WSP setup and demographics code

Lizzie Jones[1]

02/05/2021

1 WSP - Initial data exploration

1.0.0.1 About R Markdowns

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com). To generate the document of all content, click the **Knit** button. To change the output (e.g. PDF, HTML) change the 'output' at the top to any of the outputs listerd here: https://rmarkdown.rstudio.com/lesson-9.html (https://rmarkdown.rstudio.com/lesson-9.html).

1.1 Data cleaning

First I have conducted some data cleaning to identify any respondents or data points that need to be removed and explain why. First I converted the 'TimeTaken' column to a total number of seconds (SecsTaken) for easier to more easily investigate means and quantiles. I initially focussed on the fastest 10% of respondents across both surveys as they are most likely to have straightlined through the survey. I visually inspected the data, then used the 'careless' package to find evidence of straightliningm 'even-odd' consistencies, and intra-individual response variability (IRV), across the whole survey and within the multiple choice questions (particularly questions 4, 5, 13, 15, 16, 17, 22, 23, 24)

```
### Explore average time taken to complete questionnaire and check for straightlining
all_data$SecsTaken <- as.numeric(lubridate::seconds(all_data$TimeTaken)) # Create numeric column of time taken (seconds)

quantile(all_data$SecsTaken, 0.1) # Fastest 10% of all respondents = completion in 188.9 seconds/ about 3 mins
```

```
##   10% 
## 188.9
```

```
quantile(all_data$SecsTaken, 0.05) # Fastest 5% of all respondents = completion in 117.95 seconds/ about 2 mins
```

```
##     5% 
## 117.95
```

```
quantile(all_data$SecsTaken, 0.025) # Fastest 2.5% of all respondents = completion in 70.975 seconds/ about 1.2 mins
```

```
##   2.5% 
## 70.975
```

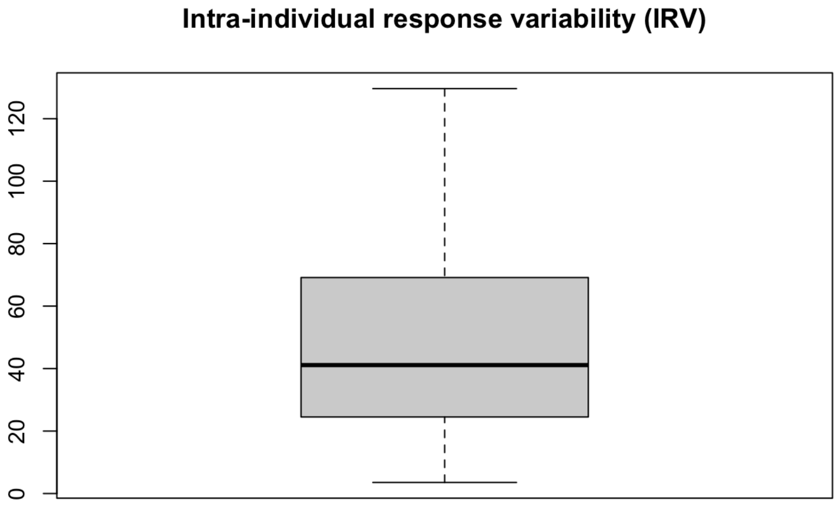
```
fastest_10 <- subset(all_data, SecsTaken < 188.8) # Sample of fastest 10% of all respondents
fastest_5 <- subset(all_data, SecsTaken < 117.95) # Sample of fastest 5% of all respondents
fastest_2.5 <- subset(all_data, SecsTaken < 70.975) # Sample of fastest 2.5% of all respondents

summary(fastest_5$SurveyType) # 96.07% of respondents in fastest 5% are from the NatRep sample
```

```
##    NatRep Proactive 
##       171         7
```

```
summary(fastest_2.5$SurveyType) # 100% of respondents in fastest 2.5% are from the NatRep sample
```

```
##    NatRep Proactive 
##        89         0
```

```
### Checking fastest 5% of respondents
# Checking the fastest 5% for straightlining across whole survey
long_fastest_5 <- longstring(fastest_5, avg = FALSE) # Identifies the longest string of identical consecutive responses for each respondent
evenodd_fastest_5 <- evenodd(fastest_5, rep(5,10)) # Calculates the even-odd consistency score
irv_fast_5 <- irv(fastest_5) # Calculates the intra-individual response variability (IRV)
boxplot(irv_fast_5, main="Intra-individual response variability (IRV)")
```



Intra-individual response variability (IRV)

```
# Checking the fastest 5% for straightlining within each set of mutliple choice questions
# Q5 diet
summary(all_data$Q5_overallscore_diet)
```

```
## Length  Class   Mode 
##      0   NULL   NULL
```

```
summary(fastest_5$Q5_overallscore_diet) ### Not a significant difference in Q5 diet score between all_data, faste
st 5% and 2.5% samples
```

```
## Length  Class   Mode 
##      0   NULL   NULL
```

```
# Q6 habitat
summary(all_data$Q6_habitat_overallscore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.4000  0.6000  0.5873  0.8000  1.0000
```

```
summary(fastest_5$Q6_habitat_overallscore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.0000  0.0000  0.6000  0.4146  0.6000  1.0000
```

```
# Overall knowledge score
summary(all_data$KnowledgeScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   2.100   3.600   3.579   5.000   7.000
```

```
summary(fastest_5$KnowledgeScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   1.000   2.000   2.042   3.000   5.000
```

```
# NCI
summary(all_data$NCI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   38.00   59.00   59.53   82.00  100.00
```

```
summary(fastest_5$NCI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   18.00   31.50   36.25   49.00  100.00
```

```
# Pro-cons behaviours
summary(all_data$ProCoBS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    4.00   12.00   16.00   15.94   20.00   28.00      22
```

```
summary(fastest_5$ProCoBS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    4.00    8.00   12.00   12.55   16.00   28.00       3
```

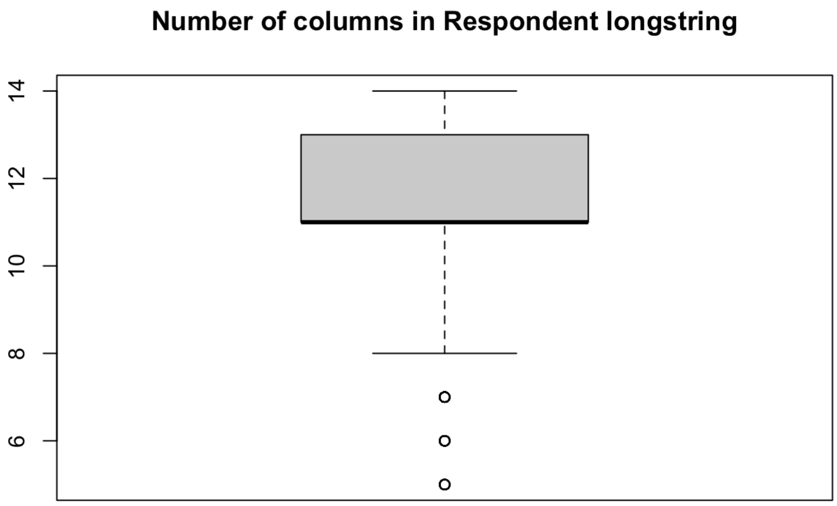
```
# Bird Interest Score
summary(all_data$BirdInterestScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.00   15.00   17.00   16.59   20.00   20.00
```

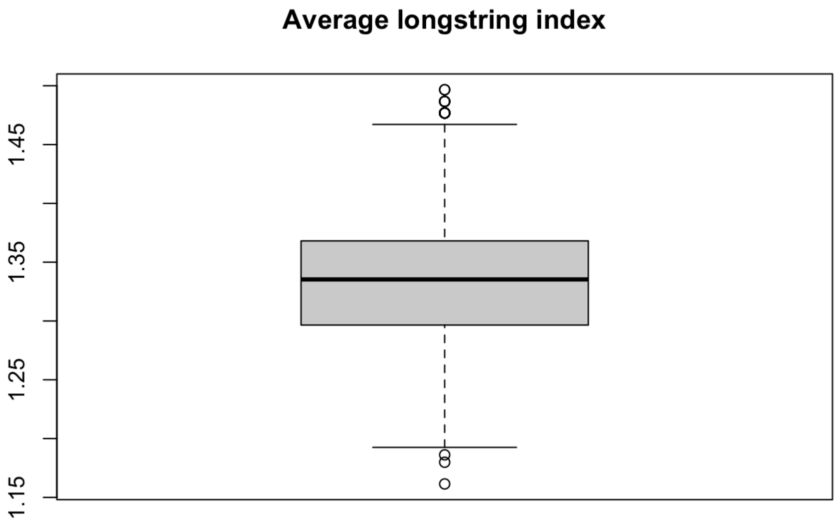
```
summary(fastest_5$BirdInterestScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.00   11.00   12.00   12.63   14.00   20.00
```

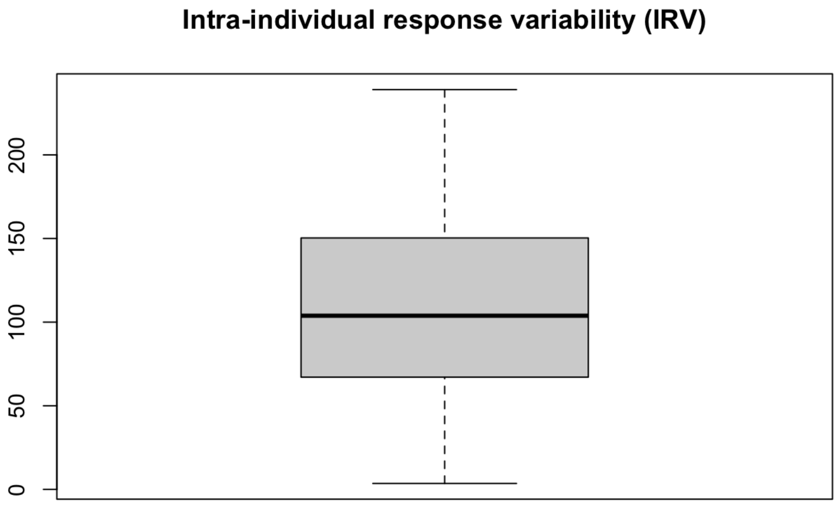
```
##### Data cleaning using the 'Careless' package
# Identifies the longest string of identical consecutive responses for each observation
careless_long <- longstring(all_data, avg = FALSE)
careless_avg <- longstring(all_data, avg = TRUE)
boxplot(careless_avg$longstr, main="Number of columns in Respondent longstring") #produce a boxplot of the longstring index
```



```
boxplot(careless_avg$avgstr, main="Average longstring index")
```



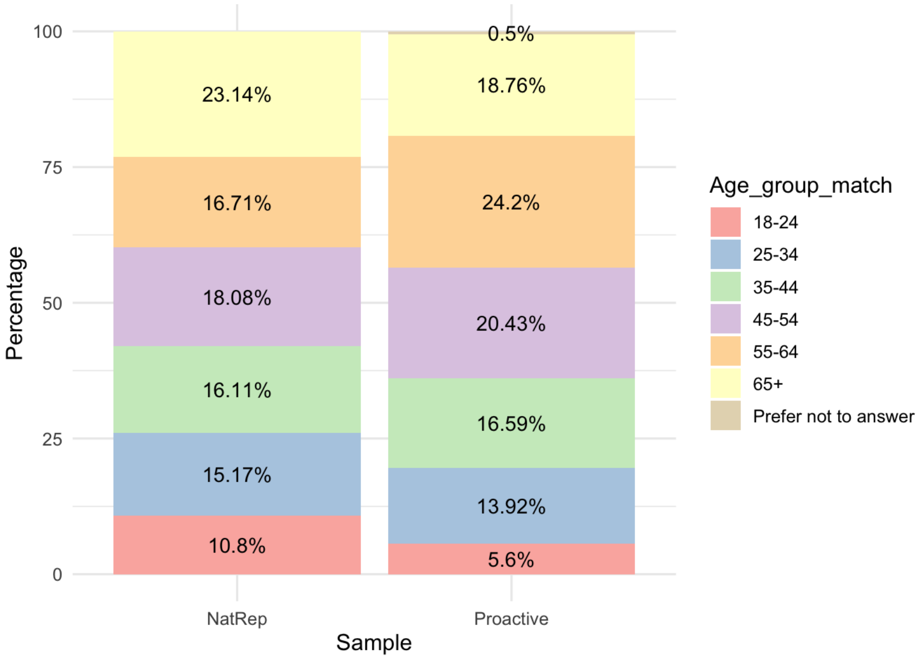
```r
# Calculates the even-odd consistency score
careless_all <- evenodd(all_data, rep(5,10))
careless_alldiag <- evenodd(all_data, rep(5,10), diag = TRUE)

# Calculates the intra-individual response variability (IRV)
irv_total <- irv(all_data)
boxplot(irv_total, main="Intra-individual response variability (IRV)")
```

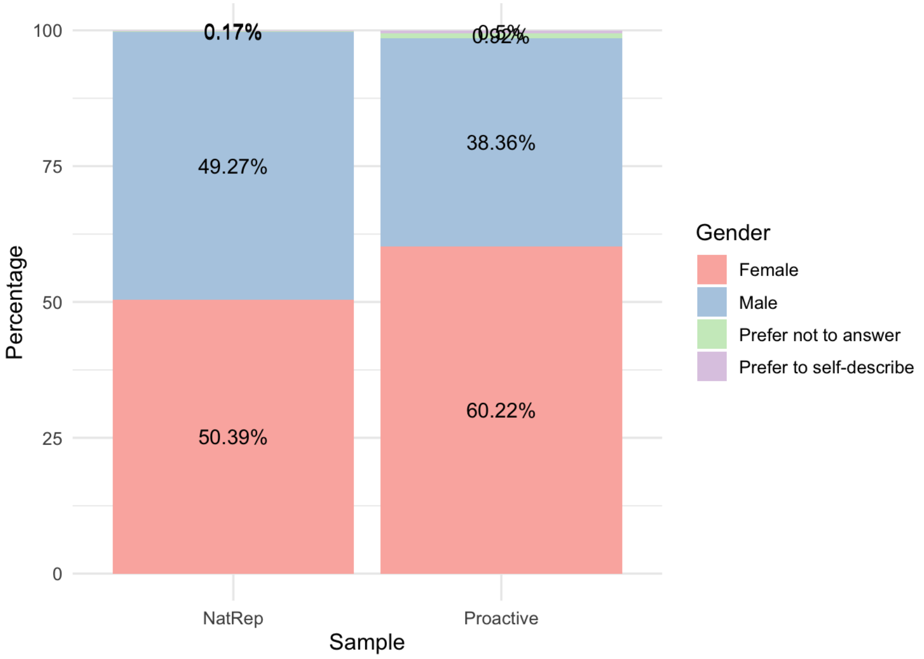


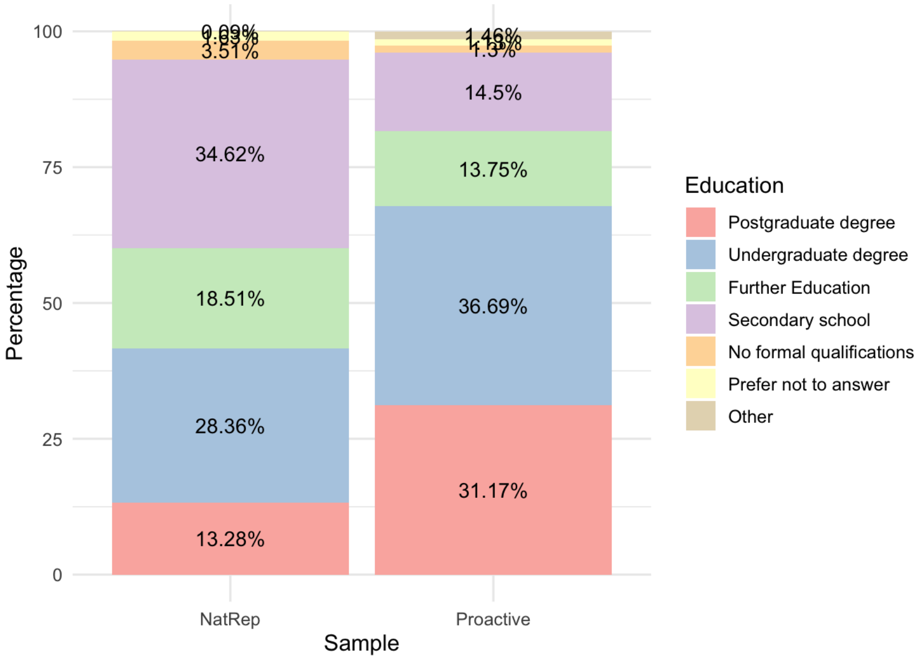
```r
#calculate the irv over all items + calculate the irv for each quarter of the questionnaire
irv_split <- irv(all_data, split = TRUE, num.split = 4)
# boxplot(irv_split$irv4) #produce a boxplot of the IRV for the fourth quarter
```

1.2 Exploring Respondent demographics

The distribution of gender and education is explored and compared between samples using stacked bar plots.







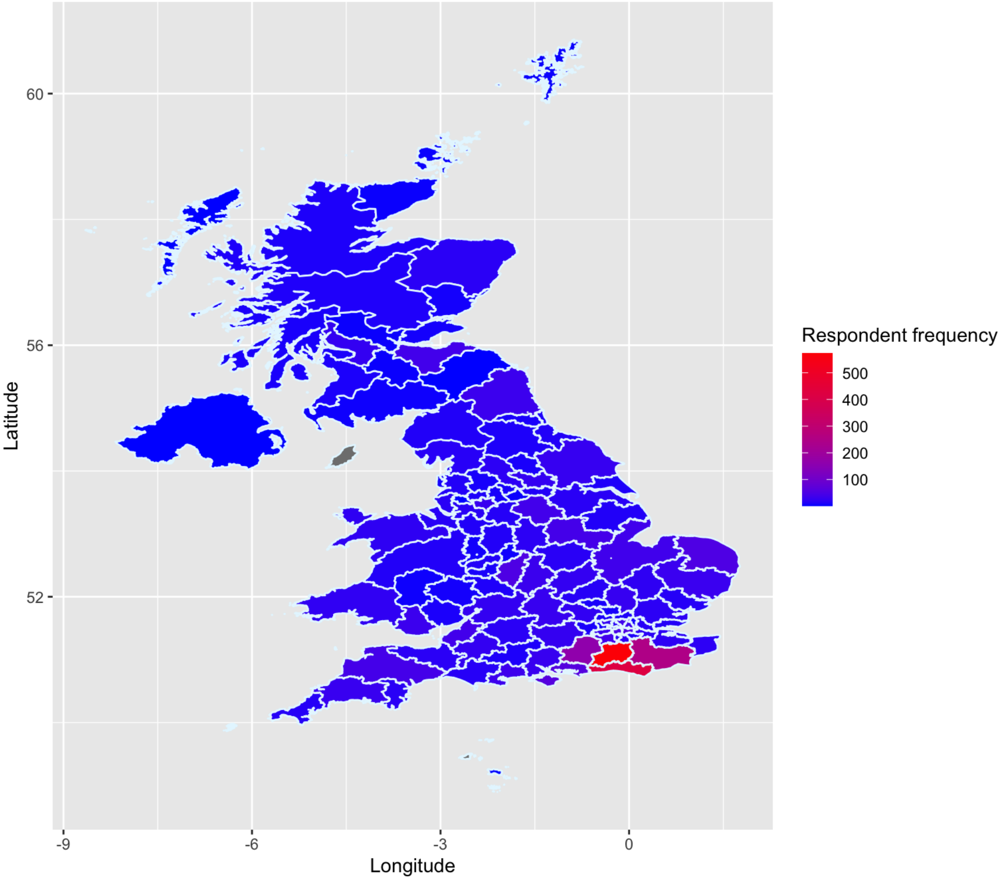
1.2.1 Respondent demographics table

The table below (created using the package “table1”) outlines the demographic characteriscs of each of the two samples, and the overall demographics of all respondents across both samples. For each demographic variable the tables provides a breakdown of the number of respondents within each level/group and the percentage.

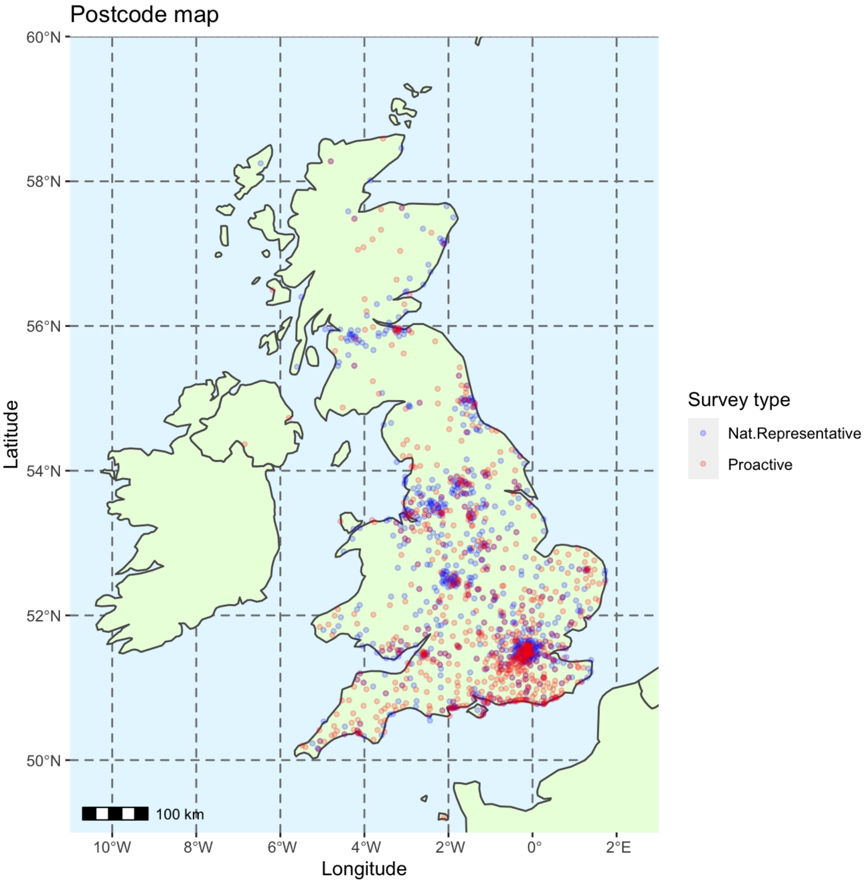
	Nationally rep. (N=1167)	Proactive (N=2393)	Overall (N=3560)
Age group			
18-24	126 (10.8%)	134 (5.6%)	260 (7.3%)
25-34	177 (15.2%)	333 (13.9%)	510 (14.3%)
35-44	188 (16.1%)	397 (16.6%)	585 (16.4%)
45-54	211 (18.1%)	489 (20.4%)	700 (19.7%)
55-64	195 (16.7%)	579 (24.2%)	774 (21.7%)
65+	270 (23.1%)	449 (18.8%)	719 (20.2%)
Prefer not to answer	0 (0%)	12 (0.5%)	12 (0.3%)
Gender			
Female	588 (50.4%)	1441 (60.2%)	2029 (57.0%)
Male	575 (49.3%)	918 (38.4%)	1493 (41.9%)
Prefer not to answer	2 (0.2%)	22 (0.9%)	24 (0.7%)
Prefer to self-describe	2 (0.2%)	12 (0.5%)	14 (0.4%)
Education			
Postgraduate degree	155 (13.3%)	746 (31.2%)	901 (25.3%)
Undergraduate degree	331 (28.4%)	878 (36.7%)	1209 (34.0%)
Further Education	216 (18.5%)	329 (13.7%)	545 (15.3%)
Secondary school	404 (34.6%)	347 (14.5%)	751 (21.1%)
No formal qualifications	41 (3.5%)	31 (1.3%)	72 (2.0%)
Prefer not to answer	19 (1.6%)	27 (1.1%)	46 (1.3%)
Other	1 (0.1%)	35 (1.5%)	36 (1.0%)
Region			
East Midlands	66 (5.7%)	61 (2.5%)	127 (3.6%)
East of England	100 (8.6%)	132 (5.5%)	232 (6.5%)
Greater London	213 (18.3%)	118 (4.9%)	331 (9.3%)
North East	47 (4.0%)	29 (1.2%)	76 (2.1%)
North West	114 (9.8%)	61 (2.5%)	175 (4.9%)
Northern Ireland	0 (0%)	3 (0.1%)	3 (0.1%)
Scotland	96 (8.2%)	56 (2.3%)	152 (4.3%)
South East	174 (14.9%)	1555 (65.0%)	1729 (48.6%)
South West	104 (8.9%)	209 (8.7%)	313 (8.8%)
Wales	58 (5.0%)	40 (1.7%)	98 (2.8%)
West Midlands	106 (9.1%)	54 (2.3%)	160 (4.5%)
Yorkshire and the Humber	89 (7.6%)	75 (3.1%)	164 (4.6%)
Area type			
Rural	225 (19.3%)	1047 (43.8%)	1272 (35.7%)
Sub-urban	548 (47.0%)	858 (35.9%)	1406 (39.5%)
Urban	394 (33.8%)	488 (20.4%)	882 (24.8%)
Release site			
Knepp	5 (0.4%)	432 (18.1%)	437 (12.3%)
Knepp-Wintershall	5 (0.4%)	265 (11.1%)	270 (7.6%)
No	1149 (98.5%)	1378 (57.6%)	2527 (71.0%)
Wadhurst	5 (0.4%)	0 (0%)	5 (0.1%)
Wadhurst Park	0 (0%)	193 (8.1%)	193 (5.4%)
Wintershall	3 (0.3%)	125 (5.2%)	128 (3.6%)

1.2.2 Respondent postcode mapping

Maps of respondent location using different methods: A. Map of first 1 or 2 alphabetical digits, (e.g. SW or N) for all participants with postcode boundaries, in which colour of area reflects density of participants per postcode region, and B. Map of first 4 digits of postcode (e.g., TN28), in which points are colour-coded according to survey type.



Map of first 2 digits of all postcodes (e.g., SW)



Map of first 4 digits of postcode (e.g.,), colour = survey type

1. University of Brighton, l.jones4@brighton.ac.uk (mailto:l.jones4@brighton.ac.uk)↩