Office for National Statistics

{78}

Survey Methodology Bulletin

January 2018

Contents

Office for National Statistics
Swyddfa Ystadegau Gwladol

The Survey Methodology Bulletin is primarily produced to inform staff in the Office for National Statistics (ONS) and the wider Government Statistical Service (GSS) about ONS survey methodology work. It is produced by ONS, and ONS staff are encouraged to write short articles about methodological projects or issues of general interest. Articles in the bulletin are not professionally refereed, as this would considerably increase the time and effort to produce the bulletin; they are working papers and should be viewed as such.

The bulletin is published twice a year and is available as a download only from the ONS website.

The mission of ONS is to improve understanding of life in the United Kingdom and enable informed decisions through trusted, relevant, and independent statistics and analysis. On 1 April 2008, under the legislative requirements of the 2007 Statistics and Registration Service Act, ONS became the executive office of the UK Statistics Authority. The Authority's objective is to promote and safeguard the production and publication of official statistics that serve the public good and, in doing so, will promote and safeguard (1) the quality of official statistics, (2) good practice in relation to official statistics, and (3) the comprehensiveness of official statistics. The National Statistician is the principal advisor on these matters.

www.ons.gov.uk

Edited by: Philip Lowthian

methodology@ons.gov.uk

A Short Guide to using Latent Class Analysis

Debbie Cooper ONS[1], Comfort Ajoku ONS[2]

1. Introduction

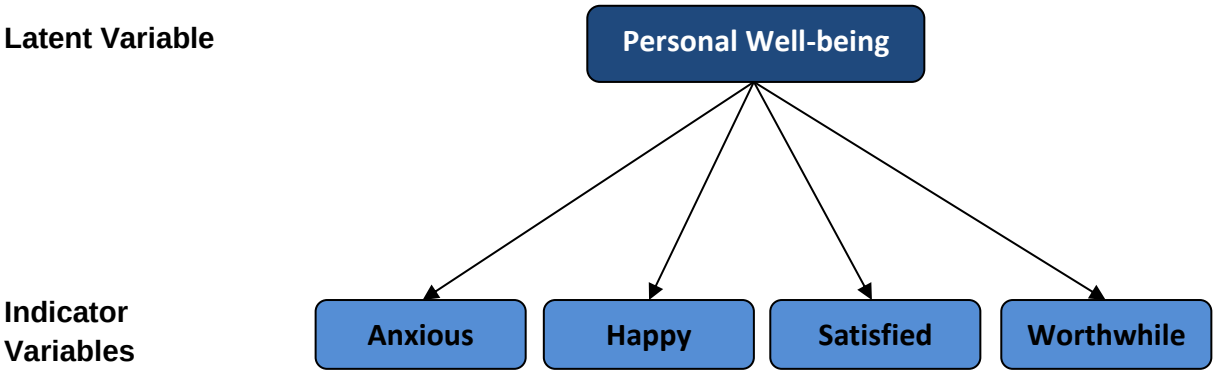
The aim of this guide is to briefly describe Latent Class Analysis (LCA) and how it can be used as well as to provide an applied example. The applied example will illustrate how to carry out an LCA using R (R Foundation for Statistical Computing, 2011) and how to interpret the LCA. Some guidance on carrying out LCA in SAS will also be provided. It is hoped that this will enable colleagues across the GSS to implement LCA in their work.

2. What is Latent Class Analysis?

LCA provides a flexible and powerful approach to categorical data analysis (McCutcheon and Hagenaars, 1997). It is a type of model-based cluster analysis that generally uses the expectation-maximisation (EM) algorithm for model estimation (for further methodological detail see McCutcheon, 1997 and Linzer & Lewis, 2011).

In numerous studies, particularly in social research, researchers are interested in latent variables (variables that cannot be measured directly) e.g. personal well-being or quality of life. These variables tend to be measured by means of a number of indicator (observed) variables. For example, the Office for National Statistics (ONS) uses four indicator variables to measure personal well-being (a latent variable) in the UK[3]:

Figure 1. Indicator variables used to measure personal well-being in the UK



In LCA, the indicator variables are categorical. LCA is used to identify patterns of responses to the indicator variables to create a set of mutually exclusive latent classes (groups of individuals or other units of analysis). Individuals in the same latent class will have similar response patterns to the indicator variables whilst individuals across latent classes tend to have different response patterns to each other. In other words, LCA splits respondents into homogenous groups (latent classes).

[1] Office for National Statistics; Debbie.Cooper@ons.gov.uk
[2] Office for National Statistics; Comfort.Ajoku@ons.gov.uk
[3] Data for personal well-being official statistics are collected by the ONS as part of the Annual Population Survey (APS). See: Office for National Statistics (2016) for data.

LCA has numerous advantages over traditional cluster analysis techniques such as hierarchical cluster analysis and K-means clustering. Some of these advantages include:

- It is model-based unlike other types of cluster analysis which tend to be distance-based. An advantage of this is that there are more formal criteria for choosing the final model when using LCA (for further information see Vermunt & Magidson, 2002).

- It is relatively easy to deal with variables having different scale types (Vermunt & Magidson, 2002).

- In most traditional cluster analysis techniques persons are assigned to clusters on an all-or-none basis. On the other hand, LCA allows membership of a person to each cluster to a certain degree allowing for fractional cluster membership (captured by posterior possibilities).

3. Applied Example

In order to better explain LCA, I will be providing an applied example of LCA using ONS personal well-being data. The aim of this example is not to provide users of personal well-being data with the UK personal well-being profiles[4] but only to illustrate how to carry out and interpret an LCA. **The results shown in this paper are not official statistics** (see Office for National Statistics (2016) for the latest personal well-being official statistics).

LCA can be carried out in many software programs such as SAS®[5], R (R Foundation for Statistical Computing, 2011), STATA[5] (StataCorp LP, 2015), Mplus (Muthén & Muthén, 2011) and Latent Gold (Vermunt & Magidson, 2013), amongst others. For the purposes of this example LCA has been carried out using R and the code is described below. Although the applied example will be coded in R, an explanation regarding how to evaluate the resulting models will be provided and this should be relevant regardless of the software program used. For information regarding how to carry out LCA in SAS please refer to Appendix 1.

This section is split into 4 subsections. The first describes the formula and command required to run an LCA in R. Following this, there is a subsection which briefly describes the personal well-being dataset used in the applied example. Next, the code specified to run the LCA on the personal well-being dataset in R is provided and described. The results of this analysis are then provided along with an explanation of how to interpret the LCA results.

[4] See Chanfreau et al. (2014) for LCA of personal well-being data from the National Survey for Wales.

[5] The LCA procedures in SAS and STATA have not been written and are not supported by SAS Institute Inc. and StataCorp LP. In order to carry out LCA in SAS and STATA, plugins were developed by The Methodology Centre (2015) at Pennsylvania University. These plugins are available to download for free from The Methodology Centre.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc

3.1 Specifying a Basic Latent Class Analysis in R

LCA can be carried out using the R package poLCA (Linzer & Lewis, 2013; Linzer & Lewis, 2011). The formula definition for a basic LCA model is as follows:

```
f <- cbind(Y1, Y2, Y3) ~ 1
```

Y1, Y2 and Y3 are the categorical variables to be included in the LCA. The "~ 1" instructs poLCA to estimate a basic latent class model.

In order to run the poLCA, the basic command used in this paper is as follows:

```
poLCA (formula, data, nclass = 2, maxiter = 50000, graphs = FALSE, na.rm = TRUE,
nrep =  10, verbose = TRUE)
```

Description of the options specified in the command above:

formula: the formula definition 'f' specified above

data: the name of the data frame to be used in the LCA

nclass: the number of latent classes to be calculated in the model. The default is 2 latent classes. poLCA assumes one set of latent classes every time it is run. Therefore, in order to obtain multiple models, each assuming a different number of latent classes, the command must be run a number of times each time specifying a different number of latent classes to be assumed. This will become clearer in the applied example provided below.

maxiter: this is the maximum number of iterations for convergence. If convergence is not achieved before reaching this number of iterations an error message will appear and the analysis will terminate.

graphs: this specifies whether a graph showing the parameter estimates should be produced. The default is FALSE. It takes quite a long period of time for R to run analysis with 4 or more latent classes on large datasets. Therefore, it might be more practical to run the analysis without producing graphs and, if required, only producing a graph for the best fitting model once the resulting models have been compared.

na.rm: this specifies how poLCA handles cases with missing values. If specified as TRUE, those cases are removed by means of listwise deletion before model estimation. If specified as FALSE, cases with missing values are retained. The default is TRUE. Linzer and Lewis (2011) suggest that it is not necessary to delete cases with missing values before estimating the model because poLCA excludes cases with missing values from the calculation.

nrep: this option is used to specify the number of times the model should be estimated using different starting values. It is preferable to set nrep to greater than 1 to ensure that the algorithm finds a global rather than local maximum of the log-likelihood function.

verbose: this indicates whether the results of the model should be output to the screen or not. The default is TRUE.

There are a number of additional options which can be included in this command. For a full list of these options please refer to Linzer and Lewis (2011).

3.2 Dataset Information

As illustrated in Section 2, four personal well-being questions[6] are asked in order to measure personal well-being in the UK, these are:

Overall, how satisfied are you with your life nowadays?

Overall, to what extent do you feel that the things you do in your life are worthwhile?

Overall, how happy did you feel yesterday?

Overall, how anxious did you feel yesterday?

The four personal well-being variables are measured on a scale of 0 to 10, where 0 is 'not at all' and 10 is 'completely'. For the purposes of this analysis, the responses to these questions are categorised as follows:

≥ 0 and ≤ 4 = 1 (Low)

≥ 5 and ≤ 7 = 2 (Medium)

≥ 8 and ≤ 10 = 3 (High)

Please note that these categories are different to those used in the national statistics published by the ONS. Fewer categories than those used for official statistics were required for this analysis in order for meaningful results to be obtained. In this case, the use of too many categories results in poor model fit as there is not enough differentiation between the categories.

[6] Personal well-being official statistics use APS data. However, other UK surveys also collect personal well-being data using the four personal well-being questions. See Office for National Statistics (2017) for further details.

3.3 Latent Class Analysis of Personal Well-being

An LCA of personal well-being data was carried out using the R package poLCA (Linzer & Lewis, 2013; Linzer & Lewis, 2011). The code to run the LCA was specified as follows:

```
library (foreign)
library (MASS)
library (scatterplot3d)
library (poLCA)

data <- read.spss("D:PWB_LCA/ wellbeing.sav")
data1 <- as.data.frame(data)
data2 <- data1[,16:19]
f <- cbind(satisth2, happyth2, anxiouth2, worthth2)~1

wellbeing2 <- poLCA (f, data2, nclass=2, maxiter=50000, graphs=FALSE, nrep=10,
verbose =TRUE)
wellbeing3 <- poLCA (f, data2, nclass =3, maxiter = 50000, graphs = FALSE, nrep =
10, verbose = TRUE)
wellbeing4 <- poLCA (f, data2, nclass =4, maxiter = 50000, graphs = FALSE, nrep =
2, verbose = TRUE)
wellbeing5 <- poLCA (f, data2, nclass =5, maxiter = 50000, graphs = FALSE, nrep =
2, verbose = TRUE)
wellbeing6 <- poLCA (f, data2, nclass =6, maxiter = 50000, graphs = FALSE, nrep =
2, verbose = TRUE)
wellbeing7 <- poLCA (f, data2, nclass =7, maxiter = 50000, graphs = FALSE, nrep =
2, verbose = TRUE)
wellbeing8 <- poLCA (f, data2, nclass =8, maxiter = 50000, graphs = FALSE, nrep =
2, verbose = TRUE)
```

The first step involves specifying the libraries. The 'foreign' package is used to import the SPSS data file into R. poLCA depends on two packages: 'MASS' and 'scatterplot3d' therefore these were specified in addition to the poLCA package. Following specification of the libraries:

- The "data <- read.spss" line reads in an SPSS file containing the personal well-being variables to be used in the analysis.
- "Data1" sets "Data" as a frame so that it can be used for analysis.
- "Data2" is a subset of "Data1" containing only the personal well-being variables to be used in the analysis.
- The function "f" takes the four personal well-being variables and models a basic latent class model with no covariates (as described in Section 3.1 the "~ 1" instructs poLCA to estimate a basic latent class model). Note that if the dataset has more variables than the f function, it will result in an error because the dimensions are not the same.
- "wellbeing2" carries out the latent class analysis for 2 classes (nclass= 2), "wellbeing3", "wellbeing4" and "wellbeing5" run the analysis assuming 3, 4 and 5 latent classes, respectively.
-

- In each of the "wellbeing" commands:
  - the function "f" is specified
  - the data frame "data2" is specified for use in the analysis
  - the number of classes to be assumed in the model is specified
  - maxiter is set to 50000 to ensure convergence is obtained
  - graphs is set to FALSE so as not to produce any graphs (the analysis runs faster this way)
  - nrep is set to 10 in order to ensure that global rather than local maxima are found. As the number of latent classes calculated by the model increases, the analysis takes longer to run. Consequently, for models with 4 or more latent classes, one may wish to start by specifying a small number for nrep (in this case nrep=2 was used for "wellbeing4" and "wellbeing5"). It is possible to determine from the output whether global maxima have been found. If global maxima have been found there is no need to run the analysis with more repetitions. However, if it becomes evident that global maxima have not been found, it would be desirable to re-run the analysis specifying a higher number for nrep in order to obtain global maxima.
  - verbose is set to TRUE in order to output the results to the screen for interpretation

3.4 Interpretation of LCA Results

The "wellbeing2" analysis run in section 3.3 outputs the following results:

```
Model 1: llik = -934514.9 ... best llik = -934514.9
Model 2: llik = -934514.9 ... best llik = -934514.9
Model 3: llik = -934514.9 ... best llik = -934514.9
Model 4: llik = -934514.9 ... best llik = -934514.9
Model 5: llik = -934514.9 ... best llik = -934514.9
Model 6: llik = -934514.9 ... best llik = -934514.9
Model 7: llik = -934514.9 ... best llik = -934514.9
Model 8: llik = -934514.9 ... best llik = -934514.9
Model 9: llik = -934514.9 ... best llik = -934514.9
Model 10: llik = -934514.9 ... best llik = -934514.9

Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$satisth2
           Pr(1)  Pr(2)  Pr(3)
class 1:  0.0008 0.1414 0.8578
class 2:  0.1695 0.7222 0.1083

$happyth2
           Pr(1)  Pr(2)  Pr(3)
class 1:  0.0174 0.1901 0.7925
```

```
class 2:  0.2519 0.5662 0.1819

$anxiouth2
           Pr(1)  Pr(2)  Pr(3)
class 1:  0.7958 0.1541 0.0501
class 2:  0.4752 0.3419 0.1829

$worthth2
           Pr(1)  Pr(2)  Pr(3)
class 1:  0.0014 0.1258 0.8728
class 2:  0.1232 0.6365 0.2403

Estimated class population shares
 0.6368 0.3632

Predicted class memberships (by modal posterior prob.)
 0.6278 0.3722

=========================================================
Fit for 2 latent classes:
=========================================================
number of observations: 303778
number of estimated parameters: 17
residual degrees of freedom: 63
maximum log-likelihood: -934514.9

AIC(2): 1869064
BIC(2): 1869244
G^2(2): 61111.46 (Likelihood ratio/deviance statistic)
X^2(2): 138365.7 (Chi-square goodness of fit)
```

As specified in Section 3.0 these results are not official statistics. They are only provided to illustrate how to carry out and interpret an LCA.

The results are interpreted as follows:

The first part of the output (Model 1 to Model 10 llik and best llik) shows that the latent class model was estimated ten times (as specified by nrep=10) using different starting values. The results assigned to "wellbeing2" will be those estimated for the model with the greatest value of the log-likelihood function (Linzer & Lewis, 2011). In this case, it seems as though the global maximum log-likelihood of -934514.9 was found on the first attempt at fitting the model. Therefore, the Model 1 results will be assigned to "wellbeing2".

The next section of output provides conditional item response probabilities, by outcome variable, for each class. This output shows the probabilities of respondents in each latent class providing a low, medium or high response to the indicator variable in question. The

rows represent the latent classes. The model assumed 2 latent classes in this case, therefore there are 2 rows. The columns indicate the categories (low, medium and high) of the indicator variable. For example, the conditional item response probabilities for the satisfaction variable produced in the 2-class model were as follows:

```
$satisth2
           Pr(1)  Pr(2)  Pr(3)
class 1:  0.0008 0.1414 0.8578
class 2:  0.1695 0.7222 0.1083
```

"satisth2" is the name of the satisfaction variable used in this analysis. In the case of this analysis, as specified in Section 3.2 "low", "medium" and "high" responses to each variable were coded as 1, 2 and 3 respectively. Therefore, in the output shown above, Pr(1) is the probability of a respondent providing a "low" response to the satisfaction variable.

The output shown above can be interpreted as follows: there is a 0.08% chance of a respondent in latent class 1 providing a "low" response to the satisfaction variable; a 14.14% chance of them providing a "medium" response and an 85.78% chance of them providing a "high" response. Therefore, overall respondents in latent class 1 are more likely to have high levels of satisfaction. On the other hand, respondents in latent class 2 are more likely to have medium levels of satisfaction (72.22%).

Taken together, the conditional probability results for all 4 personal well-being variables provided above indicate that respondents in latent class 1 are likely to have high levels of satisfaction (conditional probability = 85.78%), high levels of happiness (conditional probability = 79.25%), low levels of anxiety (conditional probability = 79.58%) and high levels of worth (conditional probability = 0.8728). On the other hand, respondents in latent class 2 are likely to have medium levels of satisfaction (conditional probability = 72.22%), medium levels of happiness (conditional probability = 56.62%), low levels of anxiety (conditional probability = 47.52%) and medium levels of worth (conditional probability = 63.65%).

The "estimated class population shares" section of the output provides the estimated proportions corresponding to the share of observations belonging to each latent class (Linzer & Lewis, 2011). Therefore, in the case of the 2-class model, the share of observations is estimated to be 63.68% in latent class 1 and 36.32% in latent class 2.

The "Predicted class memberships" is another way of estimating the size of the latent classes. This assigns observations to the latent classes using posterior probabilities. Generally, when the values for the "estimated class population shares" and "Predicted class memberships" are similar, this is an indication of good model fit. However, this congruence between values alone should not be used to assess model fit as there are other criteria which should be used to choose the best-fitting model. These are provided in the next section of output called "Fit for 2 latent classes".

The "Fit for 2 latent classes" section of the output indicated the following results:

```
number of observations: 303778
number of estimated parameters: 17
residual degrees of freedom: 63
maximum log-likelihood: -934514.9

AIC(2): 1869064
BIC(2): 1869244
G^2(2): 61111.46 (Likelihood ratio/deviance statistic)
X^2(2): 138365.7 (Chi-square goodness of fit)
```

The "number of observations", "number of estimated parameters", "residual degrees of freedom" and "maximum log-likelihood" are reported in the output. The "number of observations" specifies the number of fully observed cases (i.e. cases without missing values) that were used in the analysis. The "number of estimated parameters" indicates the number of degrees of freedom used by the model. It is worth checking that the "residual degrees of freedom" is not negative. However, a negative residual will normally result in an error message.

The next set of values is useful for comparing models with different numbers of latent classes in order to assess model fit. When models with different numbers of latent classes are compared, the model with the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) is normally chosen as the best-fitting model. This is because a lower value of the information criterion suggests a better balance between model fit and parsimony (Lanza, S.T & Rhoades, B.L., 2013). Sometimes, the AIC and BIC do not indicate the same model as the best-fitting one. In the case of basic exploratory LCA, the BIC is usually more appropriate because of the relative simplicity of the model (Lin and Dayton 1997; Forster 2000).

The Likelihood ratio and Chi-Square statistics may also be used to assess model fit. Like the AIC and BIC the aim is to select models that minimise the Likelihood ratio and Chi-Square statistics whilst also maintaining a low number of parameters. It is important to note that these tests are not useful with sparse data (which is a common problem in LCA) because they violate the chi-square distribution assumption (Linzer & Lewis, 2011; Nylund et. al, 2007). From their study investigating the various methods for assessing model fit, Nylund et. al (2007) concluded that the bootstrap likelihood ratio test is the best method for evaluating model fit, followed by the BIC. Currently, it doesn't seem as though one can calculate the bootstrap likelihood ratio test using the poLCA package. Consequently, it is recommended that, when using poLCA, the BIC is used to choose the best-fitting model. A comparison of the BIC for all models estimated with the personal well-being data is shown in Table 1.

Table 1. Comparison of BIC obtained from the various latent class models estimated using personal well-being data

Number of Latent Classes estimated	BIC
2	1869244
3	1827588
4	1816278
5	1811968
6	1809868
7	1809400
8	1809478

The results in Table 1 indicate that the model with the lowest BIC is the 7-class model. Consequently, out of all the competing models, the model with 7 latent classes fits the data best and would be chosen as the final model for interpretation.

4. Extensions and Uses of Latent Class Analysis

The LCA described above is an exploratory LCA. This is probably the simplest form of LCA. There are many extensions to this which create a multitude of uses for LCA. For example, by including restrictions you can carry out a confirmatory LCA. Covariates could also be added to an LCA. Another extension of LCA is Latent Transition Analysis (SAS software for this is also available from The Methodology Centre) which can be used to assess changes in latent classes over time. Analysis with continuous variables can also be carried out; this is usually referred to as Latent Profile Analysis.

LCA is an extremely useful way of conducting more holistic analysis because it is a type of person-centred analysis as opposed to variable-centred analysis. Apart from being used in final analysis of data, LCA has also been used to assess the quality of questionnaires and the resulting estimates e.g. it has been used to assess classification error (see Biemer & Wiesen, 2002; Biemer 2011); to assess mode effects (see Biemer, 2001); and to evaluate questionnaire items (see Kreuter, Yan & Tourangeau, 2008).

5. Conclusion

LCA is a powerful technique which has a multitude of applications to official statistics. It is becoming very popular because of its many uses and the ease with which it can be carried out. Producers of official statistics are encouraged to apply LCA to their work in order to be able to provide higher quality and more meaningful statistics to users. For those colleagues interested in learning more about LCA, the following sources may be of interest:

McCutcheon, A.L., 1987. Latent Class Analysis. California: Sage Publications.

Magidson, J. and Vermunt, J.K. Latent Class Models. In: Kaplan, D. eds. The SAGE Handbook of Quantitative Methodology for the Social Sciences. California: Sage Publications. pp.175-198.

References

Biemer, P. P., 2001. Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing. *Journal of Official Statistics*, 17(2), p.295-320.

Biemer, P. P. and Wiesen, C., 2002. Measurement Error evaluation of self-reported drug use: a latent class analysis of the US Household National Survey on Drug Abuse. *Journal of the Royal Statistical Society A,* 165(1), p.97-119.

Biemer, P. P., 2011. Latent Class Analysis of Survey Error. New Jersey: Wiley.

Forster M.R., 2000. Key Concepts in Model Selection: Performance and Generalizability. *Journal of Mathematical Psychology*, 44, 205-231.

Chanfreau, J., Cullinane,C., Calcutt, E. and McManus, S. , 2014. Wellbeing in Wales Secondary analysis of the National Survey for Wales 2012-13. [pdf]. Welsh Government Social Research. Available at: http://gov.wales/docs/caecd/research/2014/140430-national-survey-wellbeing-wales-2012-13-en.pdf [Accessed 25 January 2017].

Kreuter, F., Yan, T. and Tourangeau, R., 2008. Good item or bad—can latent class analysis tell?: the utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society A*. 171(3).

Lanza, S.T., Collins, L.M., Lemmon, D.R. and Schafer, J.L. 2007. PROC LCA: A SAS Procedure for Latent Class Analysis. *Structural Equation Modeling*, 14(4), p.671–694.

Lanza, S.T. and Rhoades, B.L., 2013. Latent Class Analysis: An Alternative Perspective on Subgroup Analysis in Prevention and Treatment. *Prevention Science*, 14(2), p.157-168.

Lin T.H. and Dayton C.M., 1997. Model Selection Information Criteria for Non-Nested Latent
Class Models. *Journal of Educational and Behavioral Statistics*, 22(3), 249-264.

Linzer, D. A. and Lewis J. B., 2013. "poLCA: Polytomous Variable Latent Class Analysis." R package version 1.4. http://dlinzer.github.com/poLCA.

Linzer, D. A. and Lewis J. B., 2011. "poLCA: an R Package for Polytomous Variable Latent Class Analysis." *Journal of Statistical Software*. 42(10): 1-29. http://www.jstatsoft.org/v42/i10

McCutcheon A.L. and Hagenaars, J.A.,1997. Simultaneous Latent Class Models for Comparative Social Research. In: Langeheine, R. and Rost, J. (eds) Applications of Latent Trait and Latent Class Models. New York: Waxmann. Pgs. 266-277.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Office for National Statistics., 2016. Personal well-being in the UK: Oct 2015 to Sept 2016 [online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/oct2015tosept2016 [Accessed 25 January 2017].

Office for National Statistics., 2016. Annual Population Survey [online]. Available through: UK Data Service Discover https://discover.ukdataservice.ac.uk/series/?sn=200002 [Accessed 25 January 2017].

Office for National Statistics., 2017. Surveys using the 4 Office for National Statistics personal well-being questions [online]. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/surveysusingthe4officefornationalstatisticspersonalwellbeingquestions [Accessed 25 January 2017].

R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. 3.0.2. [online]. Available at: http://www.R-project.org [Accessed 19 January 2016].

SAS Institute Inc. 2010. SAS/STAT™ 9.22 User's Guide. Cary, NC: SAS Institute Inc. [online] Available at: https://support.sas.com/documentation/cdl/en/statug/63347/PDF/default/statug.pdf [Accessed 19 January 2016].

StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.

The Methodology Centre. 2015. SAS Procedures for Latent Class Analysis & Latent Transition Analysis. [online]. Available at: https://methodology.psu.edu/downloads/proclcalta [Accessed 18 January 2016].

Vermunt, J. K., & Magidson, J., 2002. Latent class cluster analysis. In: J. A. Hagenaars, & A. L. McCutcheon eds. *Applied latent class analysis.* New York: Cambridge University Press, pp.89-106.

Vermunt, J. K. and Magidson, J., 2013. Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont Massachussetts: Statistical Innovations Inc. [online]. Available at: http://www.statisticalinnovations.com/latent-gold-5-1/ [Accessed 19 January 2016].

Appendix 1 – SAS® code for proc LCA

As described in Section 3, LCA can also be carried out in SAS. The Methodology Centre at Pennsylvania State University developed a SAS plugin which enables users to run LCA in SAS. It is important to note that the LCA procedure in SAS has not been written and is not supported by SAS Institute Inc.

The procedure used for running LCA in SAS is called proc LCA. In order to carry out an LCA in SAS, the code should be specified as follows:

```
proc LCA data = dataset_for_analysis;
         NCLASS number_of_latent_classes_to_be_estimated;
         ITEMS categorical_variables_to_include_in_analysis;
         Categories number_of_categories_in_each_variable;
         SEED specifies_seed_for_random_number_generator;
run;
```

The following is an example of how this code would be specified for the personal well-being data described in Section 3.2:

```
proc LCA data = PWB_LCA.wellbeing;
         NCLASS 5;
         ITEMS happyth2 anxiouth2 worthth2 satisth2;
         Categories 3 3 3 3;
         SEED 100000;
run;
```

The first line calls the 'wellbeing' dataset which is the dataset containing the personal well-being variables for analysis. Next, the number of classes to be estimated is specified. In this case we are running a latent class analysis which estimates 5 latent classes. The "ITEMS" option specifies the variables within the 'wellbeing' dataset that should be included in the LCA. Following this, the "Categories" option specifies the number of categories that each variable contains. In the case of the personal well-being dataset, each variable (happyth2, anxiouth2, worthth2 and satisth2) contains 3 categories (low, medium and high) as described in Section 3.2. Finally, a seed is specified, this is a starting value for the random number generator (the user can pick any starting value). There are a number of other options which can be included in the proc LCA e.g. MAXITER to specify the maximum number of iterations as was done with poLCA. For a full list of options and further information on carrying out LCA in SAS please refer to Lanza et. al (2007).

The proc LCA output includes the Likelihood Ratio G^2, degrees of freedom, AIC and BIC. The same methods for assessing model fit as those discussed in Section 3.4 apply here.

An investigation into the length of the Living Costs and Food Survey expenditure diary

The impact on data quality and estimates

Oscar Carrel ONS[1], Robynne Davies ONS[2]

1. Introduction

The Living Costs and Food Survey (LCF) is a household survey whose primary purpose is to collect information about expenditure on goods and services by UK households. The data collection of the LCF is split into two components; a face-to-face questionnaire followed by a self-completion expenditure diary. In 2016, the LCF underwent a National Statistics Quality Review (NSQR)[3]. This review recommended that we review the length of the two week expenditure diary. In particular, it posed the question:

"Could a shorter diary period produce the same level of accuracy whilst increasing response rates and data quality? Or would a shorter diary period reduce purchase frequency too much resulting in increased zero recording for some items and higher variability? "

This paper summarises the work we have carried out to date in response to the research question posed by the NSQR. In brief, the paper begins by studying evidence from other National Statistic Institutes (NSIs) and research organisations that have already considered this topic. It then moves on to present analysis we carried out to gain a fuller understanding of our current diary data, paying particular attention to patterns that may be an effect of data collection methods. It then presents our recreated main estimates and corresponding estimates of variability that use a subset of diary data. It concludes by summarising our findings to date and recommends areas where we believe further research would be beneficial.

1.1 Background of the Living Costs and Food Survey

1.1.1 History

A household expenditure survey has been conducted each year in the UK since 1957. From 1957 to March 2001, the Family Expenditure and National Food Surveys (FES and NFS) provided information on household expenditure patterns and food consumption for government and the wider community. In April 2001, these surveys were combined to form the Expenditure and Food Survey (EFS) which was later renamed the Living Costs and Food Survey (LCF) in 2008.

[1]Office for National Statistics; Oscar.Carrel@ons.gov.uk
[2]Office for National Statistics; Robynne.Davies@ons.gov.uk
[3]https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/nsqrseries2reportnumber3livingcostsandfoodsurvey

1.1.2 Uses

LCF data are widely used within and outside of government. The data are used to calculate the household final consumption expenditure component of Gross Domestic Product[4] and also provide weighting information for consumer price indices such as the CPIH and RPI. LCF data are used for a variety of other purposes, such as the estimation of calorie and nutrient intake by UK households[5] and the effect of taxes and benefits on income[6].

1.1.3 Fieldwork

The data collection of the LCF is split into two components; a face-to-face questionnaire followed by a self-completion expenditure diary. Information about regular expenditure, retrospective large and infrequent expenditures, income and demographics is collected within the questionnaire. The questionnaire is then followed by an expenditure diary, where all eligible adults and children[7] within a sampled household are asked to record their expenditure. Since the creation of the EFS in 2001, respondents have been asked to complete the expenditure diary over a two week continuous period that begins directly after the questionnaire has been completed. Within these two weeks, respondents are asked to record everything they spend within that time period.

Within the two week period, interviewers make at least one checking call on the household. At the end of the two week period, the interviewer completes a thorough check of the diary information, probing for further detail where necessary and adding this to the paper diary. The interviewer also arranges incentive payments. In 2015/16, a gift of £10 was made to each adult (aged 16 and over) who successfully completed the questionnaire and diary; children were given a gift of £5[8].

[4] https://www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/bulletins/consumertrends/quarter1jantomar2017
[5] https://www.gov.uk/government/collections/family-food-statistics
[6] https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/theeffectsoftaxesandbenefitsonhouseholdincome/previousReleases
[7] Children aged between 7 and 15 years of age are issued a simplified version of the diary
[8] From April 2016, the incentive for adults was increased to £20

2. Literature Review

We considered reports from other statistical and research institutes who have considered the length of the diary reporting period for a variety of consumer expenditure surveys. The findings from our investigations can be grouped into two main themes:

1. The effect diary length has on response
2. The effect diary length has on data quality and estimates

We present the evidence for each of these themes in turn.

2.1 The effect of the length of the diary reporting period on response

The NSQR identified an increase in response rates as one of main potential benefits for adopting a shorter diary length. Evidence from interviewer focus groups carried out as part of the NSQR suggested that the prospect of completing of a two week expenditure diary can lead to an outright refusal in some cases. The NSQR also highlighted the correlation between respondent burden and response, where a reduction in respondent burden (such as shorter diary reporting period) could have a positive impact on response rates.[9]

There is varying evidence regarding the relationship between the length of the diary reporting period and response. The French Household Budget Survey, conducted every 5 years, moved from a two week to a one week diary period in 2010/11 and reported a response rate of 69%, compared to 52% in 2005.[10] However, it is not certain whether this can be solely attributed to the decrease in diary reporting length as many other changes were made that year. For example, the number of questionnaires was reduced from three to two. In addition to that, the French Household Budget Survey is mandatory which makes comparison of response rates difficult.

The USA is currently planning large changes to their consumer expenditure survey[11] and commissioned the statistical survey research agency Westat to evaluate their current practices and to make recommendations for a new design. Westat recommend[12] moving to a one week diary period, and made reference to a potential increase in cooperation as a by-product of this. However, improving response is not the main reason for suggesting the reduction in the diary reporting period.

2.2 The effect of the length of the diary reporting period on data quality and estimates

The NSQR raised concerns that the length of the diary was having a negative impact on data quality. In particular, it highlighted the presence of diary fatigue in our data and the

[9] https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/nsqrseries2reportnumber3livingcostsandfoodsurvey
[10] https://www.insee.fr/en/metadonnees/source/s1333
[11] https://www.bls.gov/cex/ce_gemini_redesign.pdf
[12] https://www.bls.gov/cex/ce_gem_west_redesign.pdf

problems associated with this. Diary fatigue is a term that describes the decline in number of recorded purchases over the diary period. This is usually attributed to respondents tiring of the diary and subsequently completing it to a lower standard, or respondents forgetting to fill out their diaries. This can lead to a systematic under-reporting of data. In reducing the diary reporting period, the impact of diary fatigue could decrease, resulting in higher quality data. However, the NSQR did highlight that there was a potential for other factors to have a detrimental effect on data quality if the reporting period were to decrease, such as the reduction in purchase frequency and an increase in variability caused by a higher number of zero recorded items.

The positives and negatives that a shorter diary length has on data quality, as outlined in the NSQR, are echoed by other statistical and research institutes. For example, Westat's recommendation to move the USA Consumer Expenditure Survey's diary to a shorter reporting period is to tackle the effect of diary fatigue. Their paper states that a shorter reporting period may improve data quality if the diary reporting period is shortened[13]. However, there is also evidence to suggest that a shorter diary period can have a detrimental effect on diary fatigue. In a report[14] looking at poverty expenditure in Brazil using data from Brazil's Consumer Expenditure Survey, it was highlighted that the one week diary period was detrimental to the data due to the relatively large number of households reporting zero expenditure. In addition to this, during the period when the design for the EFS (now LCF) was being considered, a two week diary reporting period was decided upon with the aim of minimising the number of zero expenditure households.[15]

A research paper by Crossley and Winter[16] summarises the trade off a shorter diary reporting period presents for data quality; a shorter period may decrease the bias of estimates, but increase the variability.

2.3 Conclusions

Our literature review suggests that it is possible that a decrease in the diary reporting period will benefit response rates. However, due to the individual characteristics of expenditure surveys and the conditions under which they are conducted, it is difficult to predict how response may be affected for the LCF specifically. In order to gain a greater and more robust understanding of the effects it may have, a split sample trial would be needed.

There is evidence to suggest that diary fatigue has a detrimental effect on data quality, therefore a shorter diary reporting period may improve the data by decreasing diary fatigue and subsequent under-reporting. However, we have found supporting evidence to support the claim that a shorter reporting period presents different threats to data quality; namely, an increase in variability and an increase in households reporting zero expenditure for various goods and services.

[13] https://www.bls.gov/cex/ce_gem_west_redesign.pdf
[14] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.569.4724&rep=rep1&type=pdf
[15] EFS Pilot report, ONS, available on request
[16] https://core.ac.uk/download/pdf/6628272.pdf

In order to better understand the effect on data quality, the next sections of this paper explore the extent to which LCF data is affected by diary fatigue, and then considers how main estimates, variability and the levels of zero expenditure may be affected if we were to reduce the diary reporting period from two weeks to one.

3. Patterns and characteristic of current LCF diary data

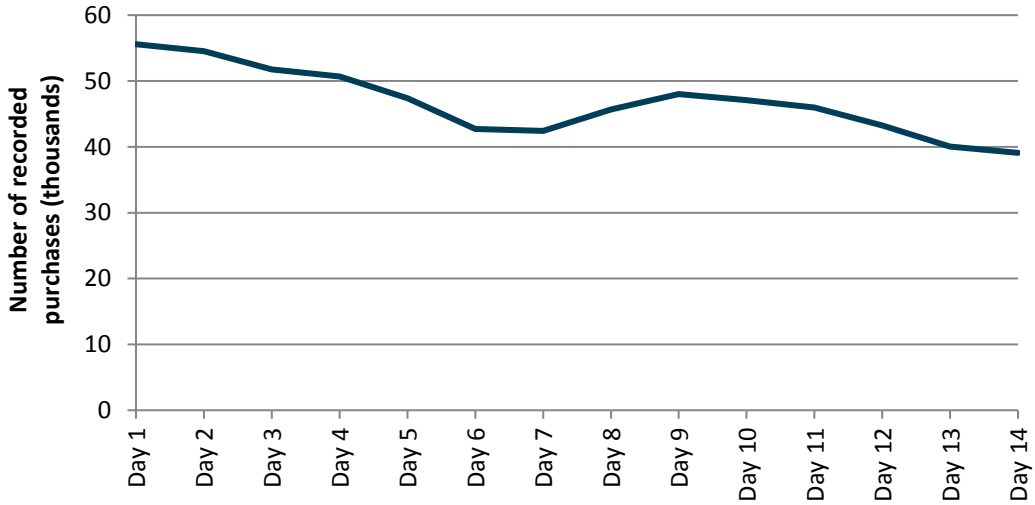
To best identify the consequences of changing the length of our diary recording period it is important to understand our current data. We therefore conducted an investigation into our 2015/16 financial year dataset in order to better understand current spending patterns and in particular investigated the prominence of diary fatigue in the data.

3.1 Diary fatigue

As mentioned previously, diary fatigue is the decline in the number of recorded purchases over the diary period. We have found evidence of diary fatigue in our current 2015/16 data: comparing the total number of items purchased in the first week of the current period compared to the second, we see a decrease of 10.4% from week 1 to week 2. It is this apparent 'loss' of data in the second week that may indicate a loss of data quality. In shortening the length of the diary, this problem may be removed to a certain extent.

Figure 1 shows that there is a steady decline in recorded purchases between days 1 to 7, and although the number of recordings increases at the beginning of the second week, it decreases again to the lowest number of recorded purchases on day 14.

Figure 1 - Total recorded purchases by day of the diary period. 2015/16 data



It is unlikely that diary fatigue affects all types of purchases in the same way. Firstly, not all purchases are captured in the diary and would therefore not be affected. Secondly, due to respondent behaviour, it is likely that some items are more susceptible to diary fatigue. In order to understand this better, we have examined diary fatigue for different subdivisions of our data.

3.1.1 Diary fatigue by COICOP category

We publish estimates of average weekly household expenditure based on the United Nations' Classification of Individual Consumption by Purpose (COICOP)[17]. Table 1 presents the number of items purchased in each week, broken down by COICOP category. The level of fatigue is presented as the percentage decrease in recorded purchases between the first and second week of the diary. To aid interpretation, we have included how much each estimate relies on diary data; this is presented as the proportion of each expenditure estimate made up by purchases recorded in the diary.

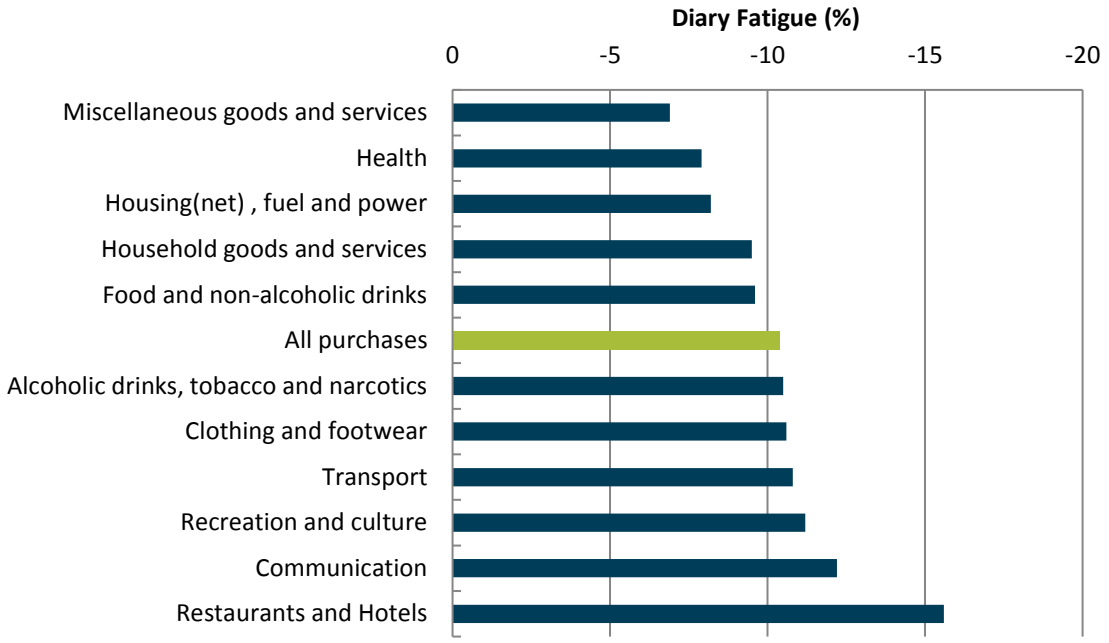
Table 1 - Diary fatigue and the contribution diary data makes to overall estimate, by COICOP category, 2015/16 data

COICOP category		Number of recorded purchases Week 1	Week 2	Percentage decrease i.e. fatigue (%)		Diary expenditure (%)
1	Food and non-alcoholic drinks	201,052	181,744	-9.6		99.7
2	Alcoholic drinks, tobacco and narcotics	8,734	7,815	-10.5		99.5
3	Clothing and footwear	8,132	7,269	-10.6		98.5
4	Housing(net), fuel and power	998	916	-8.2		11.5
5	Household goods and services	17,231	15,600	-9.5		57.2
6	Health	3,194	2,943	-7.9		92.5
7	Transport	8,619	7,685	-10.8		59.2
8	Communication	1,002	880	-12.2		13.2
9	Recreation and culture	29,125	25,861	-11.2		57.3
10	Education	87	68	-21.8		8.1
11	Restaurants and Hotels	48,967	41,345	-15.6		92.1
12	Miscellaneous goods and services	14,752	13,730	-6.9		52.5
ALL ITEMS		**344940**	**309034**	**-10.4**		**56.1**

Figure 2 ranks the COICOP categories by levels of diary fatigue from smallest to largest.

[17] https://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=5&Lg=1

Figure 2 – Diary Fatigue by COICOP category, 2015/16 data



Both Table 1 and Figure 2 show that when broken down by COICOP category, the level of diary fatigue in our data varies. The food and non-alcoholic drinks category makes up over half of all recorded purchases and we can see that this has lower than average diary fatigue. This suggests that a change in diary recording period may not affect data quality of this category that much. In contrast to this, both Restaurants and Hotels and Recreation and Culture show some of the greatest levels of fatigue, at 15.6% and 11.2% respectively. These categories are also heavily reliant on diary data. At even more detailed breakdowns of expenditure there is further variation. Even within COICOP categories, subcategories of expenditure experience different levels of fatigue. Tables 10 and 11 in the annex contain a selection of subcategories that perform notably better or worse than average.

The NSQR carried out analysis into the levels of diary fatigue in older LCF data and identified specific items affected. Although we see slightly smaller levels of fatigue than those reported in the NSQR, we see the same variables being affected more than others such as eating out and takeaways.

The reasons behind the differing levels of fatigue between categories of expenditure are unclear and are outside of the scope of this research; however, it may be beneficial to undertake further research into their causes. In particular, a better understanding of the causes of fatigue may present solutions to reduce diary fatigue that do not require changes to the length of the LCF diary. Some examples of alternative solutions may include: prompts, added to the diary to remind respondents about certain purchases that are shown to be regularly forgotten; the simplification of sections of the diary where the recording task is shown to have too great a cognitive burden on respondents or perhaps

more frequent interviewer checking calls if it is shown that regular visits would prevent respondents from forgetting the diary entirely as they grow tired of it.

3.1.2 Diary fatigue by price of items purchased

In addition to COICOP category, we investigated whether there were any patterns in diary fatigue when broken down by price category as shown in Table 2.

Table 2 - Diary fatigue by price bracket, 2015/16

Price bracket of recorded purchases	Number of recorded purchases		Percentage decrease i.e. fatigue (%)
	Week 1	Week 2	
Less than or equal to £0.25	8,218	7,283	-11.4
£0.26 - £0.50	29,825	26,872	-9.9
£0.51 - £1.00	105,809	94,251	-10.9
£1.01 - £1.50	45,526	40,988	-10.0
£1.51 - £2.00	44,487	40,128	-9.8
£2.01 - £5.00	66,254	59,102	-10.8
£5.01 - £10.00	22,509	19,810	-12.0
£10.01 - £20.00	11,384	10,301	-9.5
£20.01 - £50.00	7,515	6,605	-12.1
Greater than £50.01	2,857	2,563	-10.3

Table 2 shows that although there is a small amount of variation in the levels of diary fatigue between price categories, there is no obvious relationship between price and fatigue.

3.1.3 Diary fatigue during the first half of the two week diary

It should also be noted that moving to a one week diary recording period may not eliminate the problem of diary fatigue within the data. Table 3 demonstrates that fatigue exists within the first 8 days of our current recording period:

Table 3 - diary fatigue within first week of recording period

	COICOP category	No. items DAYS 1-4	No. items DAYS 5-8	% decrease
1	Food and non-alcoholic drinks	124,468	102,939	-17.3
2	Alcoholic drinks, tobacco and narcotics	5,431	4,499	-17.16
3	Clothing and footwear	4,724	4,557	-3.54
4	Housing(net)1, fuel and power	590	569	-3.56
5	Household goods and services	10,603	8,968	-15.42
6	Health	1,893	1,753	-7.4
7	Transport	5,208	4,575	-12.15
8	Communication	581	591	1.72
9	Recreation and culture	17,828	15,275	-14.32
10	Education	51	43	-15.69
11	Restaurants and Hotels	30,369	24,798	-18.34
12	Miscellaneous goods and services	8,981	7,780	-13.37

3.2 Other factors that may influence data quality

The NSQR and our literature review identified other factors that can have a negative impact on data quality; a reduction in the number of houses recording certain purchases and an increase in variability. Unlike diary fatigue, these factors are likely to worsen if we were to reduce the length of the diary reporting period. In order to understand this better, we analysed the prominence of these factors in our current dataset.

3.2.1 Purchase infrequency and zero expenditure households

Purchase infrequency refers to the irregularity at which a particular item is purchased. It is an important factor in identifying the ideal length of the LCF diary's recording period. If we intend to record the expenditure of an item that is purchased irregularly (or has high purchase infrequency) a longer diary gives us a greater chance of recording those purchases and will allow us to create more precise estimates. In contrast if the diary is too short the proportion of diaries recording that expenditure is reduced and this will lead to greater variation in estimates. As an indication of the effect of purchase infrequency on our estimates, we examined the number of recording households when we use the first week of the diary in isolation and compared this with a full two week dataset. We then broke this down for different expenditure categories.

In moving from a two week to a one week diary period, a further 0.4% of responding households became nil-expenditure households; that is, they recorded no purchases at all during the diary period. However, at more detailed breakdowns of expenditure the number of new nil-expenditure households is often greater. Table 3 compares the proportion of responding households that did not record any expenditure for the different COICOP categories over the different recording periods. Please note that this looks at

diary data only, which is why some categories may have unusual levels of nil-expenditure cases, such as housing and communication.

Table 3 – Percentage of nil-expenditure cases in diary data for different diary periods, by COICOP, 2015/16 diary data

COICOP category		Proportion of households with nil-expenditure (%)		Difference in proportion of nil-expenditure
		Original	1st Week	
1	Food and non-alcoholic drinks	0.5	1.8	1.3
2	Alcoholic drinks, tobacco and narcotics	38.9	51.4	12.5
3	Clothing and footwear	34.9	50.9	15.9
4	Housing(net), fuel and power	81.9	88.8	6.9
5	Household goods and services	9.1	20.3	11.2
6	Health	48.8	65.9	17.1
7	Transport	20.6	32.5	11.9
8	Communication	76.3	85.1	8.8
9	Recreation and culture	8.2	14.0	5.7
10	Education	97.7	98.5	0.9
11	Restaurants and Hotels	13.3	19.4	6.1
12	Miscellaneous goods and services	12.1	26.4	14.3

As expected, the proportion of nil-expenditure cases for all expenditure categories increases as we shorten the length of the diary. However, the extent of that increase varies. The effects for different COICOP categories, and how they compare and contrast, are more easily shown in Figure 3.

Figure 3 – Nil-expenditure cases in diary data as a proportion of achieved sample, by COICOP, 2015/16 data

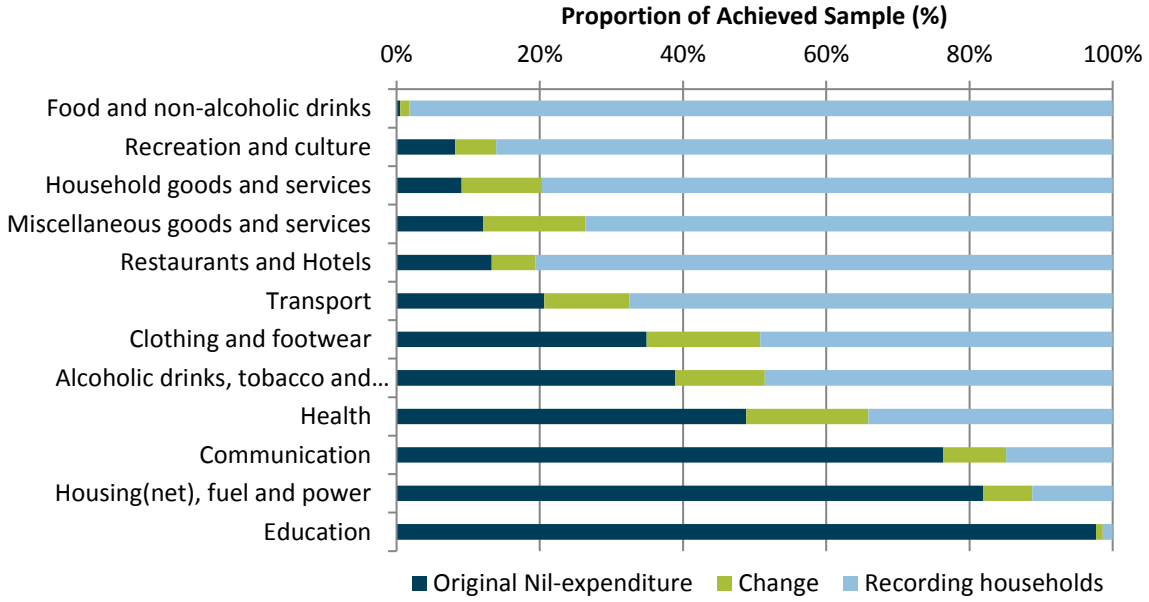


Figure 3 shows us that for some COICOP categories such as food and non-alcoholic drink, the proportion of nil-expenditure cases (despite being three times greater in the shorter diary) remains low at 1.8%. This reflects that most households purchase an item of food and drink every week. It is the categories that are purchased less regularly that see the greatest change. Health and Clothing and Footwear are categories in which purchases are less regular and they see the greatest increase in the proportion of nil-expenditure cases at 17.1% and 15.9% respectively.

The increase in the number of nil expenditure households is also reflected in the more detailed expenditure categories. We considered the most detailed levels of expenditure that are published in Table A1 as part of the annual publication of the LCF, Family Spending[18]. We found that the number of recording households for some lower level item drops considerably. As an indication of the decrease, 42% of 3rd and 4th level categories in Table A1 had a decrease in recording HH's of 30% or more, whilst 18% of categories decrease by 40% or more.

It is worth noting the number of recording purchases only serves as a proxy for the reliability of an expenditure estimate. Later sections of this paper will consider whether the reduction in purchases has increased the variability of estimates.

3.2.2 The impact the diary completion task has on respondent behaviour

The risk of changing respondent behaviour was a significant factor in the decision to retain a two week diary period when the survey's data collection was reviewed by the EFS pilot in 2000. The concern was that changing the diary length could influence

[18] https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/expenditure/datasets/componentsofhouseholdexpenditureuktablea1

shopping habits and, in particular, it was feared that a one week diary period would make it easier for a respondent to 'put-off' a shopping trip to avoid the burdensome diary task.[19]

During the course of our analysis of the LCF diary data, we encountered a potential way in which respondents' behaviour changes as a result of completing the diary task. Below we present evidence that suggests households are making an unnaturally high number of purchases towards the beginning of diary. Whilst we do not know how this behaviour might change if we were to introduce a shorter diary, it is important to understand how the current task may introduce unnatural shopping habits for respondents.

From Figure 1 (presented earlier) we saw that the rate of diary recording is greatest at the beginning of the diary recording period. This shows that whatever the day of the week, the greatest number of purchases is made at the beginning of the diary recording period. However, this is biased towards which day of the week the diary starts on. Table 4 demonstrates Tuesdays and Wednesdays are the most popular days of the week for the diary to be given to households to complete.

Table 4 - Percentage of diaries started by day of the week, 2015/16 data

	Percentage of diaries started (%)
Monday	15.0
Tuesday	22.2
Wednesday	20.8
Thursday	18.2
Friday	15.0
Saturday	6.9
Sunday	1.9

Therefore, the majority of weekends will appear in our data around days 5 and 6 and 12 and 13 when shopping behaviours may be different. To discount the bias introduced by which day of the week that the diary started on, we weighted the data to simulate an even diary placement for the seven days of the week. We can then understand the unbiased average number of purchases recorded by each household each day. Figure 4 presents this average number of recordings for each day of the diary period as though an even placement of diaries was achieved.

[19] EFS Pilot report, ONS, available on request

Figure 4 – Average Purchases per Household each day of the diary period, with even diary placement. 2015/16 data

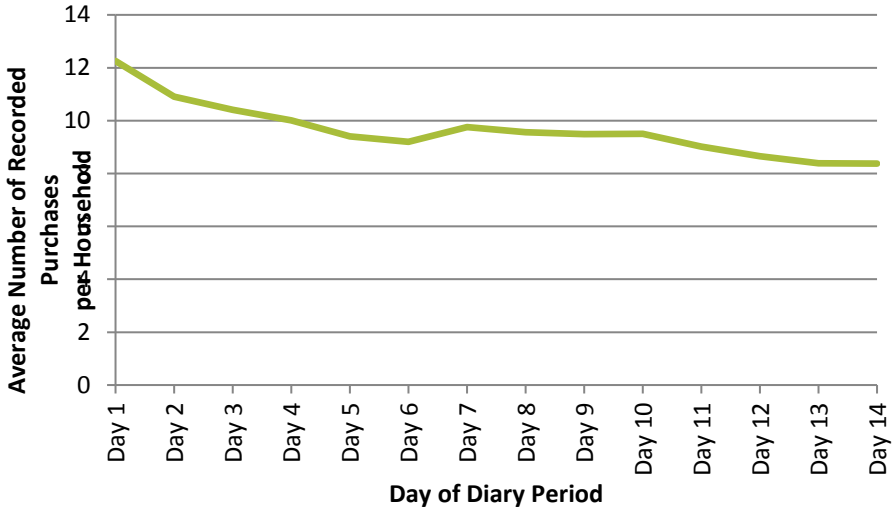
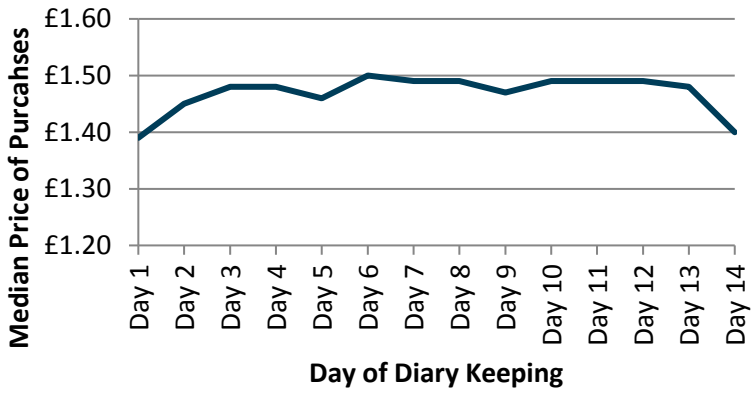


Figure 4 shows that adjusting for the uneven diary placement, it is clear that the decline in recording is sharpest towards the beginning of the two weeks. Although we do not dispute that there is evidence of diary fatigue in our data, we further explored whether this is solely responsible for the sharp decline in number of items purchased over the first couple of days. Alternative explanations could be that the act of keeping a diary influences a respondent to shop earlier than they might have otherwise or that the interviewers encourage an early shopping trip to ensure that respondents are completing the diary correctly. If this is the case, then the diary fatigue that we have explored in earlier sections may be exaggerated by this respondent behaviour.

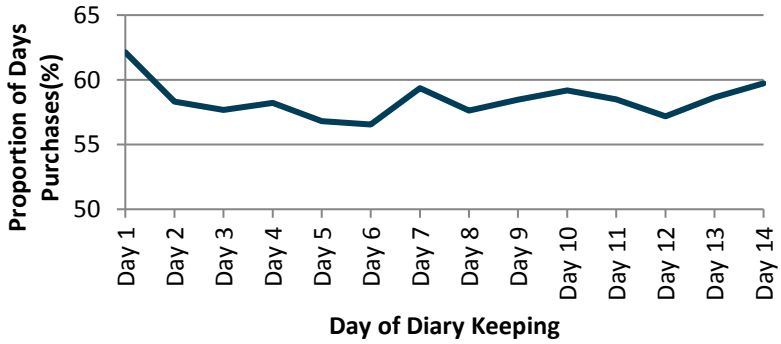
Some evidence to suggest the large proportion of purchases made on day 1 is due to an unexpectedly high number of 'big shops' can be found in the type of purchases themselves. We see from Figure 5 that the median price per item is lower for day 1 than the following days. The median price paid of items each day gives us an understanding of the type of shopping that is occurring each day. A low median price indicates respondents are purchasing a greater proportion of low price items (similar perhaps to a typical supermarket shop) and we see that the median expenditure on day 1 is the lowest of any recording day at £1.39.

Figure 5 - Median price of purchases for each day of the recording period



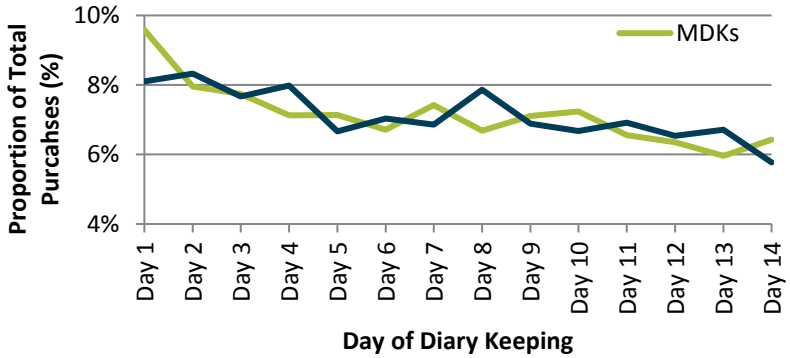
Furthermore, the types of purchases can provide more evidence. As demonstrated by Figure 6, purchases of food and non-alcoholic drinks make up a greater proportion of the purchases on day 1 than any other day. This could suggest that households are food shopping more on day 1, which we would not expect to be the case if we were recording natural shopping habits. Table 13 in the annex contains more data on this.

Figure 6 – 'Food and non-alcoholic drink' as a proportion of total purchases, by day of reporting period, 2015/16 data



Similar evidence can be seen in the behaviour of Main-Diary-Keepers (MDKs) on day 1 of the diary. The MDK is the member of the household identified as the person responsible for the majority of the household shopping. Figure 7 demonstrates that MDKs make a larger proportion of their total purchases on day 1 than any other day, in addition to making a larger proportion of purchases than non-MDKS on day 1.

Figure 7 – Proportion of Total MDK expenditure made each day of reporting period, 2015/16 data



Whilst the above evidence may suggest that respondents are changing their shopping patterns for the data collection, provided that respondents are still buying the same items over the course of the two weeks there is no effect on our estimates of expenditure. However, it could be of concern if it was determined that for some respondents this 'big shop' on day 1 is causing respondents to purchase different items throughout the diary as this will bias our estimates. It may be worth considering whether or not this is more likely to be the case in a shorter diary. It also acts a note of caution when we examine the effect of diary fatigue in the 2015/16 data, which may be exaggerated due to the high number of purchases made at the beginning of the diary recording period.

This is just one example of potential unusual respondent behaviour and is based upon an investigation on 2015/16 data collected with a two week diary period. We cannot identify how a respondent would have behaved if they were given a one week diary from our current data. A split-sample trial would be required to fully understand how respondents would behave under different conditions.

3.3 Conclusion

Our investigations into 2015/16 data have identified a number of things. Firstly, there is evidence to suggest that diary fatigue is present in our data. The diary fatigue varies from item to item and it is not immediately clear why one item is more likely to suffer from diary fatigue than other. We have also found evidence to suggest that asking households to record purchases for one week only would increase the number of zero expenditure households, which may pose problems for data quality. This will be explored further in the next section of this paper. Finally, we have found evidence to suggest that the task of completing a diary may influence the way households buy and record purchases; this suggests that changing the length of the diary recording period may influence respondents' behaviour causing a change in our data.

4. Headline expenditure estimates using the one week of diary data

4.1 Headline expenditure estimates using the first week of diary data

We have created new estimates of average household expenditure for 2015/16 excluding the purchases recorded in the second week of diary keeping. This effectively simulates data collection with a one week diary period. By comparing these new estimates with our original 2015/16 estimates we can attempt to quantify the changes to data quality that a change in diary period would cause.

It is worth noting that the new estimates are based upon a subset of data collected with a two week diary as opposed to a true one week diary and therefore won't reflect any changes in respondent behaviour that may occur, such as an increase in response rates.

Please note that the new estimates in the below tables have been recreated for comparison purposes only to inform this paper, and do not replace the original 15/16 estimates published in February 2017.

4.1.1 Changes to the magnitude of expenditure estimates

We first consider the changes to the magnitude of our estimates of average weekly household expenditure based on the first week of diary data. Please note that for our headline results, the average is the mean. We have investigated the change in estimates at all levels of aggregation and present our results below.

Change in top-Level COICOP expenditure estimates

Table 5 compares the original two week diary period expenditure estimates with the new one week diary period expenditure estimates.

Table 5 - Expenditure estimates based on 1st Week, by COICOP, 2015/16 data

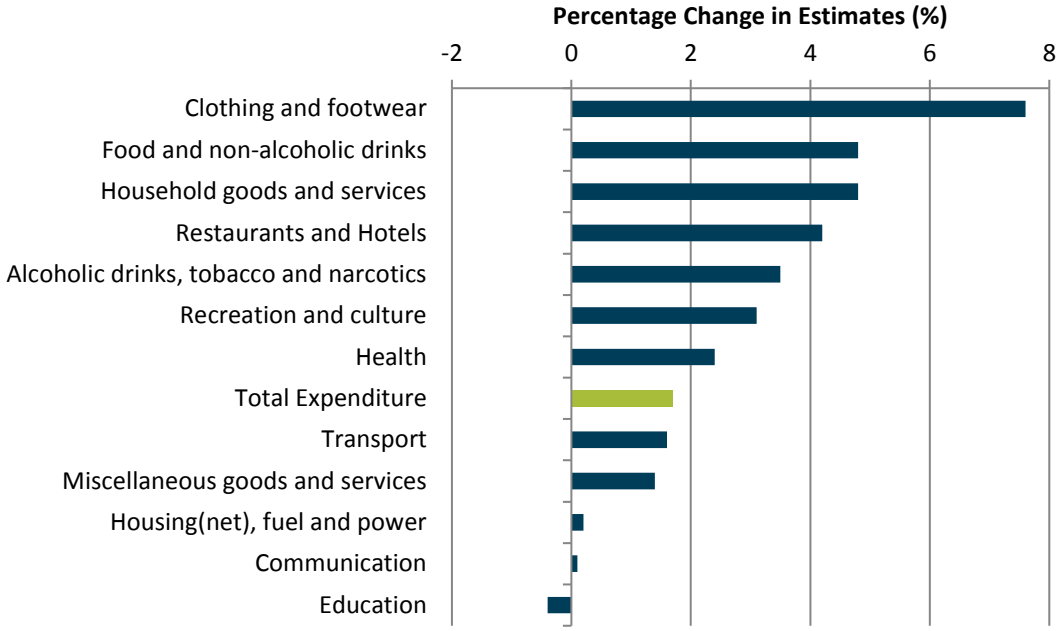
COICOP category		Expenditure Estimates (£) Original	1st Week	Percentage change in estimates (%)	95% confidence interval of Original Estimate (£) Lower	Upper
1	Food and non-alcoholic drinks	56.8	59.6**	4.8	55.8	57.9
2	Alcoholic drinks, tobacco and narcotics	11.4	11.8*	3.5	10.8	11.9
3	Clothing and footwear	23.5	25.2**	7.6	22.1	24.8
4	Housing(net), fuel and power	72.5	72.7	0.2	70.0	74.9
5	Household goods and services	35.5	37.2**	4.8	32.6	38.3
6	Health	7.2	7.4	2.4	6.0	8.4
7	Transport	72.8	73.9	1.6	69.8	75.7
8	Communication	16.0	16.0	0.1	15.6	16.4
9	Recreation and culture	68.0	70.1**	3.1	65.0	70.9
10	Education	7.0	7.0	-0.4	5.1	9.0
11	Restaurants and Hotels	45.1	47.0**	4.2	43.1	47.1
12	Miscellaneous goods and services	39.7	40.2	1.4	37.8	41.5
Total Expenditure		528.9	537.8	1.7	515.7	542.1

**significantly different to original estimate, at the 1% level, * significantly different to original estimate, at the 5% level

We see that with the shorter diary period, our estimate of total average weekly expenditure increased by 1.7% .The top-level expenditure estimates generally increase when they exclude the second week of diary data. This is unsurprising as by discounting the second week of data we have reduced diary fatigue and subsequent under-reporting in our data. Education expenditure, a category that is captured almost entirely by the face-to-face questionnaire element of the survey, was the only category to decrease; this change was not found to be significant.

Figure 8 shows the COICOP categories with the largest percentage changes to estimates.

Figure 8 – Percentage Change in Estimates of Average Weekly Expenditure by Top-level COICOP Category, 2015/16 data



Estimates for categories reliant on diary data had the largest changes, and were found to be significantly different to the original estimates. Clothing and footwear had the greatest increase of any category; this was influenced to a certain extent by a potential outlier in the first week of diary data. However, even when the outlier is treated[20], the category still has the greatest percentage increase at 5.3%.

[20] The outlier was treated by reducing the case's weight to one

Change in median expenditure estimates

An alternative estimate to consider is the median weekly expenditure of households. We considered the changes to our estimates of median weekly expenditure if we were to exclude the second week of diary data. The results are shown in Table 6.

Table 6 - Median expenditure for different diary periods, by COICOP, 2015/16 data

COICOP category		Median Expenditure by COICOP category (£) Original	1st Week	Difference (%)
1	Food and non-alcoholic drinks	49.8	50.7	1.8
2	**Alcoholic drinks, tobacco and narcotics**	**3.6**	**0.0**	-
3	**Clothing and footwear**	**8.0**	**0.0**	-
4	Housing(net), fuel and power	43.8	43.2	-1.4
5	Household goods and services	10.4	9.5	-8.9
6	**Health**	**0.2**	**0.0**	-
7	Transport	40.1	40.1	-0.1
8	Communication	13.7	13.6	-0.7
9	Recreation and culture	32.2	30.5	-5.3
10	Education	0.0	0.0	0.0
11	Restaurants and Hotels	27.3	26.1	-4.4
12	Miscellaneous goods and services	23.7	22.1	-6.5
Total Expenditure		431.8	433.4	0.4

Table 6 shows that unlike the mean estimates, the median estimates tend to decrease when we exclude the second week of diary data. There is an increased disparity between the mean and median estimates of expenditure that indicates that the distribution of expenditure is more skewed when only one week's diary data is used. Furthermore, as a result of a decrease in recorded purchases, the median expenditure of 3 categories (bold in the table) drops to zero. This means at least 50% of households spent no money on these categories when only the first week of data is used, reflecting the increase in the number of nil-expenditure households that we expected to see as a result of a shorter diary period.

Both of these factors point towards the problems posed by purchase infrequency and the greater dispersion of expenditure reported by individual households. As mentioned earlier this doesn't necessarily indicate a decrease in the quality of our estimates but should be considered alongside other indicators of data quality.

Change in lower-level COICOP expenditure estimates

The expenditure estimates at a lower level follow similar trends to the top level estimates although they are inevitably more variable. Generally, estimates tend to increase marginally (approximately 3%). Table 13 (available on request) contains details on the change in estimates for all detailed expenditure categories.

4.1.2 Changes to precision of expenditure estimates

In addition to the change in the magnitude of the estimates, we can consider the change in precision where we use the Coefficients of Variation (CVs) as a measure of precision. CVs present the standard error of an estimate as a proportion of the estimate itself and therefore give us an indication of the extent our estimates will vary about their true population values due solely to chance. Furthermore, CVs are comparable across different estimates allowing us to easily assess the difference in precision between our new one week estimates and our original two week estimates.

As with the estimates themselves we will consider the precision of our expenditure estimates at different levels of aggregation.

Change in precision of top-level estimates

The estimate of average total household expenditure does not become much less precise when we exclude the second week of the diary data. We can conclude from this that the data collected by the face-to-face questionnaire and first week of diary keeping is sufficient to retain a similar level of precision at the greatest level of aggregation.

However, broken down by top-level COICOP categories we do see the decrease in precision, indicated by the increase in CVs. As shown in Table 7, the CVs of our top-level COICOP estimates tend to increase. In particular, we see that those categories that are highly reliant on diary data (bold in table 7) such as alcohol drinks, tobacco and narcotics all see increases to their CVs.

Table 7 – Change in Coefficient of Variation based on 1st Week data, by COICOP, 2015/16 data

COICOP category		Coefficients of variation (%) Original	1st Week	Difference between CVs (%)
1	**Food and non-alcoholic drinks**	**1.0**	**1.2**	**0.2**
2	**Alcoholic drinks, tobacco and narcotics**	**2.2**	**2.6**	**0.4**
3	**Clothing and footwear**	**2.9**	**3.9**	**1.0**
4	Housing(net), fuel and power	1.7	1.8	0.1
5	**Household goods and services**	**4.1**	**4.5**	**0.4**
6	Health	8.5	7.1	-1.5
7	Transport	2.1	2.3	0.2
8	Communication	1.2	1.2	0.0
9	**Recreation and culture**	**2.2**	**2.5**	**0.3**
10	Education	14.1	14.0	-0.1
11	**Restaurants and Hotels**	**2.3**	**2.5**	**0.2**
12	Miscellaneous goods and services	2.3	2.4	0.1
Total Expenditure		1.3	1.3	0.0

We investigated the increase in sample size required in order to regain the current levels of precision for each of the main COICOP categories. Using an approximate method by using the rule that the sample error will reduce with the square of the sample size, we estimated that the sample size would need to dramatically increase in order for precision to be regained. This is described in Table 12 of the annex, where our 2015/16 sample size was 4916 households.

Change in precision of lower level estimates

The CVs of lower level estimates tend to increase and quite often to a greater extent than for the top-level estimates. More than 30% of lower level estimates that relay more detailed expenditure have their CV increase by over a third of the original. This loss of precision is of concern because important users of LCF data, such as the Department for Food, Energy and Rural Affairs and National Accounts, rely on the quality of the lower level estimates. In addition to this, the NSQR states that "the level of precision should not be allowed to decline further".

Table 13 (available on request) contains details on the change in CVs for all detailed expenditure categories. Some of the categories with subcategories suffering from high CVs based on the one week diary data from 2015/16 include:

- 5.3 - Household Appliances
- 7.1.3 - Purchase of Motorcycles
- 7.2.1 - Spares and Accessories (Within operation of personal transport)
- 9.1 - Audio-visual, photographic and information processing equipment
- 9.2 - Other major durables for recreation and culture

4.2 Headline expenditure estimates using the second week of diary data

The above estimates and CVs were calculated using the first week of the diary data, all the questionnaire data and discounted the second week of diary data. We also created estimates and CVs where we used the second week of the diary data, all the questionnaire data and discounted the first week of diary data. These are presented in Table 8.

Table 8 - Expenditure estimates based on the 2nd week only, 2015/16 data

COICOP Category		Average Expenditure (£) Original	Average Expenditure (£) 2nd Week	Percentage change in estimates (%)
1	Food and non-alcoholic drinks	56.8	53.7	-5.5
2	Alcoholic drinks, tobacco and narcotics	11.4	10.8	-5.3
3	Clothing and footwear	23.5	21.5	-8.5
4	Housing(net), fuel and power	72.5	72.3	-0.3
5	Household goods and services	35.5	33.7	-5.0
6	Health	7.2	6.9	-3.6
7	Transport	72.7	71.1	-2.3
8	Communication	16.0	15.9	-0.4
9	Recreation and culture	68.0	65.4	-3.8
10	Education	7.0	7.1	0.4
11	Restaurants and Hotels	45.1	42.8	-5.2
12	Miscellaneous goods and services	39.7	39.0	-1.8
Total Expenditure		528.9	517.8	-2.1

Table 8 shows that the estimates for main COICOP estimates decrease, which is to be expected as we are using the week of diary data where less items were recorded.

We found that the CVs of the main COICOP estimates increased when compared to the original estimates. However, the increase in CVs was smaller than when we used the first week of the diary data. Table 9 shows that that for the majority of COICOP categories, estimates that exclude the first week of the diary data are more precise than those that exclude the second week of the diary data.

Table 9 - Comparison of the increase in Coefficient of Variation for different diary periods, 2015/16 data

COICOP Category		Coefficients of variation (%)			Difference, original and 1st week	Difference, original and 2nd week	Ratio of differences
		Original	1st Week	2nd Week			
1	Food and non-alcoholic drinks	1.0	1.2	1.1	0.2	0.2	0.7
2	Alcoholic drinks, tobacco and narcotics	2.3	2.6	2.5	0.4	0.3	0.8
3	Clothing and footwear	2.9	3.9	3.4	1.0	0.6	0.6
4	Housing(net), fuel and power	1.7	1.8	1.7	0.1	-0.0	0.3
5	Household goods and services	4.1	4.5	4.3	0.4	0.2	0.4
6	Health	8.5	7.1	14.8	-1.5	6.3	4.2
7	Transport	2.1	2.3	2.2	0.2	0.1	0.6
8	Communication	1.2	1.2	1.6	0.0	0.4	29.7
9	Recreation and culture	2.2	2.5	2.6	0.3	0.4	1.4
10	Education	14.1	14.0	14.3	-0.1	0.2	1.5
11	Restaurants and Hotels	2.3	2.5	2.5	0.2	0.2	0.8
12	Miscellaneous goods and services	2.3	2.4	3.3	0.1	0.9	17.2
Total Expenditure		1.3	1.3	1.3	0.0	0.1	2.2

This suggests that expenditure itself is less variable in the second week. This poses the question: If we were to move to a one week diary period, would households' expenditures be as variable as we see in week one or as variable as in week two? It also questions whether the variability we see in the first week of the data may be due to respondent behaviour that we discussed earlier on in this report.

Nevertheless, as the CV increases regardless of the one week period that we consider, we can conclude that a one week diary would result in higher variability.

4.3 Outliers

When considering a shorter diary period, another interesting feature that should be considered is the creation of new outliers. In the case of using just the first week of data, one household that was previously not considered to be an outlier had its clothing expenditure increase drastically to the point it was having an unwanted effect on our

clothing estimate and CV. Similarly, in moving to just the second week of data, one household became a potential outlier in the health expenditure category as a handful of its high expenditure purchases had a greater influence on the estimates.

These occurrences serve as an indication that, if we were to shorten our diary reporting period, outliers will occur more frequently.

4.4 Conclusion

In recreating estimates and CVs, we can conclude that if we were to move to a shorter diary period, our estimates would change. In particular for items reliant on diary data, changes will be significant. Care would therefore have to be taken when comparing estimates over time, as the change in diary recording period may cause discontinuities in time series.

We can also conclude that reducing the diary recording period, all other things remaining equal, would cause a loss in precision; particularly for lower level estimates. This is of particular concern given the steer from the NSQR that precision should not decrease further.

5. Overall conclusion and areas for further research

5.1 Conclusion

Our investigation has identified some key findings to help us answer the NSQR question:

"Could a shorter diary period produce the same level of accuracy whilst increasing response rates and data quality? Or would a shorter diary period reduce purchase frequency too much resulting in increased zero recording for some items and higher variability? "

Assuming diary fatigue does create a bias towards estimates lower than the true estimate, a shorter diary period may improve our level of accuracy.

However, based on an equal achieved sample, the improvements to accuracy would probably be outweighed by the decrease in precision: we saw considerable increases in variability over lower-level estimates when one week of diary data was used and the NSQR is clear that precision cannot be allowed to decline any further than current levels.

If response did improve due to a shorter diary period this would recover some of the lost precision but this investigation could not estimate the changes in response. Reports from other National Statistic Institutions give a mixed impression of the effect a shorter diary period has on response, presenting little evidence that it has a positive effect. A split sample trial would be needed to assess the specific impact on LCF response rates.

Furthermore, it's unlikely that the total number of purchases can be recovered by an improved response rate alone. Therefore, some of the categories that already have a low number of purchases may fail to be captured sufficiently by the shortened diary. It may be worth considering other methods to record expenditure for certain items if a shorter diary period were to be introduced.

5.2 Areas for further research

Our literature review identified that various other statistical institutes are considering or have moved to a shorter diary period. We should continue to monitor the results of those institutes and liaise where appropriate to learn from one another.

Diary fatigue features heavily in literature as the reason for reducing the diary reporting period. We could not identify any obvious patterns to determine which categories of expenditure suffered from high fatigue. It would be beneficial to investigate this further; if we could identify the cause, perhaps we could find an alternative method for reducing fatigue such as using prompts in the paper diary.

The regularity with which people go shopping should be important to deciding how long our diary should be as we do not want to give households the opportunity of delaying a big shop. There have been additional questions added to the LCF questionnaire for 2017 to better understand UK household shopping patterns.

It would be beneficial to further understand the impact the diary and interviewer has on respondents' behaviour, so that we better understand the "day one" effect we currently see and to unmask any other effects the data collection has on data recording.

Finally, our investigations were conducted on a subset of 2015/16 data. Any research that can be carried out to assess changes to respondent's behaviour when they're completing a shorter diary may uncover results that we could not predict. Being able to answer questions such as "How would respondents behave differently when completing a shorter diary?" would be beneficial.

Annex

Table 10 – Selection of expenditure categories with better than average levels of fatigue, 2015/16 data

Expenditure category		Recorded Purchases Week 1	Recorded Purchases Week 2	Fatigue (%)
1.1.1.2.2	Buns, crispbread and biscuits	10628	10000	-5.9
1.1.1.4.1	Cakes and puddings	4708	4365	-7.3
1.1.3.3.1	Dried, smoked or salted fish and seafood	495	466	-5.9
1.1.3.4.1	Other preserved or processed fish and seafood and fish and seafood preparations	2874	2735	-4.8
1.1.5.1.1	Butter	1407	1376	-2.2
1.1.6.8.1	Dried fruit and nuts	2115	2080	-1.7
1.1.8.5.1	Edible ices and ice cream	1820	1708	-6.2
1.1.9.3.1	Baker's yeast, dessert preparations, soups	3248	3118	-4.0
2.1.1.1.1	Spirits and liqueurs (brought home)	672	659	-1.9
5.5.2.1.4	Electrical consumables	649	619	-4.6
5.6.1.1.2	Disinfectants, polishes, other cleaning materials and some pest control products	3735	3460	-7.4
5.6.1.2.1	Kitchen disposables	4995	4628	-7.3
9.5.2.1.1	Newspapers	5925	5459	-7.9
9.5.4.1.1	Stationery, diaries, address books, art materials	1372	1315	-4.2
12.1.3.1.2	Toiletries (disposable)	3819	3646	-4.5
12.1.3.1.6	Cosmetics and related accessories	2345	2258	-3.7

Table 11 – Selection of expenditure categories with worse than average levels of fatigue, 2015/16 data

Expenditure category		Recorded Purchases Week 1	Recorded Purchases Week 2	Fatigue (%)
1.1.3.1.1	Fresh, chilled or frozen fish	1220	957	-21.6
1.1.4.2.1	Low fat milk	6689	5809	-13.2
1.1.6.9.1	Preserved fruit and fruit-based products	854	709	-17.0
1.2.1.2.1	Tea	1149	959	-16.5
1.2.2.2.1	Soft drinks	7382	6289	-14.8
11.1.1.2.7	Hot take-away meal eaten at home	4001	3158	-21.1
3.1.2.3.2	Girls' outer garments (5-15)	586	426	-27.3
9.4.1.1.2	Participant sports (excluding subscriptions)	692	553	-20.1
9.4.3.1.9	National Lottery stakes	1952	1563	-19.9
9.5.1.1.1	Books	909	769	-15.4
11.1.1.2.2	Adult Eating Out – Confectionery - Shops	928	663	-28.6
11.1.1.2.4	Adult Eating Out – Soft Drinks - Shops	2053	1633	-20.5
11.1.2.1.3	Adult Eating Out – Work	2927	2476	-15.4
11.1.1.1.5	Adult Eating Out – All other hot food - Catering	1906	1594	-16.4
11.1.1.2.6	Adult Eating Out – All other cold food - Shops	2913	2402	-17.5

Table 12 –Sample size required to regain precision, first week of diary data

COICOP category		Sample size needed to regain precision to original value (number of households)
1	Food and non-alcoholic drinks	8026
2	Alcoholic drinks, tobacco and narcotics	6989
3	Clothing and footwear	10304
4	Housing(net), fuel and power	5155
5	Household goods and services	6509
6	Health	3458
7	Transport	6102
8	Communication	4916
9	Recreation and culture	6539
10	Education	4817
11	Restaurants and Hotels	6468
12	Miscellaneous goods and services	5238

Table 13 – COICOP Categories as a Proportion of Total Purchases by Day of Diary Recording, 2015/16 data

CIOCOP category	Percentage of total purchases by day of diary recording (%)														
	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	**Total**
1 Food and non-alcoholic drinks	60.9	58.0	57.4	58.0	58.4	57.4	57.5	57.7	58.6	59.4	58.8	58.7	58.9	59.6	**58.5**
2 Alcoholic drinks, tobacco and narcotics	2.4	2.6	2.6	2.6	2.3	2.7	2.5	2.6	2.6	2.6	2.5	2.7	2.4	2.2	**2.5**
3 Clothing and footwear	2.1	2.2	2.3	2.3	2.5	2.7	2.4	2.5	2.2	2.2	2.2	2.5	2.4	2.4	**2.4**
4 Housing(net), fuel and power	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.4	0.3	0.3	0.2	0.3	0.3	0.3	**0.3**
5 Household goods and services	5.0	5.0	4.9	5.1	4.9	4.9	5.2	5.1	5.1	4.9	5.1	4.8	5.0	5.2	**5.0**
6 Health	0.9	0.9	0.9	0.9	0.9	1.1	1.0	1.0	1.0	0.9	0.9	1.0	0.9	0.9	**0.9**
7 Transport	2.4	2.5	2.5	2.4	2.5	2.6	2.7	2.5	2.5	2.5	2.4	2.4	2.5	2.6	**2.5**
8 Communication	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.2	0.3	0.2	0.3	0.3	0.4	**0.3**
9 Recreation and Culture	7.9	8.5	8.5	8.7	8.4	8.5	8.7	8.7	8.4	8.3	8.1	8.4	8.3	8.3	**8.4**
10 Education	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	**0.0**
11 Restaurants and Hotels	12.9	14.6	15.3	14.4	14.2	14.1	13.8	13.6	13.3	13.4	14.1	13.4	13.3	12.3	**13.8**
12 Miscellaneous goods and services	4.2	4.3	4.2	4.2	4.3	4.5	4.3	4.4	4.7	4.3	4.4	4.5	4.4	4.4	**4.4**
20 Other	0.7	0.8	0.9	0.9	1.0	1.0	1.1	1.1	0.9	0.9	1.0	1.1	1.1	1.2	**1.0**

Using respondent centric design to transform Social Surveys at ONS.

Laura Wilson, ONS[1]

Abstract

The Office for National Statistics (ONS) is not alone in its pursuit of the Holy Grail – that is, the introduction of online data collection, with a good take-up rate – to its portfolio in a time when response rates keep us all awake at night. ONS currently runs a well established, interviewer-led, mixed-mode operation of telephone and face-to-face interviewing for its voluntary Social Surveys. We are very comfortable and very competent in that space. However, it is time for us to respond to societal demands and also directives from central government to be 'Digital by Default', and rightly so. Online surveying brings new challenges, operationally and for design, but most importantly it brings opportunities - online is not the bad guy. It's an opportunity for ONS to change its business for the better and to put the respondent at the heart of what we do as opposed the data user, as is tradition. This shift is long overdue and is spreading through UK government service design like wildfire. After all, if there are no respondents then there are no numbers for our data users. This article will describe how ONS is developing a respondent-centric approach to the respondent journey, from the moment the letter lands on the doorstep to the questionnaire experience and then the between wave engagement.

1. What is 'respondent-centred design'?

It simply means putting the respondent in the driving seat when it comes to the design of your survey experience. We must not underestimate the importance of doing this and the importance of the 'experience' aspect of the interaction when it comes to non-response in longitudinal surveys. Self-completion surveys no longer mean that we can rely on highly trained and dedicated interviewers to provide a good experience for the respondent and to achieve that response. Interviewers play a vital role in maintaining response rates (especially around hard to reach groups) and gathering the correct information for complicated concepts which are cognitively demanding. They quickly come to learn which questions are troublesome however changes to question wording are kept to a minimum in order to manage time series data concerns. Instead the interviewers will use their skills and experience to work around the challenges; this is no longer acceptable.

The data-user-centred approach has left us with questionnaires that are long, confusing and sometimes repetitive. However, worst of all is that they often feel irrelevant to the respondent, leaving them feeling like they've not really represented themselves well enough or in the way they wanted to. This poor experience is a contributing factor to

[1] Office for National Statistics; laura.wilson@ons.gov.uk

declining response rates; however, it can no longer be overlooked when the tables are turned and the respondent becomes the interviewer and the interviewee.

User-centred design is not new, the term was coined the year I was born and for a long time it has driven the development of products (e.g. websites and apps) in the tech world with the aim of creating something which has high 'usability'. It promotes the creation of 'models' and 'goals' to assist the development, including end-to-end 'user journeys' and 'happy/unhappy paths' to achieve a great experience for the user. In turn, this hopefully means the user perhaps buys something, comes back again and spreads the word about how great their experience was – each contributing to the success of your product. Almost all of those desired outcomes are relevant to the goals of an online government survey.

It also means not striving for perfection before sharing or testing something – the aim is to 'fail fast'. It is important to learn quickly what is and isn't working so that you can change it and get back out there with the public to test it again. This development approach means that you don't put the roof on the house before the walls are built. It reduces the risk of sticking with something because you've started it now and it's too far gone or will take too much time and effort to change it. It keeps the respondent as the informant of the design.

2. How are we achieving respondent centred design?

We are transforming our data collection approach and operations through the Data Collection Transformation Programme (DCTP). The programme has five major strands, they are:

1. Introducing online mode of data collection for Social and Business Surveys
2. Data integration – changing to a default administrative data approach rather than default survey
3. Survey rationalisation and redesign (including the use of administrative data)
4. Systems integration (including registers)
5. Field force modernisation

The programme will deliver a huge amount of change for the office and to the data we produce. Ultimately, we've accepted that it will result in a break in the time series as we move to use administrative data more in the production of our data. It will also be very difficult to disentangle the cause of the change and attribute it to one strand. It is therefore an opportunity to take advantage of this change and make alterations to the questionnaire content.

We are moving to a default administrative data approach whereby we look to the admin data first to see if it meets the output need before asking a question on a survey. If it's not possible to directly match 'like for like' then we're taking the approach of 'is it good enough?'. This means that we can instead include questions on the surveys which add context to the admin data to get a richer data source. This change is enabling us to rationalise the survey content; we're reducing the size of our surveys which we hope will help to reduce attrition between waves.

When it comes to designing the questionnaire we are using respondent mental models/ psychology to inform the flow and wording. By this I mean we explore topics with respondents, e.g. work, education etc. We learn about how they think of those topics, what do they consider etc. Then using this information we can construct the flow of the questionnaire which makes it user centred and provides a better experience. As part of this we're exploring with interviewers to learn about the problems associated with the current flow and wording and utilising their experience and knowledge to fix those. This work has contributed to our learning about the way that respondents think and talk about the concepts we need to measure on surveys. I'll explain this in more detail in a later section – but for now, the takeaway from this is we are designing a flow that makes sense to the respondent. We are not letting the data user need fully dictate the experience. Instead, we are creating a more relevant experience for the respondent and using words that they would use to describe their circumstances. As a result, the survey tone has moved away from being formal - some respondents have described it currently as using the "Queen's English"- to become much more conversational. We are also writing for the average reading age of the UK public; it is key that our questions are clear and easy to read to get the data we need. The questions now sound more like how you'd chat to a friend about your job or hobbies, in turn creating a questionnaire that the respondent can connect with. These questions still meet the fundamental output need just in a different way.

When it comes to the respondent materials, we are using behavioural insights to inform the design. Also, we are using language that the public would use; however, this is a delicate balance as we've found that the public expect government to talk and sound a particular way. In addition, we are providing less information overall – we used to pack our letters with every possible worry or piece of information you could think of. Recent testing has revealed that less is more – it seems people's attention spans are decreasing and they only want to be hit with the important information. They can find the rest elsewhere if they need it. The respondent is taking the lead on the design and content.

3. Why are we changing in this way?

ONS tried the translation approach to 'going online' before embarking on the DCTP change journey. By that I mean, taking what we currently have and putting that online in its original form. DCTP's predecessor was the Electronic Data Collection Programme (EDC) which ran up until the end of 2015. EDC focused on developing an online version of the Labour Force Survey (LFS) which is the UK's principal measure for employment statistics. The EDC LFS questionnaire content remained the same, in terms of the number of questions, their wording and flow. However, we did make some small changes to the content in order to make it more accessible. The UK LFS is a beastly questionnaire - it has over 600 variables and 300 derived variables and whilst not everyone would be asked all of the questions, it still amounts to a very long, dry questionnaire. It was clear from qualitative research with the public that without interviewer intervention the questions were very difficult to understand, potentially impacting the data quality and providing a negative respondent experience. Unsurprisingly, we also found that respondents don't use online help; this is not a revelation - others researching online surveys have found the same. Therefore, this translation approach meant that we were trying to plug the leak in the pipe by adding more guidance instead of fixing the hole and addressing the cause; effectively, under transformation, we are now buying a new pipe.

We concluded that the questions simply were not suitable for self-completion by a member of the public and needed a radical redesign.

The Data Collection for Social Surveys European Statistical System Network (ESSNET) was running around the same time as this work. It concluded that we should expect a break in the time series when introducing online data collection and that we should take advantage of this by optimising for the mode.

The evidence from these two sources shaped the strategic direction of DCTP and resulted in the respondent centric approach we are pursuing today. We think this approach is one of the key players in helping to combat non-response in voluntary surveys across all modes. The other is optimising for the mode and developing online first, mobile first questions which I come on to next.

4. Optimising for the mode and online first, mobile first questionnaire design – what does this mean in practice?

Having a mode specific approach to collecting answers to questions is not new – for example, the use of show cards in face-to-face surveys and not in telephone is common. We are extending this approach to include online mode specific design; however, this is where we can potentially begin to see big differences in the wording and number of questions. In order to get good quality data and achieve a positive experience we need to let go of the 'uniformity is key' rule. Although, we do ensure that respondent comprehension, irrelevant of mode is standardised. Some of our surveys are trying to collect data on very complicated concepts and when an interviewer is present they can exercise their judgement, talk around it and also utilise any information given before that question was asked to help find the correct response. When the interviewer is taken away it may no longer be viable to ask one question to get an answer to your data need. We have a principle that we work to called 'consistent but not uniform'. We are reserving the right, if we identify a need, to have alternative wording and number of questions for the online self-completion version of the questionnaire. For example, we have found that sometimes it is necessary to break a question down into two or more questions to get to the answer we need in an online collection mode. There is often too much complexity tied into one question for a respondent to answer without additional information or support. Rest assured, data users are able to map back the data to one data point so it is comparable to other modes. This is where designing for the output need comes into play (I talk about this more in the following LFS section). Caution should be exercised here as we don't want to increase the length of the questionnaire unnecessarily, especially as we are working so hard to reduce it. However, we would much rather let the respondent work quickly through three easy questions and not notice, than struggle for five minutes trying to work out how to respond to one question. A nice example of this is the age old debate of traditional grids versus no grids. Yes, the grid is a neater or shorter way to display the information but for the user (i.e. the respondent) for them, it is much harder to complete. They need to work out what it's asking, what to answer and then how to complete it. They also spend a lot longer on that one page than they would on multiple questions collecting the same information. This all contributes to the high break-off rate at grids – the poor user experience. We are not doing the hard work in this case, the respondent is. I talk about this principle in the next few paragraphs.

Another principle we follow is 'harness the power of defaults'. When a respondent reads an online question, often they won't read the full question stem. Instead they'll glance over the response options to work out what the question is asking and then revisit the question for confirmation that they understand the task. As this is the case, we've challenged ourselves to question whether we should front-load the question (i.e. the stem). Instead, should we harness this default behaviour which comes with an online interface and focus on developing the response options? We don't apply this to all questions as it doesn't work for all - but we have found success with some. It speeds up the experience for the respondent.

We're finding that designing for online-first is creating a better experience for all modes. Our approach is to develop online first and then optimise from that question set for the other modes. After all, if you can create a questionnaire that can be understood by 'average Joe' public then surely it is suitable for use for interviewer modes too? We have taken this online optimised version and shared with it the interviewers and it's been welcomed with open arms. This results in a better experience for everyone. We are in the process of adapting it, where necessary, for the interviewer led modes. Again, with support from the interviewers we believe this will create a better respondent experience and reduce attrition.

When we design the questions for online first we do so for a mobile or smartphone in the first instance and not for PC or laptop users. This follows another principle that we have which is 'we do the hard work so the respondent doesn't have to'. By this we mean that we have to invest the time and effort upfront at the design stage into really thinking whether that question is suitable for completion on a smartphone. We have to think more like a commercial organisation - if we don't commit to design in this way then why should we expect a respondent to commit to completing the survey? We are competing against social media, shopping apps, broadcasting apps for the public's time – they are all designed to be user centred. If we don't do the same then why should they spend their time completing a survey over killing 15 minutes on social media? Especially, when there's no tangible reward in most cases. The smartphone-first approach really focuses the mind of the questionnaire designer to only display what is necessary as the screen space is limited. Now I don't mean that the tail is wagging the dog here when it comes to design i.e. the small space doesn't truly dictate the content. Instead, it makes the questionnaire designer think hard about the space they have to work in and how best to use it and get the concepts across. Its major benefit is that it facilitates the creation of short, snappier, direct questions with little room for typical government waffle.

The success of this online-first and smartphone-first approach has also encouraged us to be bold and design questions that do not have any respondent guidance. By radically overhauling the question set we can design very simple, clear questions and flow. We've tested this transformed set with the public and where we've found that it breaks then we've explored to understand why – is it the question, the placement? Where we've found it's the question, we redesign it. Where we've found it's the flow, we adapt it. Only then if that fails subsequent re-testing do we consider adding in guidance; we've found that this is rarely necessary.

The end result is something that is no longer burdensome for the respondent to complete. This in turn creates a more engaging and easy-to-use survey which hopefully means that the respondent will be happy to return to at future waves.

5. How are we researching the above approaches?

The respondent-centric design and the mobile-first may seem like a fluffy approach or overkill but it is much needed. We've adopted this approach and created respondent user journeys and the respondent happy path. We have already seen the benefits of this approach from a recent pilot. This has helped us to identify and detail the tasks, barriers and the workflow of the survey, in turn, allowing us to identify research questions for each step. We then design for each of these so that, for example, each potential barrier is not ignored but is instead overcome through user centred design. User journeys and user stories facilitate the user centred design.

We are researching this approach with the public using a mixture of new and old qualitative methods. For example, focus groups, in-depth interview, cognitive testing, usability testing and pop-up testing. We pick the right tool for the job at the time and aim to test at least once a month. The pop-up testing is a great way to explore ideas in the initial stages – to socialise designs and iron out any major creases before going into more intensive and costly questionnaire testing, such as cognitive testing. You can also speak to high numbers of people in a short amount of time and often people that you would not normally manage to capture in other types of testing. We've embraced this way of developing and found it extremely beneficial.

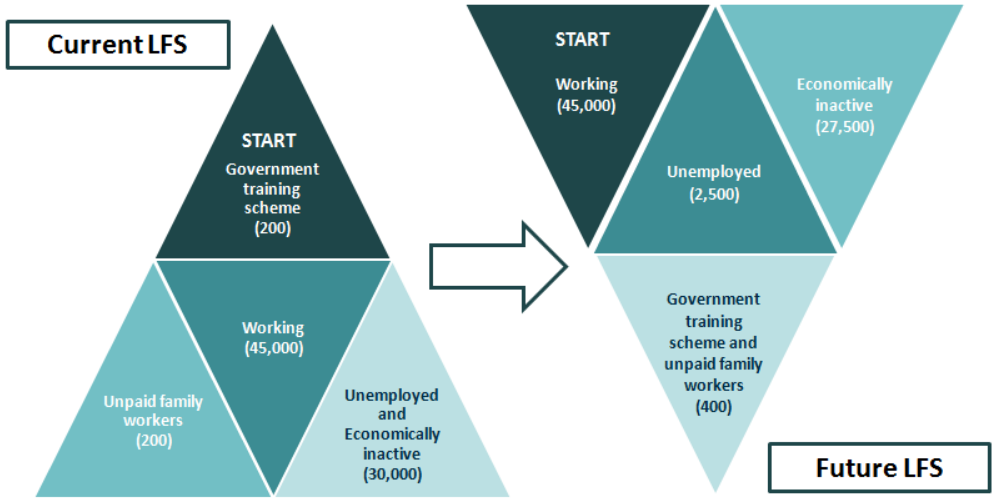
6. What does all of the above mean for the LFS?

Firstly, we are working on a subset of the LFS questions and referring to it as the Labour Market content. The transformation of ONS Social Surveys means that the model for collecting LFS data may change. In true agile form, we know that these questions are key and are the best place to start and gain some insights to inform future work. That being said, effectively, in employing all of the above approaches we've turned the traditional LFS on its head in order to move to create a future Labour Market questionnaire rather than a future 'LFS'. Like I said, we've identified the key variables that our analysts cannot live without and used these to create our rationalised future Labour Market questionnaire. This question set is much reduced; the count currently stands at circa 100 questions. We can't take on the whole LFS at once so we are breaking it down and working on small sections at a time. We have a period of reflection to see when future content could fit and allow ourselves to move the content around if we think it could be beneficial. We then asked our analysts what they *need* to output on, more generally, and not what they currently output. This gave us the freedom to change the questions and to design for the output need whilst not being constrained by the current wording and flow. We know there is a better, more respondent friendly way to get the data they need.

The next step was to design with data. By this I mean we looked at the volumes of the people being routed through sections of the questionnaire at any one time. See the diagram, fig. 1. The 'Current LFS' depicts the LFS structure and uses shading moving from dark to light to show the flow. As you can see from the numbers appended to each section, this questionnaire keeps a lot of people in the survey until the very end. This is a frustrating experience for most of these respondents as they are forced to be taken on the scenic route when they could have told you up front their situation and taken the motorway to the end. Looking at this along with the mental models and interviewer

advice I mentioned earlier, we redesigned the flow. The 'Future LFS' shows the end result and how we've turned it on its head. The numbers are inverted; we have classified the majority of respondents earlier than in the exiting questionnaire. This is creating a more relevant experience for the respondent and reducing frustrations. We take the approach of disqualifying respondents from an outcome. For example, they may have told us that they have a job (but in reality they may be waiting to start their new job). Our follow-up questions reveal why they weren't working in the reference week – that's because they've yet to start their job. By ONS definitions, these people are then reclassified as unemployed but as far as the respondent is concerned they have told us they are employed. The respondent then feels happy to have been recorded as this and is content with their experience.

Figure 1. LFS before and after



7. How far up the mountain are we? How's the trek been so far? What's the view looking like? These are my concluding remarks...

It has been challenging to get to this point in our development journey. However, now this work is endorsed and underway we are making great progress. We are confident in the approach we're taking and our key analysts are supportive. Our design is evidenced-based and if something isn't working we drop it and find an alternative solution. It's been a difficult journey but enduring the translation approach evidenced the need for the transformation approach.

The view is looking good; in July 2017 we achieved the highest take-up rate of any voluntary, non-incentivised, fresh sample, online social survey in the UK to date, achieving 19.9%. In September 2017 we ran another test, we fed in the learning from July and managed to improve the take-up rate, achieving 22.5%.

Our respondent-centric approach appears to be working and the feedback from respondents and interviewers is overwhelmingly positive. I appreciate that this approach is radical and controversial but I hope it inspires and encourages other survey organisations, both public and private, to consider its use in their design process. At ONS we're tackling non-response using an approach that we feel is breaking down one barrier at a time.

I would like to thank the following team members for their dedication, self-motivation and commitment to this work – Emma Dickinson (emma.dickinson@ons.gov.uk) and Alex Nolan (alex.nolan@ons.gov.uk). You have impressed managers immensely with your diligence and the endless hours that you have spent working on this project.

A History of Inflation Measurement

Jeff Ralph, ONS[1]

1. Introduction

The monthly consumer price inflation figures from the Office for National Statistics (ONS) are some of the most closely followed of all official statistics (ONS 2017); the publication of the statistical bulletin on the second Tuesday of each month is always reported prominently in the day's news (BBC 2017). The various consumer price inflation measures find wide application both as important indicators of the performance of economy and for adjusting wages and benefits (ONS 2016).

Inflation measures are based on a substantial data collection exercise and a sophisticated methodology. The ONS technical guide to Consumer Price Indices presents an overview of how the measures are constructed – it runs to 140 pages; the international manual provides a more detailed description and runs to 566 pages (ONS 2014 and ILO 2004). Both the methodology and the practical implementation are complex and to get to the current state of theory and practice has taken extensive development over a long period.

The overall story of this development is the subject of a recent book (O'Neill et al, 2017); this article describes five important elements of the overall story:

- The origins of the basket of goods and services
- The use of expenditure weights in a weighted index formula
- The measurement of household expenditure
- The organisations responsible for producing inflation measures
- The increasing uses of consumer price indices

Each of these topics played an important role in reaching the position we are in today.

2. What Needed to be Developed?

It is instructive to start by identifying the essential ingredients of a modern inflation measure. Summarising into a single sentence, the price change of each commodity in a representative basket of goods and services from a reference time period to the current period is measured and the changes combined, weighted by the proportion of expenditure by households on the commodities that the items in the basket represent.

An appropriate sample of goods and services is needed based on what consumers buy ensuring good coverage of the whole range of consumer goods and services on sale. The design for the collection of prices for instances of items in the basket ensures that comparable goods are priced through the year and that prices from different types of outlet in different locations are included.

[1] Office for National Statistics; Jeff.Ralph@ons.gov.uk

The need for a representative basket of goods and services to price at two time periods was a crucial early insight; while the number of items in the basket has increased over time, the concept is the same. Another important step was the realisation that price changes for items should be weighted according to the extent of household expenditure on them.

While the understanding that weighting data derived from household expenditure was an important part of constructing inflation measures, the practical collection of these data presented a significant challenge and took a considerable period to become established. The origins of capturing household expenditure are found in the investigation of the extent of poverty carried out by social reformers.

The early development of inflation measures was achieved by individuals who collected their own data; however, it gradually became apparent that the extent of data needed to ensure a reliable measure of overall price change was beyond the capabilities of individuals, however motivated they might be. An official body was required to carry out the work; this also took time to be established and is an important part of the overall history of development.

The evolution of the uses of inflation measures is important too as the uses have driven the development. The onset of the First World War and the rapid increase in prices led the Government to adjust wages for some essential workers by a measure of the level of prices. Once established this use spread to other workers. Gradually, adjustments using a price index, known as "indexation", was extended to various thresholds for tax and benefits and is now an essential part of modern practice.

The following sections describe these five topics in more detail; a much more comprehensive description can be found in O'Neill et al, 2017.

2.1 The Origins of the Basket of Goods and Services

The problems caused by volatile prices have a long history. Documents from the time of Hammurabi (2150 BCE) record attempts to control and accommodate price change. Over the subsequent centuries many other similar attempts were made, usually with limited success (Schuettinger, 1979). Over long periods of time, as well as volatility, the gradual increase in prices was recognised together with the problems this caused. In times of war and extreme natural events, prices tended to rise steeply and when this affected essential goods, it caused great concern for rulers, governments and citizens. The need to understand the change in the purchasing power of a set sum of money over time became pressing.

A very early example was provided by Bishop William Fleetwood in 1707 (O'Neill et al, 2017, p48). An Oxford College stipulated that a fellow would have to vacate his position if his income exceeded £5 a year – a rule specified in 1440. It was clear that the level of prices had risen significantly since that time and the bishop was asked to advise on the "equivalent" sum of money in year 1700 prices.

The bishop had a long-held interest in the "course of prices". In his investigation of this phenomenon, he studied prices for a variety of goods and services over a 600-year

period – he looked at the prices of corn, bacon, mutton and ale. He also considered the value of wages for workmen and servants, for example, the day rates for mowing an acre of meadow. To address the specific problem of the college rule, he chose four types of good – corn, meat, drink and cloth – goods considered relevant to a college fellow. By comparing the prices of these items between the two time periods he could determine an equivalent value to the original threshold amount for a college fellow. The bishop made two other important observations; firstly, prices should not be taken from a particular time period which would be advantageous to any party and secondly that the time periods should cover a sufficient duration to accommodate fluctuations in the prices – he chose twenty years.

When comparing the average of prices for each of the four types of item, he found each had increased by a factor of between five and six, so £5 in 1440-1460 was equivalent to £25-£30 in 1686-1706. The bishop considered another aspect of the problem before he reached a final conclusion; in modern terms, was the college statute meant to be an absolute or a relative measure? He decided it was the latter.

Several other instances of comparisons of the value of money across time periods in the 18th Century followed the pricing of a basket of goods and services, with the range of goods and services gradually extending in number, though their contents are best described as idiosyncratic. At this early period in the development, the prices collected were a mixture of wholesale and retail, with the former dominating.

Although the benefits from having a measure of the "general level of prices" were clear from the 18th century onwards, it wasn't until the early 20th century that an official price index was published and then, it was in a far from perfect form. The first official consumer price index was started in 1914 and was called the cost of living index; it was created as measure of the level prices for working class households with a basket containing 23 goods. It was used to adjust the wages of some workers considered to be essential to the war effort – a practice established to limit industrial unrest in response to rapid price rises as a consequence of war. Over the period of the First World War, prices almost doubled. This use of a price index to adjust wages in response of the prices of goods increasing became established and has continued ever since (O'Neill et al 2017, p118).

The number of items in the basket gradually increased over time. In 1947, the number had increased to 200; this became 600 by 1993; the 2017 basket contained 716 goods and services. The items in the basket are clearly a small subset of the total range of goods available in the consumer marketplace; however, these items are chosen carefully to be representative and are adjusted annually. For the update to the 2017 basket, only a few changes were required, which is the usual practice. There were 16 items added, 11 removed and 8 modified. While it might seem that more items would be better, there is a cost to collecting prices, so the overall number of items changes only slowly (Gooding, 2017).

The basket concept is followed across the world by all National Statistics Institutes. It was a vital development in the measurement of the general level of prices. While the concept was established early, the appropriate size of the basket took a long time to be determined.

2.2 The Use of Expenditure Weights

The pricing of the items in the representative basket at two time periods provides one of the two essential sets of data that are needed for measuring the general level of prices. The proportion of household expenditure for each item (or the range of items it represents) forms the other crucial data component. The need to combine price changes with a weighting to accommodate the relative spending on each item or group of items was an important insight.

The first documented indication that the price changes for some items should be more influential than others was seen in the writings of Arthur Young in 1812 (O'Neill et al, p55). In constructing his own measure of overall price change, he collected prices of a range of items and counted the price change for items in accordance with their "relative value". Wheat was counted 5 times, oats and barley twice, provisions 4 times, day labour 5 times; wool, coal and iron once each. He combined the price changes and divided by the sum of the weights. Using this approach gave a measure of overall price change that was significantly lower than previous estimates.

Despite his apparent insight, it is not thought that Young appreciated the importance of weighting in estimating an overall measure of price change. The Scottish economist Joseph Lowe is given the credit for understanding its importance; he is described by many as the "father of Index Numbers". In his book, The Present State of England (Lowe, 1832), he described the impossibility of predicting price fluctuations and lists their "injurious effects"; he makes a clear call for a measure of the "power of purchase" or the "power of procuring articles for consumption" and lists its uses both for national and individual purposes.

The use of weights was explained and clearly described by Lowe and includes a discussion on whether different sets of weights might be needed for different types of households. Although he didn't explicitly write a price index formula is his book, his writing is sufficiently clear that the formula used all round the world today for combining price changes is credited to him:

$$P_{Lowe}^{0,t} = \frac{\sum_{i=1}^{N} p_{ti} q_{bi}}{\sum_{i=1}^{N} p_{0i} q_{bi}}$$

Where the price of the i^{th} item at time period "s" is given by p_{si}, the quantity consumed q_{si} and the time period "s" stands for the price reference period (s=0), the current time period (s=t), or some other period (s=b).

Over the course of the following 185 years, many other index number formulas have been proposed and almost all use weighting information to combine price changes of different items. Although many of these formulas are more complicated than that proposed by Joseph Lowe, the Lowe formula remains almost universally used as a good compromise between theoretical validity and practicality (Ralph et al, 2015).

One of the strengths of the Lowe formula is that the weight reference time period, the "b", is very flexible. It is usually a period of time before the price reference period (the "0" in the formula) and is again usually a longer period than a month. In the UK, for the Consumer Prices Index including Housing (CPIH) measure, the weight reference period is the calendar year "y-2", where "y" is the current year[2]. The length of the weight reference period allows for sufficient data to be collected on household expenditure (Ralph et al, 2015).

The recognition that a measure of overall price changes would require weighting was another very significant development in the methodology.

2.3 Household Expenditure Data

While it was important to recognise the need for weights from a theoretical viewpoint, it left the significant challenge of the practical capture of data. The collection of household expenditure data was started by individuals investigating the extent of poverty. At the end of the 18th century, Davies and Eden were motivated by the rising price of bread and wanted to understand the impact on agricultural labourers and their families. With these families spending three quarters of their income on food, the impact was significant; Davies and Eden recommended that wages should be adjusted by a measure of costs. A different investigation was carried out by Edward Smith who was interested in identifying minimum dietary standards in 1860; he carried out a survey comprising 370 family budgets and found that the average diet was below the required standard as understood at the time (O'Neill et al, 2017, p197).

Towards the end of the 19th century, two social reformers financed private investigations of household income and expenditure. Charles Booth and Seebohm Rowntree collected data from households in London and York respectively. They found that about 30% of households had barely adequate provision to survive – this figure was about ten times the official number of registered paupers. They would have liked to extend their investigations to cover other parts of the country, but the cost was prohibitive. They believed that they needed to capture data from every household – this limited the extent of their work. Arthur Bowley, a statistician at the London School of Economics, explained that this was unnecessary and a carefully chosen one-in-twenty sample would provide sufficiently accurate information. His subsequent survey in Reading in 1913 is arguably the first use of random sampling in a social survey in the UK.

Pioneering individuals established the value of household budget data; however, they didn't have the resources to capture the information on a national basis – that required an official body to take responsibility. The Board of Trade was the organisation who took on this role. Official enquiries started very modestly; following some very small surveys at the end of the 19th century, the Board of Trade carried out a survey of 114 agricultural workers in 1902 (O'Neill et al, 2017, p95).

The years immediately following saw the Board of Trade dedicate significant resources to both household budgets and the collection of prices; this followed a request from the prime minister, Arthur Balfour, for improved data on trade, wages and the cost of living.

[1] The weight reference period for the Retail Prices Index is the start of July of year "y-2" to the end of June of year "y-1", where "y" is the current year

In 1903, the Board published two reports, the first established the extent of knowledge on wholesale and retail prices; the second described a more extensive survey of household expenditure for food, achieving 286 usable returns, though most were from London. The first report was an important step – it was the first time that retail prices had been brought together in an official report. It made clear the limitations of available price data and the wide variety of sources of variable quality, including data from trade associations, farmers' clubs and hospital and asylum records. The extent of important elements of household expenditure such as clothing and rent was very limited; the report assembled the data that were available. Taken together, the two reports allowed the Board to produce a measure of the change in the cost of living for working class families for London only for the years 1877 to 1900. It showed a decline in the (food) cost of living of nearly 30 per cent between 1877 and 1903 (O'Neill et al, 2017, p102).

A more extensive survey of households was carried out in 1904, capturing data from 1944 households and presented a much more even geographical distribution. Although it was still a relatively small sample to cover the whole of Great Britain and Ireland, it represented a significant improvement over previous studies. Questionnaires were sent out in July, August and September 1904. Expenditure items included fourteen food types, rent, clothing, fuel and light. The Board used historical data for the non-food items to produce individual time series of index numbers back to 1877.

The cost of living index numbers derived from the 1903 work on food for London were now revised to include the wider sample of households and included non-food goods. Although the combined time series of cost of living index numbers showed a fall from 1877 to 1900 as before, the values fell by only 17 percent. While food costs had fallen sharply over the period, other items such as rent had increased. This household expenditure data, with some updates in 1914, was used as the weighting information for the cost of living index for the decades following.

While new products came into the marketplace between the wars, there was little enthusiasm in official circles for carrying out another expenditure survey. It wasn't until 1937-38 that a new survey was commissioned and carried out. However, the onset of the Second World War stopped the implementation of new expenditure weights which weren't incorporated into the cost of living index until the war was over.

The need for the regular updating of household expenditure was firmly established in the 1950 and an annual survey carried out from 1957. Looking back, it seems extraordinary that the consumption pattern from 1904/1914, based on "working-class families" only, was incorporated into the cost of living index unchanged until after the Second World War.

2.4 An Organisation to Produce the Index

The topics identified above are developments of consumer price methodology; the topic in this section is different, but is no less important. While the early history of measuring the general level of prices was the story of insightful individuals, it became clear during the 19th century, that the scale of activity required to produce reliable, national measures would need the resources of the state. It was an important development to appoint an appropriate body to collect and process the data and produce an index that

had the confidence of the organisations and individuals who use it. The responsibility for producing the index had to be developed as did the methodology.

The early history was dominated by highly motivated and insightful individuals; several of them have already been mentioned already – Bishop Fleetwood, Arthur Young and Joseph Lowe. They collected their own data from whatever sources they could find. By the middle of the 19th century, a more systematic approach began to emerge. It was still dependent on enlightened individuals and private organisations; however, the need for regular, consistent data collection was recognised. The Economist magazine published a commodity price index in 1864, with data going back to 1845. In Hamburg, at about the same time, the Chamber of Commerce published price and quantity data for more than 300 commodities.

From about 1880 onwards, an official body, the Board of Trade, took on the responsibility for establishing data on household expenditure and prices. The scale of work started modestly, with small numbers of households, gradually extending the work over the next few decades as political priorities raised the profile of the need for better data on the levels of wages and the cost of living.

From 1916, employment exchanges were responsible for collecting prices as part of the Ministry of Labour, which took over responsibility for parts of the functions of the Board of Trade, including producing some of the statistics. In 1939, it was renamed the Ministry of Labour and National Service. Various further changes of name occurred over the subsequent decades and eventually, the responsibility was included into the remit of the Central Statistical Office in 1989, and then the Office for National Statistics in 1996 (Ward et al, 1991).

2.5 The Uses of Consumer Price Indices

The development and use of consumer price indices have gone hand in hand. As the uses have increased, so the pressure to ensure that the methodology is the best it can be has grown. One way of viewing this is that as the amount of money affected by consumer price inflation measures increases, so the focus on developing the measures grows too; this process continues to this day.

The earliest applications of a measure of the level of prices were to convert the purchasing power of a set sum of money from one time period to another. Wider potential uses were recognised in the 19th century though at that time there were only measures of the level of prices produced by a few individuals. With the outbreak of the First World War, prices rose sharply and the potential for industrial unrest grew. The government at the time was very keen to ensure that essential work supporting the war effort should not be interrupted by rising food prices and the hardship that would follow. They wrote to Trade Boards urging them to adjust wages in line with the cost of living, as measured by the working-class, cost of living index. This action established the principle of adjusting wages in line with a measure of the general level of prices.

After the war, the adjustment of wages by a measure of prices was extended to more workers; the Ministry of Labour estimated that 3 million workers were covered by what were called "sliding scales" by 1922. The application of these scales included the reduction of wages where the level of prices fell. As prices fell through the 1920s many

such formal agreements were abandoned, though wages tended to roughly follow prices. From the mid-1930s, prices increased and the Trades Unions pressed for wage increases. Also at this time, the methodology of the measure came under scrutiny, with the employers' and employees' sides criticising the measures in different ways (O'Neill et al, 2017, p124).

The use of indexing was gradually extended. Index-linked Government Bonds were introduced in 1981; a billion pounds of index linked gilts were issued with the Retail Price Index as the adjustment measure. A report from 1986 identified seven major uses of the Retail Prices Index:

- assessing changes in the standard of living of consumers
- monitoring the effectiveness of counter-inflation policies
- calculating the purchasing power of after-tax incomes, interest payments etc.
- deflating statistics, such as the value of retail sales
- uprating social security benefits, state pensions, the capital value of some National Savings
- gilt-edged securities and the level of tax thresholds
- providing proxy measures to stand for more specific price indicators; for example, amounts covered by insurance
- pay bargaining

By this point, it was clear that the uses of the index had grown substantially. Some of the uses of the index were put on a statutory basis in 1992 with the Social Security Administration Act. For example, the adjustment of the state pension and incapacity benefit in line with a measure of prices were statutory, while other benefits such as child benefit and job seekers' allowances were discretionary. For those which lacked statutory adjustment, the Secretary of State was still obliged to examine the value of these benefits in the light of the change in the level of prices and the general economic position (O'Neill et al, 2017, p9).

A new use was introduced in 1992 as the Government decided that inflation would be targeted as part of their monetary policy, following the departure of the UK from the European Exchange Rate mechanism. The choice of measure for this was a variant of the Retail Prices Index (RPI), RPIX, which excluded mortgage rates; the target was set at 2.5%

In 2003, the Chancellor announced that the inflation target would switch to the Consumer Prices Index (CPI), with the target set at 2%. The difference between the RPI and CPI based targets followed from differences in their construction (O'Neill et al, 2017, p230). While this change didn't cause much controversy, subsequent changes were received rather differently. In the 2010 budget, the Chancellor announced that state pensions, housing benefits and tax allowances would be adjusted in line with the CPI, which usually had a lower value. This change meant that benefits would grow more slowly. The discussion of whether the RPI or the CPI is a more suitable measure for indexing continues to this day.

3. Conclusions

This paper has identified 5 important aspects of measures of consumer price inflation and has examined their development. While they represent different aspects of the overall story, they each show the extensive work that has been required to get to the position we are in today.

The overall development has gone through several distinct phases, starting off with pioneering individuals who established the fundamentals with the state taking over responsibility for the large-scale collection of data and the continual improvement of the methodology.

The current position is one where development is carried out through international collaboration. Good practice has been established on an international basis and researchers across the world join forces to develop the fields of index numbers and price statistics.

While we have reached an elevated position with complex methodological standards and sophisticated data collection and processing operations, both those producing and using our modern measures sit on the shoulders of very many individuals and teams who worked to get us to where are today.

References

BBC 2017, News: Inflation steady despite food price increases, http://www.bbc.co.uk/news/business-41982269

Gooding, P. (2017). Consumer price inflation basket of goods and services: 2017. Newport: Office for National Statistics. www.ons.gov.uk/releases/consumerpriceinflationbasketofgoodsandservices2017

ILO, 2004, Consumer price index manual: theory and practice, http://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331153.pdf

Lowe, J. (1823). *The present state of England in regard to agriculture, trade and finance: with a comparison of the prospects of England and France*. London: Longman, Hurst, Rees, Orme & Brown. 2nd edition.

O'Neill R et al, 2017, Inflation: history and measurement, O'Neill R, Ralph J and Smith P A, Palgrave Macmillan, https://www.palgrave.com/gb/book/9783319641249

ONS 2014, CPI technical manual, http://webarchive.nationalarchives.gov.uk/20160108054359/http://www.ons.gov.uk/ons/guide-method/user-guidance/prices/cpi-and-rpi/cpi-technical-manual/index.html

ONS 2016, Users and uses of consumer price statistics, Lewis J, https://www.ons.gov.uk/economy/inflationandpriceindices/methodologies/usersandusesofconsumerpriceinflationstatistics

ONS 2017, UK Consumer Price Inflation: November 2017, https://www.ons.gov.uk/economy/inflationandpriceindices/bulletins/consumerpriceinflation/november2017

Ralph, J., O'Neill, R., & Winton, J. (2015). *A practical introduction to index numbers*. Chichester: John Wiley & Sons. https://www.wiley.com/en-gb/A+Practical+Introduction+to+Index+Numbers-p-9781118977811

Schuettinger, R.L. & Butler, E.F. (1979). *Forty centuries of price and wage controls – how not to fight inflation*. Washington D.C.: The Heritage Foundation

Ward, R. & Doggett, T. (1991). *Keeping score*. London: Central Statistical Office.

Methodology Advisory Service (MAS)

The Methodology Advisory Service is a service of the Office for National Statistics (ONS); it aims to spread best practice and improve quality across official statistics through methodological work and training activity. The ONS has about one hundred methodologists - highly qualified statisticians and researchers; their primary role is to provide expert support, advice and methodological leadership to the ONS in producing and analysing National Statistics.

Methodology staff are arranged into Centres of Expertise, each comprising a team of specialists who keep abreast of research and developments in their area of expertise through contacts with academia, other national statistical institutes and the wider research community. Many of these Centres have international reputations and present research and applied work at conferences and at other meetings of experts in their fields. Examples of these centres are Sample Design and Estimation and Time Series Analysis.

The Methodology Advisory Service has a remit to extend the services of ONS methodologists beyond ONS into other public sector organisations. Every year, MAS carries out projects with customers addressing a wide range of statistical requirements. As well as calling on methodology staff, MAS can also draw on the wider expertise of statisticians, researchers and subject area specialists across the ONS. Further expertise is available though links with Universities.

Contact ***MAS@ons.gov.uk***

GSS Methodology Series

Latest reports in the GSS Methodology Series:

38. *100 Years of the Census of Production in the UK, Paul Smith*
39. *Quality of the 2010 Electoral Register in England & Wales, Neil Hopper*
40. *Modelling sample data from smart-type electricity meters to assess potential within Official Statistics, Susan Williams and Karen Gask*
41. *Using geolocated Twitter traces to infer residence and mobility, Nigel Swier, Bence Komarniczky and Ben Clapperton*
42. *Assessing the Generalised Structure Preserving Estimator (GSPREE) for Local Authority Population Estimates by Ethnic Group in England, Solange Correa-Onel, Alison Whitworth and Kirsten Piller*

Reports are available from:

https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/currentmethodologyarticles

Forthcoming Courses

GSS Statistical Training Programme

A series of government specific short courses (between 0.5 and 2 days in length) delivered by methodological experts in the field. These courses are delivered at ONS sites in London, Newport and Titchfield.

For further information on the available courses see the Statistical Training Service prospectus:

https://gss.civilservice.gov.uk/learning-and-development/training-events/training-co-ordinated-by-the-statistical-training-service/

or contact gss.capability@ons.gov.uk

The current timetable for 2017/18 is available for downloaded through the GSS Learning Curriculum in the above link.

Details of additional opportunities for learning can also be found in in the training events page. In summary these are:

MSc in Data Analytics for Government

This is available at the following universities: University College London, Oxford Brookes University and Southampton University. More details can be found via this link.

https://gss.civilservice.gov.uk/blog/2017/07/msc-data-analytics-government/

MSc in Official Statistics

Available at Southampton University. This has been replaced by the MSc. in Data Analytics but will continue for those already enrolled.

The degree in Official Statistics is part of the network of Master programmes provided at European level. Further details can being via this link.

https://ec.europa.eu/eurostat/cros/content/emos_en

European Statistical Training Programme 2018

The purpose of the European Statistical Training Programme (ESTP) is to provide statisticians the opportunity to participate in international training courses, workshops and seminars at postgraduate level. It comprises courses in Official Statistics, IT applications, Research and Development and Statistical Management. More information on the core program for 2018 can be found on the Eurostat website

http://ec.europa.eu/eurostat/documents/747709/6103606/2018_ESTP_catalogue_final.pdf/3d416601-b8f1-4cf4-8aee-9ce8891c6ea9

Enquiries

The Survey Methodology Bulletin is usually published twice a year, in Spring and Autumn. Copies of many previous editions are available electronically at:

http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/index.html

If you would like to be added to the distribution list please email ONS Methodology at:

methodology@ons.gsi.gov.uk

Or write to us at:

Philip Lowthian
Survey Methodology Bulletin
2nd Floor
Office for National Statistics
Drummond Gate
London
SW1V 2QQ

ons.gov.uk
visual.ons.gov.uk