



ONC(SC)00/15

ONE NUMBER CENSUS STEERING COMMITTEE

2001 HARD TO COUNT INDEX

1. This paper details the research and evaluation work carried out to inform the decision on the Hard to Count (HtC) index that will be adopted for the 2001 One Number Census methodology. The index is the primary stratifier within both the design of the Census Coverage Survey and the One Number Census estimation process.
2. **The Steering Committee is asked to:**
 - a) **note the paper;**
 - b) **agree the recommendations that:**
 - **the variables for inclusion in the index are:
unemployed;
multi-occupied;
private rented;
language difficulty; and;
imputed.**
 - **the hard to count score is the sum of the proportions of the variables;**
 - **the number of levels in the index is 3; and;**
 - **the index is split into a 40%, 40%, 20% distribution at the national level.**

**Owen Abbott
Census Division, Room 4200W
Office for National Statistics
Segensworth Road
Titchfield
Fareham
PO15 5RR**

2001 ONC HARD TO COUNT INDEX

1. Introduction

- 1.1 As detailed in ONS(ONC(SC))00/11, the Hard to Count (HtC) index provides a stratification tool for the first stage of the Census Coverage Survey (CCS) Design, to partition the sample of postcodes into groups which should have a similar underenumeration pattern. In other words, the index will be based on how difficult an area is expected to be to enumerate in the 2001 Census.
- 1.2 It is expected that underenumeration will, at a local level, be higher in areas characterised by certain social, economic and demographic characteristics. For example, the 1991 Census follow up survey provided evidence that people in dwellings occupied by more than one household (known as multi-occupancy) will have a relatively high probability of not being enumerated (Heady *et al*, 1994).
- 1.3 These characteristics can be combined in some way to 'score' all 1991 Enumeration Districts, and form an index at the national level. The primary reason for using 1991 EDs is that a wide range of information from the 1991 Census is available at a micro level, whereas this is not the case for postcodes.
- 1.4 This paper outlines the results of further research that was undertaken to ensure that the index chosen for 2001 was both a good indicator of census coverage and an efficient stratification tool.
- 1.5 The research is presented under 4 headings:
 - The calculation of the hard to count score;
 - The variables that should be included;
 - The number of levels; and;
 - The national level HtC distribution.

2. The calculation of the score

- 2.1 Two different methodologies were used to derive the hard to count score during the development of the index. The first, described in Brown *et al* (1999) ranked the EDs with respect to each variable and then assigned normal scores based on these ranks. The use of the normalised ranks prevented any of the variables having undue weight in an EDs score. This was seen as undesirable, and so a simple summation of the proportions of each variable was developed for the 1999 Census Rehearsal.
- 2.2 Further discussion and consultation has established a preference for the second option, as it is a more intuitive and simpler approach. The first method loses information from extreme observations, and thus an area with the highest level of multi-occupancy in the country (but no private renters etc) might not be classified as a hard to enumerate area. In the interests of brevity, further evaluation of these options will not be carried out in this paper.
- 2.3 It is proposed to sum the proportions of the variables in the derivation of the Hard to Count score in the 2001 One Number Census.

3. The variables for inclusion

3.1 Introduction

- 3.1.1 In the past three years, the following 1991 Census variables have been considered for inclusion in the Hard to Count index:
- Multi-occupied households;
 - Private rented households;
 - Unemployed Persons;
 - Young migrant persons;
 - Persons whose first language was not English; and;
 - Imputed households.
- 3.1.2 The prototype index first suggested in ONS(ONC(SC))97/10 used five variables based upon the characteristics found to be associated with the 1991 Census undercount by ONS and the Estimating with Confidence project (Simpson *et al.*, 1997):
- Multi-occupied households;
 - Private rented households;
 - Young migrant persons;
 - Persons whose first language was not English; and;
 - Imputed households.
- 3.1.3 A second, simpler version was developed for use in the 1999 Census Rehearsal. It used three variables:
- Multi-occupied households;
 - Private rented households; and;
 - Young migrant persons.
- 3.1.4 Each of the proposed variables (multi-occupied, private rented, unemployed, young migrants, language difficulty and imputed) is discussed further below:

3.2 Discussion of proposed variables

Multi-occupation

- 3.2.1 Evidence from recent Census and CCS tests have indicated that multi-occupancy is still very much a problem when finding households. Using 1991 data should not cause a problem, as this variable is likely to be quite stable and will not have changed significantly since 1991.

Private rented households

- 3.2.2 Areas of high private rented households are likely to be those that contain large numbers of students, young single workers and cheap housing. Again, these are the types of places where recent experience indicates that coverage is likely to be a problem.

Unemployed

3.2.3 While it is acknowledged that unemployment patterns have changed dramatically since 1991, it must be remembered that we are not looking for a variable that measures the proportion of unemployed – we are looking for something that has a similar distribution to the likely 2001 undercount. Those areas that had high unemployed in 1991 are still the types of areas where we can expect underenumeration. This may be for a number of reasons, but it is likely that this variable is picking up areas of council or ex-council housing. There were a few areas like this in the 1999 Census Rehearsal, and there were some problems with making contact and getting participation in these areas. It is also important to note that council housing is not likely to be picked up in the other proposed variables.

Young Migrants

3.2.4 Migration has always been linked with underenumeration. This is because it picks up areas where either there is a mobile young population, or a large population of students. These are the type of people that are not likely to be contacted by Census enumerators and make an effort to complete a Census form, although the use of a postal methodology may reduce this possibility.

Imputed Households

3.2.5 The variable measures the level of absent households that were imputed in 1991 – these are possibly areas where the population were either very mobile, difficult to contact or very evasive. Again, there are likely to be problems with these types of households in 2001.

Language difficulty

3.2.6 It is not easy to see what additional factors the variable will bring to the HtC index that is not included within the other variables. For instance, large ethnic households that are prone to underenumeration might live in rented or multi-occupied accommodation. Hence this is the least attractive variable of those proposed, although it will still be useful to evaluate whether it is a good predictor of census coverage.

3.3 Analysis of proposed variables using 1999 Rehearsal Data

3.3.1 The 1999 Census Rehearsal provided data with which to undertake explorations of the link between the proposed variables and the likely coverage levels in the 2001 Census. The analysis makes the assumption that the patterns found in the rehearsal data will be similar in 2001. Three measures that provided information about different aspects of coverage were used – coverage of addresses, coverage of persons within counted households and household non-response.

Coverage of addresses

- 3.3.2 Within a randomly chosen sample of 70 EDs, the differences between the 1999 Census enumerator household listing and the CCS household listing were used to indicate where the Census enumerators missed a household – and so an estimated household coverage level could be computed.
- 3.3.3 Unfortunately, this did not provide sufficient data to carry out an in depth statistical analysis of the link between the coverage of addresses and each of the proposed variables in paragraph 3.1.1. However, an examination of the broad trends within the data did provide some indication that multi-occupancy and private rented household were important factors for the coverage of addresses. This is intuitively sensible, as discussed previously it is areas of high multi-occupancy and renting where households are most likely to be missed.

Coverage of persons within counted households

- 3.3.4 The second measure used 1999 Rehearsal data output from the ONC matching system to derive an estimate of the coverage of persons within counted households. The number of unmatched CCS people within a matched household was used to indicate where the Census had missed a person. This was only possible for the CCS areas, and not for all the 1999 Rehearsal areas.
- 3.3.5 This variable was compared with each of the proposed Hard to Count variables, split into 10 groups at the national level. **Annex A** contains the resulting boxplots. The width of the boxplots is an indication of the number of observations for the group (the more observations, the wider the boxplot), and the blue line links the means for each group. A distinct downward trend from left to right would indicate that the variable is good at explaining coverage – if an area has a high proportion of a variable then we expect the coverage to be low.
- 3.3.6 The graphs suggest that the language difficulty, unemployed and imputed variables have the most pronounced downward trend, indicating that they may be useful predictors. A multiple regression analysis was undertaken to explore the relationships further. The coverage within households variable was used as the response (Y) variable, and the explanatory (X) variables were the proportions of the six proposed HtC variables in paragraph 3.1.1. Three outliers/influential observations, identified in the boxplots, were removed from the analysis to ensure that the results were sensible.
- 3.3.7 All of the variables were fitted and then removed using backwards elimination until only significant variables remained. The results are displayed in **Annex B**. This analysis provides some evidence that the important variables are the proportion of unemployed and the proportion of persons who were not born in an English speaking country (language difficulty). However, the regression analysis is not particularly reliable, as the model has only explained about 19% of the variation in the data, as indicated by the adjusted R squared statistic.

Form return rate

- 3.3.8 The third measure of coverage was the form return rate for the Census Rehearsal. This is a proxy for the likely non-response patterns in the 2001 Census. If the

assumption that the level of non-response due to the voluntary nature of the rehearsal is the same everywhere, then this measure will be highly correlated with household coverage. A similar set of analyses to those above were produced using this measure. However, the results are likely to be more reliable since there are more data, because the form return rates can be calculated for all Rehearsal areas than just a sample.

3.3.9 Boxplots showing the link between the variables and the form return rate are given in **Annex C**. The majority of the variables are showing a downward trend in form return rates, indicating that they are possibly good predictors. The most prominent seem to be the unemployed, imputed and language difficulty variables.

3.3.10 As before, a multiple regression analysis was used to explore the relationships. The results are contained in **Annex D**. They provide evidence that all of the variables with the exception of young migrants are good predictors of form return rates. The variable with the most significant result is unemployed which has the highest F statistic (136.4). This model is a lot more reliable than the former since it explains 49% of the variability.

3.4 Conclusions

3.4.1 The results discussed above have indicated that of the variables suggested, the young migrants variable should not be considered further, since it was not an important predictor of the form return rate. The evidence has suggested that all of the remaining variables should be included in the Hard to Count index.

3.5 Recommendation

3.5.1 Based on the results of these analyses, **it is recommended that the 1991 Census variables that should be included in the 2001 HtC index are:**

- **unemployed;**
- **multi-occupied;**
- **private rented;**
- **language difficulty; and;**
- **imputed.**

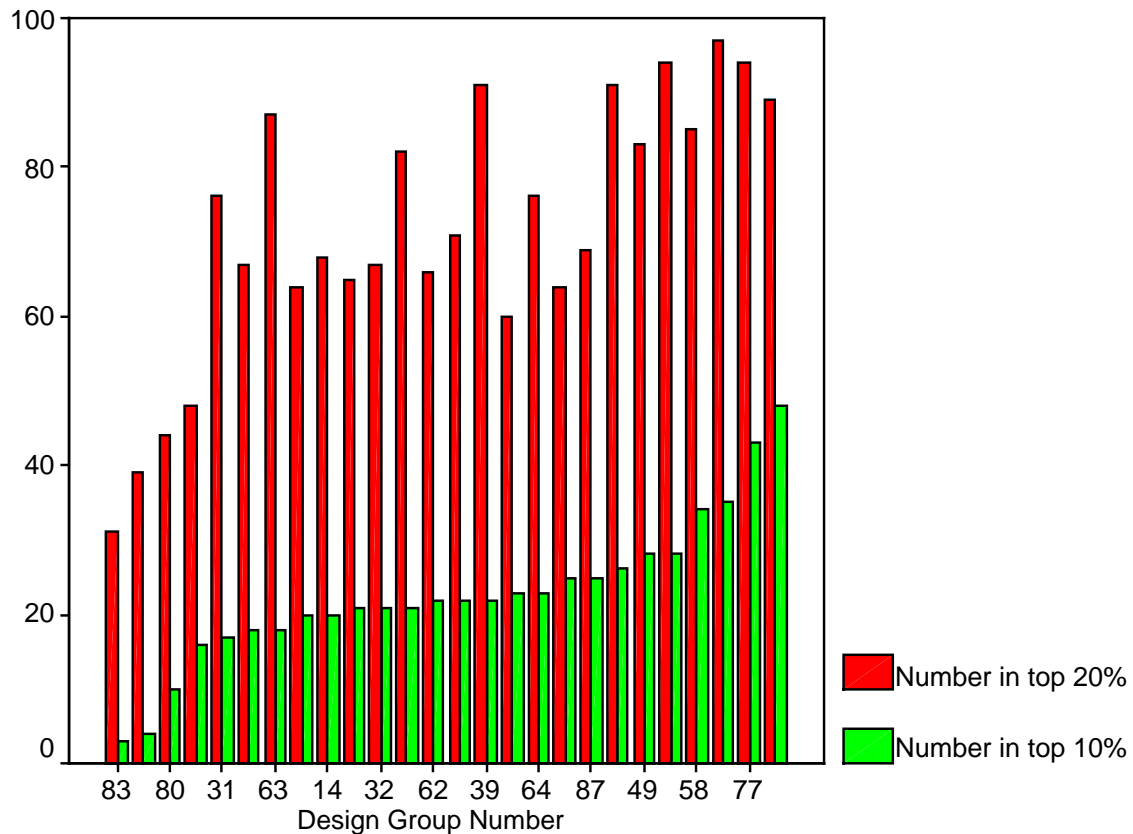
4. The number of levels

- 4.1 It is desirable from a Design and estimation perspective to minimise the number of index levels within a Design Group. The reasons for this are twofold. Firstly, the sample design is likely to be more efficient, as within a Design Group the population of each HtC strata may be too small if there are a large number of strata. Secondly, the use of less strata in the estimation process will reduce variance estimation problems as there should be more sample within each HtC group. Therefore, it is proposed to use a simple robust 3 level classification – which can be thought of as ‘easy’, ‘medium’ and ‘hard’ to enumerate areas.
- 4.2 As we are developing a robust approach to measuring underenumeration, 3 levels will provide greater protection from mis-specification of the index and reduce sample size problems. **It is therefore recommended that the number of levels in the index is 3.**

5. The distribution of the index

- 5.1 Four alternative distributions were examined:
- 70%, 20%, 10%
 - 60%, 30%, 10%
 - 50%, 30%, 20%
 - 40%, 40%, 20%
- 5.2 A simulation study was undertaken to evaluate the impact of the distribution on the precision of population estimates. The results indicated that there was no discernible difference between the four options tested. This result indicates that the distribution itself is not particularly important, provided that the HtC score is a good predictor of undercount. However, if the score fails to be highly correlated with coverage then we must have some protection against this. Therefore, to reduce the effect of any mis-specification a robust distribution should be adopted. The most robust of the options presented above is the 40%, 40%, 20% distribution.
- 5.3 The population sizes within each strata must also be examined, since the design of the coverage survey will not be efficient if the populations are too small. The sampling fraction within each Design Group will be roughly 4%. To enable the production of variance estimates we need at least 3 EDs within each HtC strata. Therefore the populations should not be less than roughly 75 EDs within each HtC strata for an efficient design.
- 5.4 The size of the Design Group ED populations within the top 10% and top 20% of the HtC score were compared as these are considered to be the most important. The Design Groups that will be used in 2001 are detailed in ONS(ONC(SC))00/10.
- 5.5 **Figure 1** displays the comparison for the Design Groups where the population in the top 20% is less than 100 EDs. The chart clearly shows that the populations are worryingly low for the top 10% - there are 9 Design Groups (out of 101) that have populations as low as 20. This compares to the lowest of 31 for the top 20%.

Figure 1: Number of EDs within the top 10% and top 20% of the HtC Score by Design Group.



5.6 Therefore, to avoid problems with small population sizes and mis-specification **it is recommended that that the 40%, 40%, 20% distribution be adopted for 2001.**

6. Recommended Approach for 2001

6.1 Based on the research and evaluation presented in this paper, it is recommended that the Hard to Count index used for the sample selection and ONC estimation processes in the 2001 Census is derived using the following methodology:

The Hard to Count score is

$$HtC_{score} = \frac{multiocc\ HHs}{total\ HHs} + \frac{imputed\ HHs}{total\ HHs} + \frac{priv.\ rent\ HHs}{total\ HHs} + \frac{unemployed}{total\ pers} + \frac{CoB\ is\ non\ english\ speaking\ persons}{total\ pers}$$

6.2 The EDs are ordered by the HtC score and split into a 40%, 40%, 20% distribution at the national level, with the top 20% being the EDs with the highest hard to count score.

References:

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999) A methodological strategy for a One Number Census. *Journal of the Royal Statistical Society A* **162**, 247-267.

Heady, P., Smith, S. and Avery, V. (1994) Census Validation Survey coverage report, OPCS.

ONS(ONC(SC))97/10 – Design of the Census Coverage Survey

ONS(ONC(SC))00/10 – Design Groups for 2001

ONS(ONC(SC))00/11 – One Number Census Methodology

Simpson, S., Cossey, R. and Diamond, I. (1997) *1991 population estimates for areas smaller than districts*. Population Trends, 90.

ANNEX A – Boxplots of persons within counted household coverage for each proposed HtC variable

The Y axis is the Persons with households coverage, and the X axis is the Hard to Count variable split into 10 groups at the national level – with level 10 being the hardest to count group.

Figure 1: Imputation Index

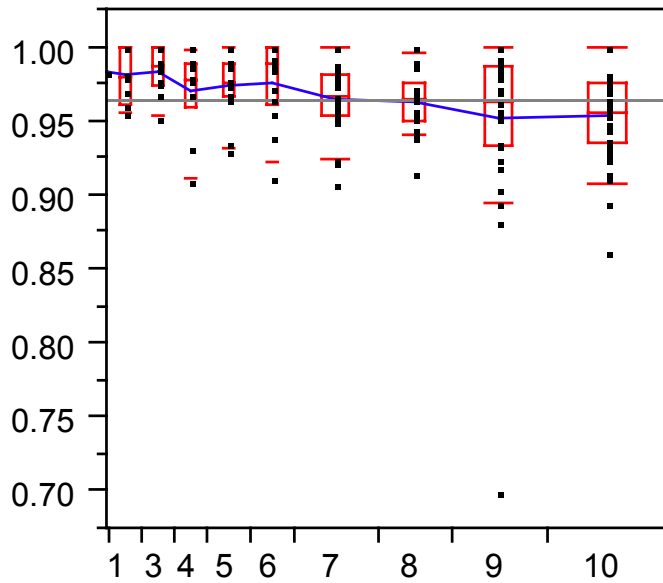


Figure 2: Private Rented Index

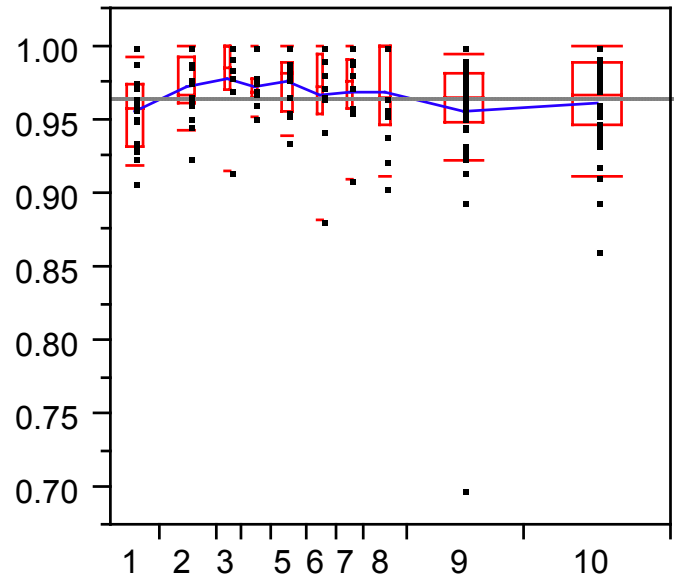


Figure 3: Multi-occupied Index

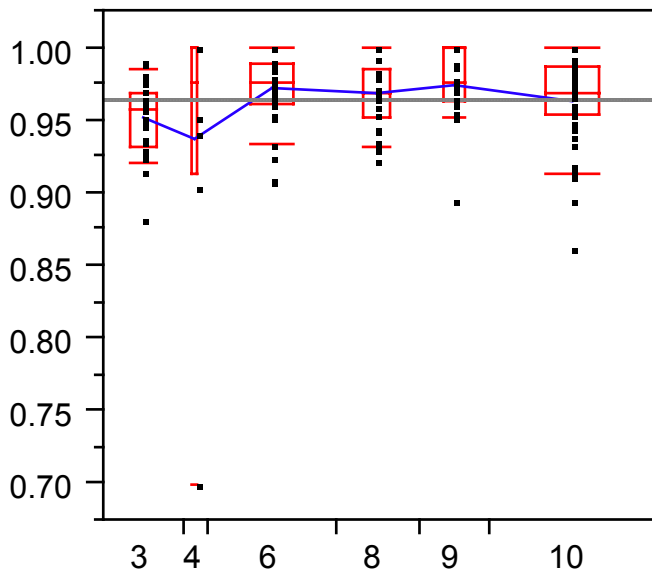


Figure 4: Young Migrants Index

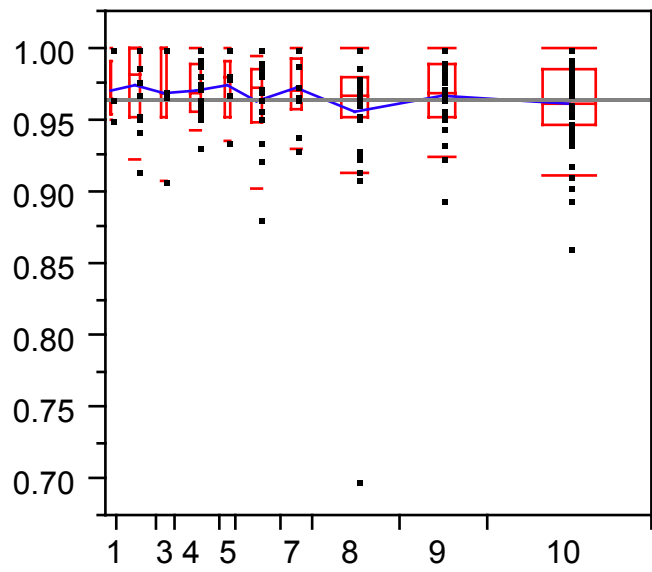


Figure 5: Unemployed persons index

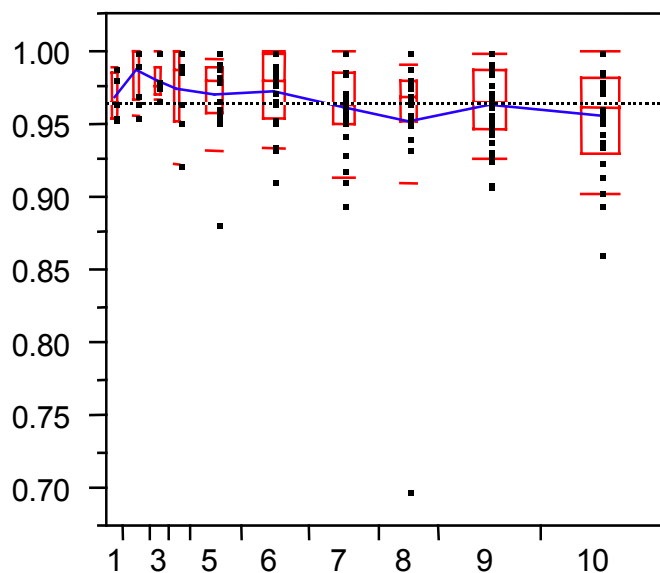
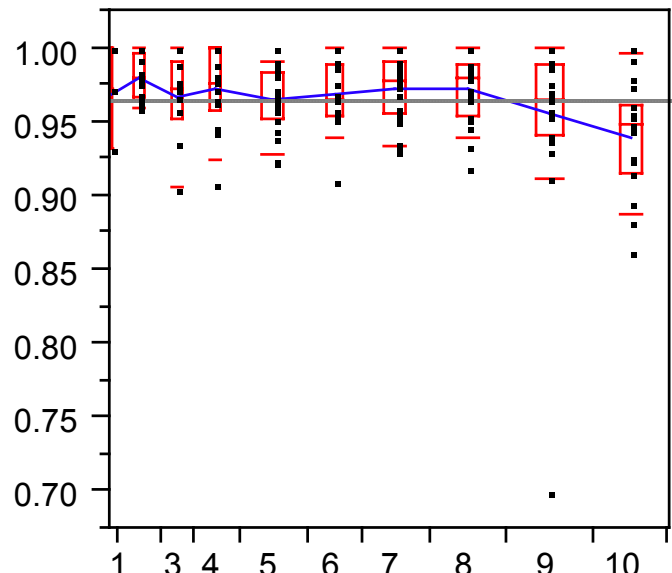


Figure 6: Language difficulty index



ANNEX B – Multiple regression analysis using persons within counted households coverage as the response variable

Response Variable: Persons within household coverage

Summary of Fit

R Squared	0.199372
Adjusted R Squared	0.189609
Root Mean Square Error	0.024106
Observations	167

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
UNEMPLOYED	1	1	0.00303241	5.2182	0.0236
LANGUAGE DIFFICULTY	1	1	0.01626803	27.9942	<.0001

ANNEX C – Boxplots of 1999 Rehearsal form return rates for each proposed HtC variable

The Y axis is the form return rate, and the X axis is the Hard to Count variable split into 10 groups at the national level – with level 10 being the hardest to count group.

Figure 1: Imputation Index

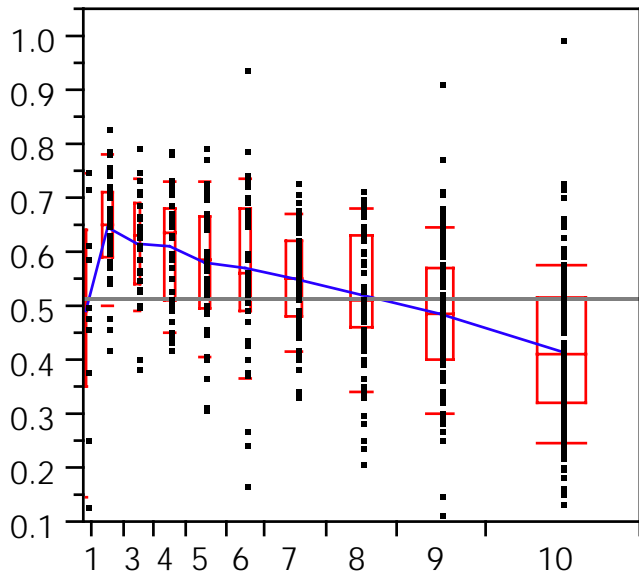


Figure 2: Private Rented Index

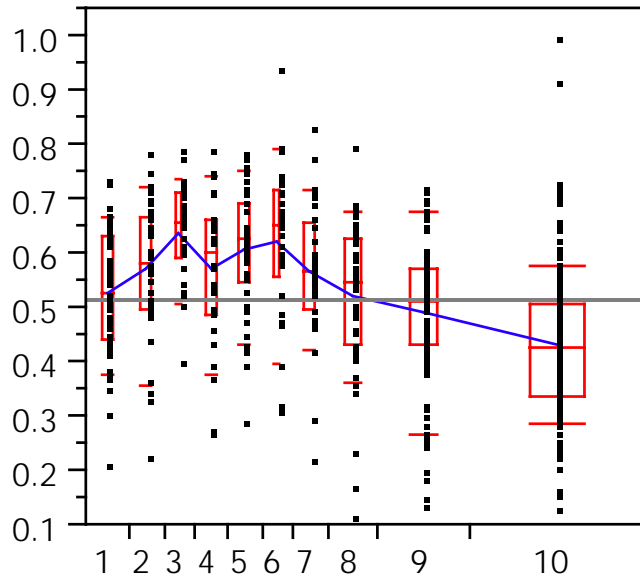


Figure 3: Multi-occupied Index

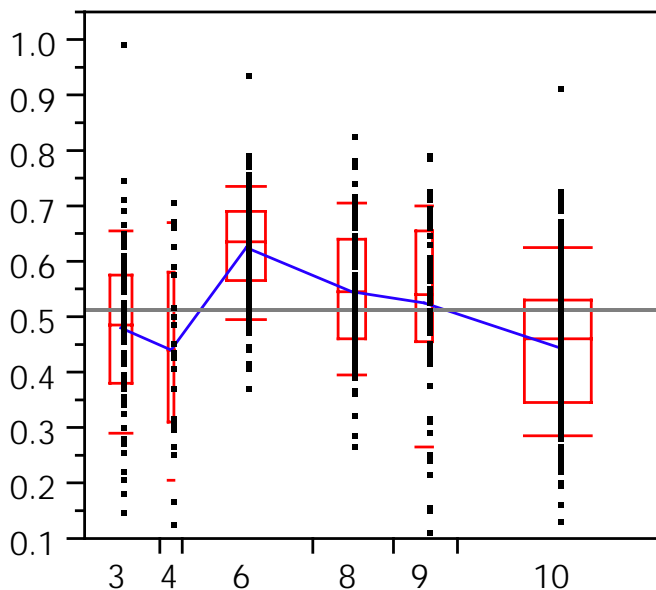


Figure 4: Young Migrants Index

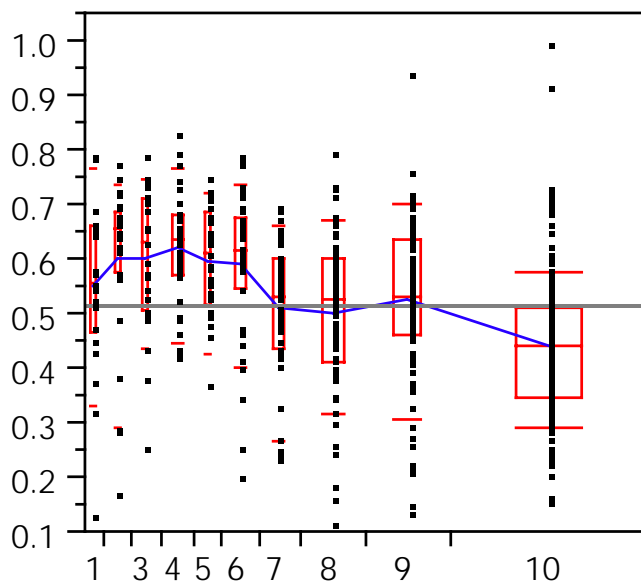


Figure 5: Unemployed persons index

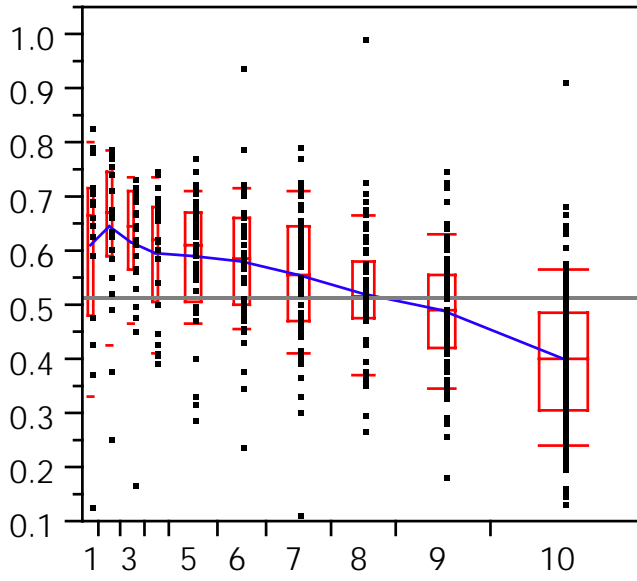
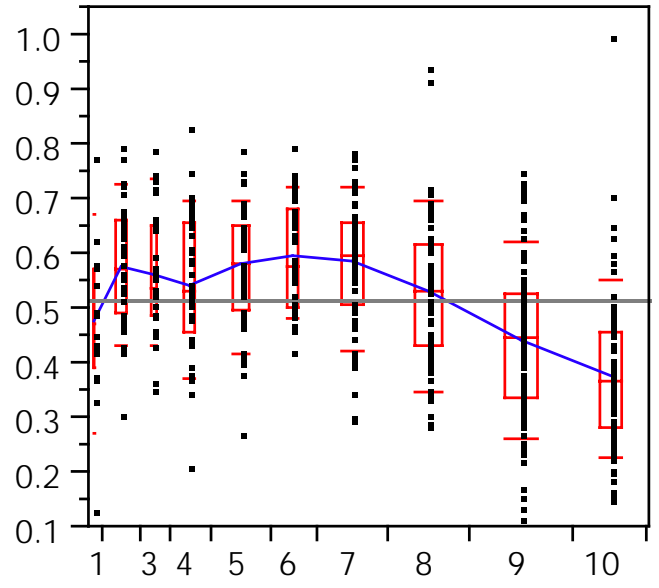


Figure 6: Language difficulty index



ANNEX D - Multiple Regression results using 1999 Rehearsal form return rates as response variable

**Response: 1999 Census Rehearsal Form Return Rate
Summary of Fit**

R Squared	0.494983
Adjusted R Squared	0.490325
Root Mean Square Error	0.102237
Observations	548

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
IMPUTED	1	1	0.5725229	54.7739	<.0001
PRIVATE RENTED	1	1	0.2850001	27.2663	<.0001
MULTIOCCUPIED	1	1	0.1609873	15.4018	<.0001
LANGUAGE DIFFICULTY	1	1	0.2403741	22.9969	<.0001
UNEMPLOYED	1	1	1.4255780	136.3867	<.0001