The impact of communication campaigns on diversity data disclosure: A randomised controlled trial

Statistical Analysis Plan

Evaluator (institution): Behavioural Insights Team

Principal investigators: Shoshana Davidson, Dr Laure Bokobza

Table 1: Project overview

PROJECT TITLE	The impact of communication campaigns on diversity data disclosure: A randomised controlled trial
DEVELOPER (INSTITUTION)	Kier Group plc
EVALUATOR (INSTITUTION)	Behavioural Insights Team
PRINCIPAL INVESTIGATORS	Shoshana Davidson, Laure Bokobza

Table 2: SAP version history

VERSION	DATE	REASON FOR REVISION	AUTHORS
1.2 [LATEST]			
1.1			

1.0 [ORIGINAL]	27 July 2026		Hannah Chappell

Table of contents

Introduction 4

Sample size and power calculations 6

Analysis 10

Primary outcome analysis 15

Secondary outcome analysis 16

Exploratory analysis 16

Robustness checks 18

Imbalance at baseline 19

Missing data 20

Covariates 20

Outcomes 20

Compliance 20

Presentation of outcomes 21

Appendix 1: Power analysis code 22

Introduction

This randomised controlled trial, funded by the Youth Futures Foundation (YFF), seeks to estimate the effects of communication campaigns aimed at increasing disclosure of diversity characteristics (such as disability status, gender identity, transgender status, social mobility, and other personal characteristics) among employees at Kier Group plc. This trial will target all employees with a valid Kier email address as of June 2026.

This evaluation will be an RCT, with stratified randomisation conducted at the employee level. The aim of the impact evaluation is to assess the impact of a behaviourally-informed communications package (as compared to a standard communications package) on disclosure of in-scope diversity characteristics. The primary analysis focuses on employees ages 16-24, while the secondary analysis focuses on employees ages 25 and above. The evaluation also includes an implementation and process evaluation (IPE) and economic evaluation. These are detailed in the Trial Protocol and are not part of the Statistical Analysis Plan (SAP).

Table 3: Trial design

TRIAL DESIGN, INCLUDING NUMBER OF ARMS	Two-arm, individually-randomised trial
TRIAL TYPE	Effectiveness
EVALUATION SETTING	Emails sent to Kier employees in July 2026
TARGET GROUP	Kier employees with a valid work email address
TREATMENT CONDITION	Behaviourally-informed communications package
COMPARISON CONDITION	Standard communication campaign

UNIT OF RANDOMISATION	Employee (individual)
STRATIFICATION VARIABLES	Previous disclosure (complete disclosure [i.e., disclosed all 9 characteristics] vs. incomplete disclosure [i.e., disclosed 0-8 characteristics]); age range (16-24; 25+) Age group (16-24; 25+)
PRIMARY OUTCOME	Number of in-scope characteristics disclosed by young Kier employees (aged 16-24) in Kier's MS Form, one month after the start of the campaigns. This outcome uses the "core" definition of disclosure, which excludes 'prefer not to say' responses.
SECONDARY OUTCOME	Number of in-scope characteristics disclosed by Kier employees aged 25 and over in Kier's MS Form, one month after the start of the campaigns (core definition)
EXPLORATORY OUTCOME(S)	The following will be analysed for employees aged 16-24 and 25+ separately: 1. Whether employee disclosed all in-scope characteristics one month after the start of the campaigns (core definition) 2. Whether employee disclosed (1) neurodivergence, (2) disability status, (3) social mobility, and (4) ethnicity one month after the start of the campaigns (core definition) a. Whether employee disclosed

	disadvantaged group status[1] for each of these characteristics b. Whether employee disclosed non-disadvantaged group status for each of these characteristics[2] 3. Number of in-scope characteristics disclosed one month after the start of the campaigns. This outcome uses the "expanded" definition of disclosure (i.e. including 'prefer not to say' responses). 4. Whether employee disclosed all in-scope diversity characteristics one month after the start of the campaigns. This outcome uses the "expanded" definition of disclosure, which includes 'prefer not to say' responses.

Sample size and power calculations

BIT will use all employees with a valid Kier email address as of June 2026 as the sample for randomisation. 757 employees between the ages of 16 and 24 had valid Kier email addresses, while 8,885 employees ages 25 and above had valid Kier email addresses at the time of randomisation.

BIT conducted power calculations for the primary outcome (number of in-scope characteristics disclosed). These calculations were performed at the individual employee level as randomisation took place at that level. Analysis was conducted in R and the code can be found in Annex A in the Trial Protocol.

[1] Employees are considered to have disadvantaged group status if they disclose being neurodiverse, having a disability, having low social mobility, or having a non-white ethnic background.

[2] Employees are considered to have non-disadvantaged group status if they disclose being neurotypical, not having a disability, not having low social mobility, or having a white ethnic background.

Table 4: Summary of power calculation assumptions and inputs

Assumption		Rationale
Alpha (significance level)	5%	Standard assumption.
Power	80%	Standard assumption.
Planned sample size	757 for primary outcome; 8,885 for secondary outcome	757 employees ages 16-24 had valid Kier email addresses, while 8,885 employees ages 25 and above had valid Kier email addresses.
Characteristics included	9	All in-scope characteristics are included in the outcome construction: disability status, neurodiversity, ethnicity, gender identity, transgender/gender reassignment, sexual orientation, social mobility, religion and carer status.
Mean number of characteristics declared (conditional on disclosing)	8.1 for primary outcome, 6.9 for secondary outcome	The mean number of characteristics disclosed for employees ages 16-24 (primary outcome) and those aged 25 and up (secondary outcome) are calculated using disclosure data (sourced from Kier's internal HR system) from October 2025.
Standard deviation of the number of characteristics declared (conditional on disclosing)	0.9 for primary outcome, 1.27 for secondary outcome	As with the mean number of characteristics disclosed, standard deviations are calculated using disclosure data from October 2025.

Assumed MS Form response rate for standard communications arm	15%, 30%, 50%	BIT's prior experience with email campaign-based trials suggests that a response rate of 4-6% is typical for "cold" campaigns (i.e., unsolicited emails). We anticipate a much higher response rate (i.e., around 30%) for this trial, given that the communications will come from trusted messengers within Kier and email campaigns of this nature are relatively common within the organisation.[3]

MDESs are expressed in two ways: (1) the number of additional characteristics disclosed per employee, and (2) as a percentage of the gap to full disclosure.

Table 5: Power calculations

	Assumed MS Form response rate for standard communications arm		
	15%	**30%**	**50%**
Primary outcome: Number of in-scope characteristics (out of 9 possible characteristics) disclosed by 16-24-year-old Kier employees (*N* = 757)			
NUMBER OF EMPLOYEES (DIVIDED EQUALLY BETWEEN ARMS)	378		
BASELINE: Mean number of characteristics declared per employee	1.22	2.43	4.05
MINIMUM DETECTABLE EFFECT SIZE (MDES) AS	0.59	0.76	0.84

[3] A comparable study that BIT previously conducted with the engineering firm Arup saw the proportion of staff with fully completed forms increase from 55% to 85%, closing around 64% of the gap to full disclosure. The targets for the present study fall below this, suggesting the required effect sizes are plausible.

NUMBER OF ADDITIONAL CHARACTERISTICS DISCLOSED PER EMPLOYEE			
MDES AS PERCENTAGE OF GAP TO FULL DISCLOSURE	7.6%	11.6%	16.9%
Secondary outcome: Number of in-scope characteristics (out of 9 possible characteristics) disclosed by Kier employees aged 25 and over (*N* = 8,885)			
NUMBER OF EMPLOYEES (DIVIDED EQUALLY BETWEEN ARMS)	4,442		
BASELINE: Mean number of characteristics disclosed in control group	1.04	2.07	3.43
MINIMUM DETECTABLE EFFECT SIZE (MDES) AS NUMBER OF ADDITIONAL CHARACTERISTICS DISCLOSED PER EMPLOYEE	0.15	0.19	0.21
MDES AS PERCENTAGE OF GAP TO FULL DISCLOSURE	1.9%	2.8%	3.8%
Linear model estimated by OLS			

See "Sample size and power calculations" in the Trial Protocol for further discussion of the results of the power calculations.

Analysis

BIT will perform analysis on an intention-to-treat (ITT) basis. All analyses have been specified a priori (before data collection of outcomes took place), and will be conducted in R.

Table 6 summarises the outcomes and analysis plan. More details on outcome data sources are available in the Trial Protocol.

Table 6: Summary of planned analyses and outcomes

UNIT OF RANDOMISATION		Employee
UNIT OF ANALYSIS		Employee
STRATIFICATION VARIABLES		Age group (16-24; 25+), previous disclosure (binary: all characteristics disclosed vs. partial or no disclosure)[4]
PRIMARY OUTCOME	VARIABLE	Number of in-scope characteristics disclosed by young Kier employees (aged 16-24) in Kier's MS Form, one month after the start of the campaigns. This outcome uses the "core" definition of disclosure.
	MEASURE (INSTRUMENT, SCALE, SOURCE)	The fields associated with each characteristic will be converted into a binary indicator of disclosure (0 = no response or "prefer not to say"; 1 = "active" disclosure response), and these indicators will be summed across all in-scope characteristics. The primary outcome measure will therefore be continuous, ranging from 0 (no disclosure responses provided for any characteristic) to 9 (disclosure responses provided for all in-scope

[4] 'Previous disclosure' refers to diversity characteristics captured in Kier's HR system as of October 2025.

		characteristics). This will be calculated using information captured in MS Forms) on disclosure of each in-scope characteristic.
	ANALYSIS (REGRESSION MODEL, COVARIATES)	Linear model. Previous disclosure (binary: all characteristics disclosed vs partial disclosure or no characteristics closed) is used as a stratification variable. In addition, number of characteristics previously disclosed (integer, 0-9) is also included as a covariate.
SECONDARY OUTCOME	VARIABLE	Number of in-scope characteristics disclosed by Kier employees aged 25 and over in Kier's MS Form, one month after the start of the campaigns (core definition).
	MEASURES (INSTRUMENT, SCALE, SOURCE)	Continuous, ranging from 0-9, calculated using information from MS Forms on disclosure of each in-scope characteristic. Scale construction and data source will be the same as for the primary outcome.
	ANALYSIS (REGRESSION MODEL, COVARIATES)	Linear model. Previous disclosure (binary: all characteristics disclosed vs partial disclosure or no characteristics closed) is used as a stratification variable. In addition, number of characteristics previously disclosed (integer, 0-9) is also included as a covariate.
EXPLORATORY OUTCOMES	VARIABLE	1. Whether employee disclosed all in-scope characteristics one month after the start of the

		campaigns (core definition) 2. Whether employee disclosed (1) neurodivergence, (2) disability status, (3) social mobility, and (4) ethnicity one month after the start of the campaigns (core definition): a. Whether employee disclosed disadvantaged group status for each of these characteristics b. Whether employee disclosed non-disadvantaged group status for each of these characteristics 3. Number of in-scope characteristics disclosed one month after the start of the campaigns. This outcome uses the "expanded" definition of disclosure (i.e. including 'prefer not to say' responses). 4. Whether employee disclosed all in-scope diversity characteristics one month after the start of the campaigns (expanded definition)
	MEASURES (INSTRUMENT, SCALE, SOURCE)	For employees aged 16-24 and 25+ separately: 1. Binary, coded as 1 if the employee disclosed all diversity characteristics under the code definition, and 0 if their response

		is blank or "prefer not to say". 2. Binary, coded as 1 if the employee disclosed their (1) neurodivergence, (2) disability status, (3) ethnicity, and (4) social mobility under the core definition, and 0 if their response is blank or "prefer not to say". a. Binary, coded separately for each characteristic as 1 if the employee disclosed disadvantaged group status (i.e. disabled, neurodiverse, disadvantaged ethnicity, low social mobility), and 0 if their response is blank, "prefer not to say" or disclosing majority group status (i.e. not disabled, neurotypical, high social mobility).[5] b. Binary, coded separately for each characteristic as 1 if the employee disclosed non-disadvantaged group status (i.e. not disabled, neurotypical, majority ethnicity, high social mobility), and 0 if their response is blank, "prefer not to say" or disclosing

[5] An employee would be considered as disabled if they answer "Yes" to all three disability questions. An employee would be considered as neurodiverse if they answer "Yes" to this question: "Do you consider yourself to be neurodiverse?". An employee would be considered to have low social mobility if they answer "Routine, manual and service / Semi-routine manual and service / Technical and craft / Long-term unemployed".

		disadvantaged group status (i.e. disabled, neurodiverse, disadvantaged ethnicity, low social mobility). 3. The fields associated with each characteristic will be converted into a binary indicator of disclosure (0 = no response; 1 = either "active" disclosure or "prefer not to say" response provided), and these indicators will be summed across all in-scope characteristics. This outcome will therefore be continuous, ranging from 0 (no disclosure responses provided for any characteristic) to 9 (active or "prefer not to say" responses provided for all in-scope characteristics). 4. Binary for each employee, coded as 1 if the employee disclosed all nine in-scope characteristics under the expanded definition (i.e. not blank fields), and 0 otherwise. All data will be sourced from MS Forms.
	ANALYSIS (REGRESSION MODEL, COVARIATES)	Logistic model to analyse (1) whether employees disclosed all in-scope characteristics one month after the start of the campaigns (core definition), (2) whether employee disclosed [1] neurodivergence, [2] disability status, [3] social mobility, and [4] ethnicity one month after the start of the campaigns (core definition),

		and if so, whether employees disclosed (2a.) disadvantaged or (2b.) non-disadvantaged group status for each of these characteristics, and (4) whether employees disclosed all in-scope diversity characteristics one month after the start of the campaigns (expanded definition). Linear model to analyse (3) number of characteristics disclosed one month after the start of the campaigns (expanded definition). For all analyses, previous disclosure (binary), previous disclosure (integer, 0-9), and employee tenure (categorised as [1] less than 1 year; [2] 1-2 years; [3] 2-5 years; [4] 5-10 years; [5] 10 or more years)[6] as of randomisation (4th June 2026) are included as covariates.[7]

Primary outcome analysis

For the primary outcome (number of in-scope characteristics disclosed by employees ages 16-24), we will estimate the following ordinary least squares regression model:

$$Y_i = \alpha + \beta_T T_{1i} + \beta_2\, previous\ disclosure\ binary_i + \beta_3\, previous\ disclosure\ continuous_i + \beta_4\, tenure_i + \epsilon_i$$

Where:

- Y_i is the number of in-scope characteristics disclosed for *employee*$_i$

[6] Categories are mutually exclusive.

[7] As age group is included as stratification variable, along with previous disclosure (binary), this would typically also be included as a covariate. However, as all analyses are conducted on the two age groups separately, it would be redundant to include age group as a covariate.

- T_{1i} is a binary indicator for employee i being allocated to the treatment group
- β_T is the coefficient of interest, representing the ITT effect
- $previous\ disclosure\ binary_i$ is a binary indicator representing the previous disclosure (stratification variable), coded as 1 if employee i previously disclosed all diversity characteristics and coded as 0 if employee i previously partially disclosed or disclosed no characteristics
- $previous\ disclosure\ continuous_i$ is a continuous variable capturing the total number of diversity characteristics previously disclosed by employee i, ranging from 0 to 9
- $tenure_i$ is a categorical variable capturing the duration of employee i's employment at Kier
- ε_i is the error term

The coefficient of interest is β_T, which is the ITT estimator of the change in number of characteristics disclosed due to being allocated to the behaviourally-enhanced communications package compared to the standard communications package. Heteroskedasticity-robust standard errors will be used for all linear models.

Secondary outcome analysis

For the secondary outcome, we will repeat the analysis using the same regression specification as the primary analysis, except the participant group will be employees ages 25+.

Exploratory analysis[8]

In addition to the primary and secondary analyses, BIT will consider the following outcomes as exploratory analyses for ages 16-24 and 25+ separately.

Whether employee disclosed all in-scope characteristics one month after the start of the campaigns (core definition)

For this outcome, we will estimate the following logistic regression model:

$$logit(p_i) = \alpha + \beta_T T_{1i} + \beta_2\ previous\ disclosure\ binary_i + \beta_3\ previous\ disclosure\ continuous_i$$

[8] All exploratory analyses will be conducted twice: once restricted to employees aged 16–24 and once restricted to employees aged 25 and over.

$$+\beta_4\, tenure_i + \epsilon_i$$

Where:

- p_i is the probability of employee *i* disclosing all in-scope characteristics
- $logit$ is the logit function, such that $logit(p_i) = ln(p_i / 1 - p_i))_i$, where ln is the natural logarithm
- T_{1i} is a binary indicator for employee i being allocated to the treatment group
- β_T is the coefficient of interest, representing the intention-to-treat effect

Whether employee disclosed the following characteristics: (1) neurodivergence status, (2) disability status, (3) ethnicity and (4) social mobility, one month after the start of the campaigns

We will repeat the analysis using the same regression specification as the first exploratory analysis, except the outcomes will be binary variables indicating whether an employee disclosed each of the four stated characteristics or not. These will be treated as four separate outcomes.

a. Whether employee disclosed disadvantaged group status for each of these characteristics

We will repeat the analysis again using the same regression specification as the first exploratory analysis, except the outcomes will be binary variables indicating whether an employee disclosed disadvantaged group status for each of the four relevant characteristics. Again, this will be treated as four distinct outcomes.

b. Whether employee disclosed non-disadvantaged group status for each of these characteristics

We will repeat the analysis using the same regression specification as the first exploratory analysis, except the outcome will be binary variables indicating whether an employee disclosed non-disadvantaged group status for each of the four relevant characteristics. This will also be treated as four separate outcomes.

Alongside these exploratory analyses, we will also compare the proportion of employees in each arm who disclosed any characteristics; the proportion of employees in each arm who did not disclose any characteristics; and the

proportion of employees per arm who disclose disadvantaged status for the four relevant characteristics.

It is important to note that we cannot answer the question of whether employees who identify as belonging to a disadvantaged group are more likely to disclose their diversity characteristics than non-disadvantaged employees. This is because disadvantaged group status is only observable for employees who have chosen to disclose it; for employees who do not disclose, we do not know whether they belong to a disadvantaged group or not. Any comparison of disclosure rates by group would therefore be limited to employees who have already disclosed and so can not tell us anything about the disclosure behaviour of employees who chose not to disclose at all. By examining unadjusted disclosure proportions across all assigned employees rather than conditioning on those who choose to disclose, the descriptive statistics (i.e., proportions) capture sample-wide changes in overall participation and group composition free from response selection bias.

Number of in-scope characteristics disclosed, one month after the start of the campaigns (expanded definition)

We will repeat the analysis using the same regression specification as the primary analysis, except the outcome will be the number of disclosed characteristics using the expanded definition of disclosure (i.e., including 'prefer not to say' responses).

Whether employee disclosed all in-scope diversity characteristics one month after the start of the campaigns (expanded definition)

Finally, we will repeat the analysis using the same regression specification as the first exploratory analysis, except the outcome will be a binary variable indicating whether an employee disclosed all characteristics (including 'prefer not to say' responses) or not.

Robustness checks

BIT will conduct the following robustness checks for the primary and secondary analysis:

- **Supplemental analyses using zero-inflated model:** BIT will conduct exploratory secondary analyses using a zero-inflated model (zero-inflated Poisson), given the anticipated high proportion of zeros in the

outcome data.[9]

- **Quasi-binomial model:** As a further robustness check, BIT will also estimate the primary and secondary analyses using a quasi-binomial model. The outcome variable (number of in-scope characteristics disclosed) is bounded between 0 and 9, reflecting the fixed number of characteristics available for disclosure, and is therefore more naturally conceptualised as a binomial process (i.e., *k* disclosures out of 9 possible characteristics) than as an unbounded count.

Imbalance at baseline

To check for balance **at randomisation**, BIT will report the following characteristics separately for the treatment and control groups:

- Number of employees (unit of randomisation);
- Number of employees in each age group (16-24; 25+);
- Number of employees who previously disclosed all diversity characteristics

To check for imbalance between treatment groups at the **analysis stage**, BIT will calculate normalised differences for each of the following:

- Tenure at Kier (continuous - days of tenure [start date - 1st July 2026, i.e., trial launch]);
- Age group (stratification variable - binary for being aged 25+ instead of 16-24);
- Number of diversity characteristics previously disclosed;
- Previous diversity disclosure (stratification variable - binary for having disclosed all characteristics)

While imbalance in observed characteristics between treatment groups may indicate differential attrition, this could also be due to these characteristics being unbalanced at randomisation. If the normalised difference from any of the tests above exceeds 0.25 between treatment groups, BIT will treat this as if it is a differential attrition issue and apply Lee bounds to determine the

[9] Long, D. L., Preisser, J. S., Herring, A. H., and Golin, C. E. (2014). A Marginalized Zero-Inflated Poisson Regression Model with Overall Exposure Effects. *Stat Med*, 33(29), 5151-5165. DOI: 10.1002/sim.6293

robustness of the results for the primary analysis.

Missing data

Covariates

No missing data are expected for age group, previous diversity disclosure or employee tenure. Age group and previous disclosure (binary) are provided at randomisation and used as stratification variables. Employee tenure (i.e., start date at Kier) and previous disclosure (dichotomous variables for each in-scope characteristic) will be provided alongside the outcome data in early August 2026.

Outcomes

Missing data can arise if an employee does not open the MS Form or opens it but does not complete the form. In either case, this is not treated as missing data. Instead, non-completion is coded as 0, representing non-disclosure of diversity characteristics. However, missing data can also arise if an employee leaves Kier during the trial. Kier's data team will flag individuals who have left the organisation when sharing communications campaign send-out summary data, and these employees will then be removed from the dataset prior to analysis. Attrition from employee turnover will be reported in the evaluation report. Though we do not anticipate differential attrition in employee turnover, we will test for this and report on the number of employees who leave Kier from each treatment arm during the trial and any statistically significant differences in attrition between groups.

Compliance

In this trial, "compliance" with the interventions has several dimensions:

- **Implementation of randomisation:** Each package of communications (standard vs behaviourally-enhanced) should be sent to the correct employees.
- **Spillovers across groups**: Employees in each intervention group may forward emails to peers in the other treatment group, or discuss the content with one another in person.

To check that randomisation has been implemented correctly, Kier will provide BIT with a summary dataset after each email communication wave (initial communication, 1st reminder, 2nd reminder). BIT will match the pseudonymised employee IDs in these datasets against the randomised dataset, to confirm that each employee received the communication corresponding to their assigned condition. We will note and quantify any discrepancies in the evaluation report.

Spillovers will be investigated during the Implementation and Process Evaluation (IPE). During the IPE interviews, employees will be asked whether they discussed the interventions with colleagues or were forwarded email communications by peers. Although these interviews will only be conducted with a subset of the employee population, they will provide insight into the extent to which spillover effects may have impacted the trial's results.

Presentation of outcomes

For outcomes estimated using a linear model, we will present the estimated treatment effect using Hedges' g (dividing the estimated effect from the regression by the pooled standard deviation of the outcome).

When presenting the results graphically, the height of the bar related to the standard communications group will be the unadjusted mean for the standard group; the height of the bar related to the behaviourally-enhanced communications group will correspond to the unadjusted mean for the standard group with the estimated treatment effect applied to it. 95% confidence intervals will be presented around the point estimates for the intervention group.

For outcomes estimated using a logit model, the estimated treatment effects will be in log-odds. BIT will convert estimated effects in log-odds terms into percentage-point estimates by applying the inverse logit transformation to the log-odds coefficients (i.e. invlogit(logit(mean of the control group) + estimate from the logistic regression) - mean of the control group). This approach allows interpretation of the findings on the original probability scale (i.e., likelihood of disclosing diversity characteristics). We will also present risk ratios.

When presenting the results graphically (e.g. in a bar graph), the height of the bar related to the standard communications group will be the proportion of employees who disclose diversity characteristics in the standard communications group; the height of the bar related to the behaviourally-enhanced communications group will correspond to the proportion for the standard communications group with the estimated treatment effect for the intervention group applied to it. 95% confidence intervals will be presented around the point estimates for the intervention group.

Appendix 1: Power analysis code

```
## Script in R markdown

#List key assumptions for power calcs

alpha <- 0.05

power <- 0.8

n_characteristics <- 9 # amended to include ethnicity question

N_total_young <- 757

N_total_other <- 8885

prop_zeros_low <- 0.85 #this is 1 - response rate (5% in this line)

prop_zeros_med <- 0.70

prop_zeros_high <- 0.50

# We assume that conditional on clearing the hurdle, an employee will disclose

# an average of 8.1 (16-24s) or 6.9 (25+) out of 9 characteristics.

mean_outcome_young_non_zeros <- 8.1

mean_outcome_other_non_zeros <- 6.9


# Using our Binomial distribution rule for 8 traits with a 50% fill rate:

# SD = sqrt(9 * 0.9 * 0.1) =

sd_outcome_young_non_zeros   <- 0.9

# SD = sqrt(9 * 0.76 * 0.24) =

sd_outcome_other_non_zeros   <- 1.27


#3.1 Primary analysis - number of characteristics declared, full sample

#Make power calcs function for primary analysis

mdes_cont <- function(N_total, ## mdes_cont (values are assigned below)

                      n_arm,

                      n_characteristics,

                      prop_zeros,
```

```
                mean_outcome_non_zeros,
                sd_outcome_non_zeros,
                alpha,
                power){
                #R_squared){ #ignore/set to 0 as no pre-data

  #Get mean and sd of overall outcome (including zeros)
  mean_outcome <- (1 - prop_zeros) * mean_outcome_non_zeros
  print(paste0("Baseline mean number of characteristics declared per employee = ",
round(mean_outcome,2)))

  var_outcome <- (1 - prop_zeros) * (sd_outcome_non_zeros^2) +
            (prop_zeros * (1 - prop_zeros) * (mean_outcome_non_zeros^2))
  sd_outcome <- sqrt(var_outcome)

  #Get MDES
  nComparisons <- choose(n_arm, 2)
  alpha_adj <- alpha / nComparisons ## BF adjustment
  d <- pwr.t2n.test(n1=floor(N_total/n_arm), n2=floor(N_total/n_arm), sig.level=alpha_adj,
power=power)$d ### calculating standardized effect size
  MDES <- d * sd_outcome
  print(paste0("MDES as n of characteristics per employee = ", round(MDES,2)))

  #***Convert MDESs into different interpretable units
  #% increase on baseline
  MDES_perc <- 100*MDES/mean_outcome
  print(paste0("MDES as % increase on baseline = ", round(MDES_perc,1)))

  #% of gap between mean and max disclosure (useful for interpretation)
```

```
  MDES_percent_gap <- 100*MDES/(n_characteristics - mean_outcome)

  print(paste0("MDES as % of gap to max disclosure = ",round(MDES_percent_gap,1)))


  #% of an arm who need to update all characteristics

  MDES_total_n_characteristics <- MDES * floor(N_total/n_arm)

  MDES_total_N_need_to_update_all <- MDES_total_n_characteristics/n_characteristics

  MDES_perc_in_arm_need_to_update_all <- 100 *
(MDES_total_N_need_to_update_all/floor(N_total/n_arm))

  print(paste0("MDES as % of employees in arm who need to update all characteristics = ",

            round(MDES_perc_in_arm_need_to_update_all, 1)))

}
```