sites/default/files/2019-09/PAG_Advisory_Committee_Report.pdf

Paluch, A. E., Bajpai, S., Bassett, D. R., Carnethon, M. R., Ekelund, U., Evenson, K. R., Galuska, D. A., et al. (2022). Daily steps and all-cause mortality: A meta-analysis of 15 international cohorts. *The Lancet Public Health, 7*(3), e219–e228. https://doi.org/10.1016/S2468-2667(21)00302-9

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research, 34*(5 Pt 2), 1189.

Pratt, M. (2021). Health state life expectancies by national deprivation deciles, England. *Office for National Statistics (ONS)*. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthinequalities/bulletins/healthstatelifeexpectanciesbyindexofmultipledeprivationimd/2017to2019

Public Health England. (2018). *National diet and nutrition survey*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/699241/NDNS_results_years_7_and_8.pdf

Public Health England. (2020). *National diet and nutrition survey*. https://assets.publishing.service.gov.uk/media/5fd23324e90e07662b09d91a/NDNS_UK_Y9-11_report.pdf

Purnell, J. Q., Gernes, R., Stein, R., Sherraden, M. S., & Knoblock-Hahn, A. (2014). A systematic review of financial incentives for dietary behavior change. *Journal of the Academy of Nutrition and Dietetics, 114*(7), 1023–1035. https://doi.org/10.1016/j.jand.2014.03.011

Raleigh, V., & Holmes, J. (2021). The health of people from ethnic minority groups in England. *The King's Fund*. https://www.kingsfund.org.uk/publications/health-people-ethnic-minority-groups-england

Rowlands, A. V., Dempsey, P. C., Gillies, C., Kloecker, D. E., Razieh, C., Chudasama, Y., Islam, N., et al. (2021). Association between accelerometer-assessed physical activity and severity of COVID-19 in UK Biobank. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes, 5*(6), 997–1007. https://doi.org/10.1016/j.mayocpiqo.2021.08.011

Sagelv, E. H., Hopstock, L. A., Morseth, B., Hansen, B. H., Steene-Johannessen, J., Johansson, J., Nordström, A., et al. (2023). Device-measured physical activity, sedentary time, and risk of all-cause mortality: An individual participant data analysis of four prospective cohort studies. *British Journal of Sports Medicine, 57*(22), 1457–1463. https://doi.org/10.1136/bjsports-2022-106568

Scarborough, P., Bhatnagar, P., Wickramasinghe, K. K., Allender, S., Foster, C., & Rayner, M. (2011). 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, 33*(4), 527–535. https://doi.org/10.1093/pubmed/fdr033

Scheelbeek, P., Green, R., Papier, K., Knuppel, A., Alae-Carew, C., Balkwill, A., Key, T. J., et al. (2020). Health impacts and environmental footprints of diets that meet the Eatwell Guide recommendations: Analyses of multiple UK studies. *BMJ Open, 10*(8), e037554. https://doi.org/10.1136/bmjopen-2020-037554

Shavit, Y., Roth, Y., & Teodorescu, K. (2021). Promoting healthy eating behaviors by incentivizing exploration of healthy alternatives. *Frontiers in Nutrition, 8*(June). https://doi.org/10.3389/fnut.2021.658793

Schmidt, S. C. E., & Woll, A. (2017). Longitudinal drop-out and weighting against its bias. *BMC Medical Research Methodology, 17*(1). https://doi.org/10.1186/s12874-017-0446-x

Singh, B., Olds, T., Curtis, R., Dumuid, D., Virgara, R., Watson, A., Szeto, K., et al. (2023). Effectiveness of physical activity interventions for improving depression, anxiety and distress: An overview of systematic reviews. *British Journal of Sports Medicine, 57*(18). https://doi.org/10.1136/bjsports-2022-106195

Sport England. (2023, April). *Active lives adult survey November 2021/22 report*. https://sportengland-production-files.s3.eu-west-2.amazonaws.com/s3fs-public/2023-04/Active%20Lives%20Adult%20Survey%20November%202021-22%20Report.pdf?VersionId=In4PN2X02DZ1LF18btgaj5KFHx0Mio9o

Steene-Johannessen, J., Hansen, B. H., Dalene, K. E., Kolle, E., Northstone, K., Møller, N. C., Grøntved, A., et al. (2020). Variations in accelerometry measured physical activity and sedentary time across Europe—Harmonized analyses of 47,497 children and adolescents. *International Journal of Behavioral Nutrition and Physical Activity, 17*(38). https://doi.org/10.1186/s12966-020-00930-x

Strohacker, K., Galarraga, O., & Williams, D. M. (2013). The impact of incentives on exercise behavior: A systematic review of randomized controlled trials. *Annals of Behavioral Medicine, 48*(1), 92–99. https://doi.org/10.1007/s12160-013-9577-4

The Institute for Health Metrics and Evaluation. (2019). *Global burden of disease (GBD)*. https://www.healthdata.org/research-analysis/gbd

Thompson, F. E., & Subar, A. F. (2017). Dietary assessment methodology. *Nutrition in the Prevention and Treatment of Disease*.

UK Active. (2023, September 20). Mental health the main motivation for exercise but three-quarters of Brits don't know recommended activity levels - National Fitness Day survey. https://www.ukactive.com/news/mental-health-the-main-motivation-for-exercise-but-three-quarters-of-brits-dont-know-recommended-activity-levels-national-fitness-day-survey/

Waugh, A. (2021). *Wholesome Wave: Impact evaluation report for the Depth2020 Fruit and Vegetable Incentive Program* (Public Health Theses, 2109).

Yang, Y., Zhao, L.-G., Wu, Q.-J., Ma, X., & Xiang, Y.-B. (2015). Association between dietary fiber and lower risk of all-cause mortality: A meta-analysis of cohort studies. *American Journal of Epidemiology, 181*(2), 83–91. https://doi.org/10.1093/aje/kwu257

Yao, J., Lim, N., Tan, J., Müller, A. M., van Dam, R. M., Chen, C., Tan, C. S., & Müller-Riemenschneider, F. (2022). Evaluation of a population-wide mobile health physical activity program in 696,907 adults in Singapore. *Journal of the American Heart Association, 11*(12). https://doi.org/10.1161/jaha.121.022508

58 Victoria Embankment
London EC4Y 0DS
+44 (0)20 7438 2500
info@bi.team
𝕏 @B_I_Team
www.bi.team

Registered office: 58 Victoria Embankment, London EC4Y 0DS.